University of Alberta

Measurement properties of primary outcome measures in pediatric RCTs: inadequate reporting, insufficient validation, or both?

by

Zafira Karim Bhaloo

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Medical Sciences - Pediatrics

Fall 2013
Edmonton, Alberta

Dedication

Life, like research, is constantly evolving with new discoveries, new experiences, new challenges and new lessons to be learned. In this ever-evolving state of change, I have been blessed with a constant strength and inspiration…my family. I dedicate this thesis to my family, those who have an unwavering belief in me and who have taught me the greatest lessons of life. Family is the home we carry in our hearts, I am truly grateful for always having a place to call home.

Abstract

Randomized controlled trials (RCTs) are the gold standard for high quality evidence generation pertaining to treatment effectiveness. RCT results are used in knowledge synthesis and evidence-based practice. The validity of RCT results is dependent upon the validity of the primary outcome measures. Primary outcomes and the measurement properties of the primary outcome measures should be clearly reported to enable confident interpretation of results. A systematic review of pediatric RCTs published in high impact journals found inadequate reporting of primary outcomes and the measurement properties of their outcome measures. Furthermore, quality assessments of the measurement properties and the methodology of the studies in which the properties are evaluated suggest a need for further validation. The issue is thus twofold: inadequate reporting paired with insufficient validation. The awareness of this dual issue can encourage higher reporting standards and improved conduct of trials to ultimately enhance the quality of primary research.

Acknowledgement

"I have come to the conclusion that there is no greater form of preparation for change than education. I also think that there is no better investment that the individual, parents, and the nation can make than an investment in education of the highest possible quality. An education for success in the modern world must be enabling, and it must be outward looking." – His Highness the Aga Khan

There are countless people to thank for their continuous support, patience, and guidance as they invested in me and my education throughout the years.

I would like to sincerely thank my supervisor, Dr. Sunita Vohra, for her ongoing support, constant encouragement and for introducing me to the realm of possibilities within research. Her passion for research is refreshing and contagious and served as motivation for me throughout my research.

I would like to thank my supervisory committee, Dr. Lisa Hartling and Dr. Caroline Terwee, for their insight, knowledge and expertise. Their feedback always proved very valuable as I sought to complete my research.

I would also like to acknowledge and thank the entire Complementary and Alternative Research and Education (CARE) team. The staff and graduate students were always willing to share their experiences and skills to support me in my research. They showed great compassion, patience and collaboration.

Last but certainly far from least, I would like to thank those dearest to me – my friends and family. These individuals have served as role models, guides, teachers and empathetic colleagues. To my mom, thank you for always being my source of strength and endurance and always believing that I can achieve my goals. To my sister, Danisha, thank you for patiently listening to my frustrations, hopes, successes and unwarranted fears. You are my constant well of optimism and often the voice I needed to hear to

reinstill belief in myself. I am grateful to my family for their prayers and unwavering belief that I would get everything done. To my friends who have always lent empathetic ears, comforting shoulders and with whom I have shared endless hours of discussions, thank you for giving me reasons to laugh both in joy and frustration.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

RCT: Randomized Controlled Trial

CONSORT: Consolidated Standards of Reporting Trials

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

PORTal: Primary Outcomes Reporting in Trials

CENTRAL: Cochrane Central Register of Controlled Trials

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RDAI: Respiratory Disease Assessment Instrument

RACS: Respiratory Assessment Change Score

IFB: Infant Feeding Behaviours

CDRS-R: Children's Depression Rating Scale – Revised

CGI-I: Clinical Global Impressions – Improvement Subscale

ADHDRS-IV-Teacher:Inv: Attention-Deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator Administered and Scored

GCJ: Global Clinical Judgements

PIPP: Premature Infant Pain Profile

CARIFS: Canadian Acute Respiratory Illness and Flu Scale

GMFM: Gross Motor Function Measure

GMFCS: Gross Motor Function Classification System

BSID-II: Mental Development Index of the Bayley Scales of Infant Development II

MDI: Mental Development Index

CY-BOCS: Children's Yale-Brown Obsessive-Compulsive Scale

ODDI: Oblique Diameter Difference Index

ACPT: Auditory Continuous Performance Task

WRAML: Wide Range Assessment of Memory and Learning

WPPSI-R: Wechsler Preschool and Primary Scale of Intelligence-Revised

WRAVMA: Wide Range Assessment of Visuo-Motor Abilities

CY-PSPP: Children and Youth Physical Self-Perception Profile

m-YPAS: Modified Yale Preoperative Anxiety Scale

ABC: Aberrant Behaviour Checklist

NAPI: Neurobehavioural Assessment of the Preterm Infant

MIC: Minimal Important Change

ICC: Intra-Class Correlation Coefficient

SEM: Standard Error of Measurement

AUC: Area Under the Receiver Operator Characteristic Curve

SD: Standard Deviation

NIPS: Neonatal Infant Pain Scale

CRIES: Crying, Requires Oxygen, Increased Vital Signs, Expression, and Sleepless Measure

EASI: Emotionality, Activity, Sociability, and Impulsivity

STAIC: State-Trait Anxiety Inventory for Children

ENNAS: Einstein Neonatal Neurobehavioural Assessment Scale

SNAP-PE: Score for Neonatal Acute Physiology

NTISS: Neonatal Therapeutic Intervention Scoring System

NBRS: Neurobiological Risk Score

SPIRIT: Standard Protocol Items for Randomized Trials

VAS: Visual Analog Scale

LoA: Limits of Agreement

QMT: Quantitative Muscle Testing

PEDI: Pediatric Evaluation of Disability Inventory

MACS: Manual Ability Classification System

CFCS: Communication Function Classification System

KABC-MPC: Kaufmann Assessment Battery for Children Mental Processing Composite

IRT: Item Response Theory

BPI: Behaviour Problems Inventory

Chapter 1
Introduction

Randomized Controlled Trials

A clinical trial is considered to be the most rigorous method to determine whether or not a given intervention or treatment has a proposed effect[1]. Randomized controlled trials (RCTs) are viewed as the gold standard for high-quality evidence that assesses treatment effectiveness. These RCTs are at the top of the hierarchy of research design, augmented by knowledge synthesis efforts to collate their high quality research evidence (Figure 1-1). RCTs have been described as "one of the simplest but most powerful tools of research."[2] These rigorous studies have key features that differentiate them from other study designs, such as the random allocation of participants to the intervention groups, the blinding of patients and trialists to treatment allocation, and the controlled setting in which all groups are treated identically except for the intervention of interest, to name a few[3]. RCT results are preferentially used by health care providers, researchers, policy makers and other decision makers, and as such, substantial effort, time, and resources are allocated to their conduct. The quality of reporting of RCTs has gained much interest and initiatives such as the Consolidated Standards of Reporting Trials (CONSORT) aim to "alleviate the problems arising from inadequate reporting of randomized controlled trials (RCTs)"[4-11]. The reporting of primary outcomes and their outcome measures in trials is of particular interest and importance.

Primary Outcomes

Every RCT is conceived with the goal of answering a question or meeting a specific objective. While trialists may wish to address several questions, the RCT usually focuses on one main or primary question that the trialists are most interested in answering. It is with this question in mind that trialists seek to design their trial. This primary question or outcome is therefore the variable that is measured during the trial and is considered to be "the outcome of greatest importance"[12]. The primary outcome may be obvious in most RCTs, but despite

this, it is not always explicitly stated[4, 7, 10]. The primary outcome must be well reported as it is the variable on which the sample size calculation is based and forms the basis of results reported[1]. These outcomes can include changes in symptoms, adequate clinical response, duration of disease, change in the scores of instruments used, development and even death. Regardless of the nature of the primary outcome, it is imperative that it is reported clearly. It is suggested that clinical trials are "only as credible as their outcomes"[13], therefore if these outcomes are not reported the strength of the RCT itself becomes compromised.

Outcome Measures and their Measurement Properties

An outcome measure is the tool used to measure the outcome of interest, usually at various time points in a study[4, 5]. This tool may be a scale, questionnaire, instrument, or scoring system. Since outcome measures measure the outcome, they form the basis for the generation of trial results that are subsequently used in evidence-based practice. The validity of these outcome measures and their appropriate selection for a particular study is therefore a critical element of RCTs. If an inappropriate, invalid measure is selected and used in a study, the credibility of the results is compromised.

Outcome measures continue to be developed and there exist a large variety from which to choose. When selecting or evaluating a particular outcome measure, the measurement properties of the measure are compared. Measurement properties are not inherent traits of the outcome measure but rather are to be considered within the context of a particular study, its population and the context in which it is applied. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group reached consensus on the terminology and definitions of measurement properties in an international Delphi study[14]. Three domains of measurement properties were identified: reliability, validity, and responsiveness. Each domain contains measurement properties and a total of nine measurement properties are defined. Briefly, the reliability of a measure is defined as the degree to which patient scores remain the same for repeated measurements

over time, across persons and with different items, so long as the patients have not changed. Validity describes the extent to which an outcome measure measures what it is supposed or meant to measure. Of note, an outcome measure may be reliable but not valid. That is, it may measure something consistently but it may be measuring the wrong construct. Lastly, responsiveness is defined as the ability of an outcome measure to assess change over time when true change has occurred.

Thesis Objective

As RCTs are the internationally accepted gold standard for research and subsequent knowledge synthesis, it is important that their conduct and reporting is adequate. RCTs are published in large numbers in high impact journals yet concern exists about their reporting, in particular the reporting of their primary outcomes and the measurement properties of their measures[4-7]. This problem has been assessed to a limited degree in specified clinical areas. The magnitude of the problem of reporting has not yet been assessed across disciplines.

This thesis aims to assess the reporting of primary outcomes across pediatric disciplines as well as the reporting of the measurement properties of the primary outcome measures. More specifically, the objective of the thesis is to gauge the reporting of primary outcomes in pediatric RCTs and the completeness and accuracy with which measurement properties of primary outcome measures are reported. In order to do this, a systematic review of pediatric RCTs published in high impact journals between 2000 and 2010 was conducted. The RCTs reporting a single primary outcome were further assessed regarding reporting of measurement properties. Next, the studies reporting measurement properties of the outcome measure were assessed to determine if the authors' reports were accurate. Finally, the studies that used an outcome measure but did not report the measurement properties were evaluated to determine whether authors failed to report measurement properties of measures that have been validated or whether there is a lack of validation as a primary reason for lack of reporting. The highly

sensitive Terwee methodological PubMed search filter was used to identify relevant studies on measurement properties[15]. Retrieved studies were compared to the author's citations to determine whether the authors provided relevant citations. The COSMIN checklist with 4-point rating scale was used to assess the methodological quality of the studies assessing measurement properties[16-18]. A modified version of Terwee quality criteria was used to assess the quality of the measurement properties[19].

Chapter Based Thesis Objectives

Chapter 2, Primary Outcomes Reporting in Trials (PORTal): a systematic review of pediatric randomized controlled trials, aims to assess how many pediatric RCTs report a single primary outcome. Of the RCTs reporting a single primary outcome, those using an outcome measure are further assessed to examine whether measurement properties of the outcome measures are reported with citations.

Chapter 3 further examines the pediatric RCTs that report measurement properties of their primary outcome measures. The objective is to assess the accuracy of reporting of measurement properties for these outcome measures. The methodological quality of the studies in which the properties were evaluated is critically appraised and the quality of the measurement properties is also assessed.

Chapter 4 further examines the pediatric RCTs that do not report measurement properties of their primary outcome measures. The objective is to assess whether authors fail to report measurement properties when they are available to cite or whether the issue is a lack of validation of the measures such that no measurement properties can be cited. The methodological quality of the studies in which the properties were evaluated is critically appraised and the quality of the measurement properties is also evaluated.

Chapter 5 summarizes whether the issue is one of inadequate reporting, insufficient validation or a combination of the two. The larger implications of inadequate reporting and validation are discussed and future research recommendations are provided.

Thesis Objective	Methods	Thesis Chapter
1. Assess primary outcome reporting and the reporting of measurement properties of primary outcome measure	Systematic Review	Chapter 2
2. Identify if authors accurately report measurement properties.	Apply a highly sensitive search filter to identify studies on measurement properties. Terwee quality criteria and COSMIN 4-point rating scale are used to assess quality of properties and their studies.	Chapter 3
3. Identify if authors fail to report measurement properties when they are available to cite.	Apply a highly sensitive search filter to identify studies on measurement properties. Terwee quality criteria and COSMIN 4-point rating scale are used to assess quality of properties and their studies.	Chapter 4

Table 1-1 Thesis Objectives

Figure 1-1 Hierarchy of Research Designs

References

[1] Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. New York: Springer, 1998. Print.

[2] Stolberg HO, Norman G, Trop I. Fundamentals of Clinical Research for Radiologists. Randomized Controlled Trials. *American Journal of Roentgenology*. 2004; 183:1539-1544.

[3] Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? *BMJ* 1998; 316:201.

[4] Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement Issues in Trials of Pediatric Acute Diarrheal Diseases: A Systematic Review. *Pediatrics* 2010; 126:e222-e231.

[5] 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(4): e96.

[6] Reid GT, Walter FM, Brisbane JM, Emery JD. Family History Questionnaires Designed for Clinical Use: A Systematic Review. *Public Health Genomics*. 2009; 12:73-83.

[7] Zhang B, Schmidt B. Do we measure the right end points? A systematic review of primary outcomes in recent neonatal randomized clinical trials. *J Pediatr*. 2001;138(1): 76-80

[8] Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: Introduction. *J Rheum* 1993; 528–530.

[9] Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010;340:c365.

[10] Sinha IP, Williamson PR, Smyth RL. Outcomes in clinical trials of inhaled corticosteroids for children with asthma are narrowly focussed on short term disease activity. *PLoS One* 2009;4(7): 362-76.

[11] Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001; 285:1992-5.

[12] Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Ann Int Med* 2010:152. Epub 24 March.

[13] Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: Introduction. *J Rheum* 1993; 528–530.

[14] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737-745.

[15] 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;18:1115-1123.

[16] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Qual Life Res* 2010; 19: 539-549.

[17] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2011; 21:651-657.

[18] Cosmin | cosmin Retrieved 5/13/2013, 2013, from http://cosmin.nl/.

[19] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, and de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34-42.

Chapter 2
Primary Outcomes Reporting in Trials (PORTal): a systematic review of pediatric randomized controlled trials

Zafira Bhaloo, *Masters Candidate*[1], Denise Adams, *Research Associate*[1], Yali Liu, *Doctorate Student Visitor*[1], Namrata Hansraj, *Summer Student*[1], Lisa Hartling, *Assistant Professor*[2], Caroline B Terwee, *Assistant Professor*[3], Sunita Vohra, *Professor*[1,4]

[1]CARE Program, Edmonton Continuing Care Center
Unit 8B, 11111 Jasper Ave
Edmonton AB T5K 0L4
[2]Alberta Research Center for Health Evidence, Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta
Edmonton Clinic Health Academy
11405-87 Avenue
Edmonton, AB T6T 1C9
[3]VU University Medical Center, Knowledgecenter Measurement Instruments, Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
van der Boechorststraat 7
1081 BT Amsterdam
[4]Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta

Correspondence to: S Vohra svohra@ualberta.ca

BACKGROUND

Randomized controlled trials (RCTs) represent the gold standard for evidence about treatment effectiveness for health care providers, researchers, policy-makers and other decision-makers. RCTs are preferentially included in knowledge synthesis efforts such as systematic reviews and meta-analyses, which inform decision-makers at every level. Many RCTs are published annually in high impact journals; however, there is growing concern with regards to the reporting of outcomes and consequently the reporting of the measurement properties of the outcome measures, namely their validity and reliability[1-4]. As clinical trials are "only as credible as their outcomes"[5], a lack of reporting and validation implies that tremendous expense, effort, and resources may not being used optimally.

An outcome is a measurable variable that should be clearly stated by the authors and an outcome measure is the tool used for measuring the outcome (scales, questionnaires, instruments, or scoring systems – we describe these collectively using the term "outcome measure")[1]. The measurement properties of an outcome measure, i.e. validity, reliability and responsiveness provide information regarding the measure's intended purpose, its performance and accuracy, and its ability to detect a true change. When selecting which outcome measures to use in any given study or when evaluating the use of a particular measure, the measurement properties are often compared. Inadequacies related to primary outcome reporting and their consequent impediment on the conduct of knowledge synthesis efforts has been discussed in light of selective outcomes reporting[6].

The issue of selective outcomes reporting is secondary to a larger issue of trials that fail to identify any primary outcome at all. The inadequate reporting of outcomes in the pediatric population has been discovered often while investigating outcomes selection within a specified clinical area. In systematic reviews of RCTs within pediatric subspecialties, authors consistently fail to report identifiable primary outcomes[1,4,7].

Although it is recognized that the "prespecification of a single primary outcome based on biologic credibility, clinical importance, and potential responsiveness to the intervention" is the best approach, the reader is more often "offered a shopping list of end points"[4]. Along with the poor reporting of primary outcomes, the validation of outcome measures is also poorly reported or missing altogether. Limited studies state the use of a validated instrument or its formal evaluation against some sort of reference standard and the few that do fail to provide evidence of these measurement properties with citations[1,3].

A variety of initiatives[5-10] have been developed to address some of the issues of inadequate reporting and validation. To assess the magnitude of this problem across pediatric disciplines, we conducted a systematic review of a random sample of pediatric RCTs published in ten high impact journals between 2000 and 2010. The main aim was to examine (1) how many RCTs reported a primary outcome, (2) the number of primary outcomes reported, (3) how many RCTs reported the measurement properties of the instruments used, and (4) the relevant citations provided for the measurement properties reported. Our primary interest was assessing outcome measures, since these have been identified as in need of further study. A secondary aim was to examine other key pediatric trial metrics and their reporting, such as information about the population (participant ages, condition(s) under study, sample size and calculation), intervention and control group(s).

METHODS

Search Strategy

With the help of an experienced health research librarian, electronic searches in MEDLINE, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL) databases were undertaken. All searches used the respective journals name; six general medicine journals (*New England Journal of Medicine, Journal of the American Medical Association, Lancet, Annals of Internal Medicine, British*

Medical Journal, and *Plos Medicine*) and four pediatric journals (*Journal of the American Academy of Child and Adolescent Psychiatry*, *Pediatrics*, *Journal of Pediatrics*, and *Archives of Pediatrics & Adolescent Medicine*). Searches were limited by publication type (RCTs), publication year (2000-2010), respective pediatric filters, and the English language. The full search strategies for each database can be found in Appendix 2-1.

Study Selection

We included studies that (1) were RCTs, i.e studies that randomly allocated participants to interventions, and included parallel, cross-over, factorial and N-of-1 designs, (2) comprised of a single phase trial (or single step intervention) in a single publication as it is difficult to extract data for trials with multiple phases and steps that may contain different methods/interventions/outcomes in each phase/steps (multiple steps may also result in multiple primary outcomes and thereby skew our findings), (3) included only the pediatric population (less than 21 years of age) as it is unlikely outcome measures have been validated for both adults and children, (4) were of any intervention type, and (5) were published in the previously identified ten high impact journals between 2000 and 2010. We excluded: (1) studies that were diagnostic or screening in nature as this initiative was focused on improving reporting based on CONSORT guidelines, and other reporting guidelines exist for diagnostic and screening studies; and (2) self-described pilot studies, which may not place the same emphasis on primary outcome measure selection and reporting. A random sample (20%) of studies was selected and the titles and abstracts were screened by independent reviewers (ZB, YL, NH) for potential inclusion. Full texts of the selected articles were then retrieved and each article was independently assessed by the same reviewers for inclusion based on the pre-defined criteria. Disagreements were resolved with a senior team member (DA) and consensus was sought. The inclusion and exclusion screening form used is provided in Appendix 2-2.

Data Extraction

Three independent reviewers (ZB, YL, NH) used a standardized data extraction form (Appendix 2-3) and extracted variables including: journal, publication year, design of RCT, age values reported, condition and intervention of interest, sample size and calculation, specification of at least one primary outcome, the number of primary outcomes, outcome measure used, measurement properties reported for the outcome measure, and evidence supporting the reported measurement properties.

An explicit report or reference to a primary outcome was searched in the abstract and full text of all included studies. As an additional measure, the "find" tool was also used to identify any mention of a primary outcome or similar terminology within the text that may have been overlooked by reviewers. As per the CONSORT statement[11], "the primary outcome measure is the pre-specified outcome considered to be of greatest importance to relevant stakeholders". Great flexibility of terminology was accorded for the identification of primary outcome(s) (e.g. main outcome, primary end point, primary objective) and terminology used across RCTs was recorded.

Studies reporting a single primary outcome were further assessed for the report of an outcome measure. An outcome measure is identified as "a scale, scoring system, instrument, questionnaire or other tool used for measuring an outcome[1]." Measurement properties of the outcome measures reported were identified based on the COSMIN group's (COnsensus-based Standards for the selection of health Measurement INstruments) published standardized terminology, definitions, and taxonomy of measurement properties for the evaluation of instruments based on international consensus[9]. The "find" tool was also used to identify any mention of a measurement property within the text that may have been overlooked by reviewers. For studies reporting on the measurement properties of the outcome measures used, citations and bibliographies were searched for evidence to support

these reports. Any discrepancies in data extraction were noted and resolved through joint discussions with a senior team member (DA).

Data Analysis

This systematic review does not evaluate the effectiveness or safety of a particular intervention but rather focuses on reporting, therefore risk of bias and meta-analysis are not necessary or relevant. Data were entered into and analyzed using STATA. Results are described using descriptive statistics (summary scores, proportion, frequency) and presented as percentages.

RESULTS

Our electronic search yielded 2229 unique references (Figure 2-1). The titles and abstracts of a random 20% sample (n=445) were screened. Of these 445 articles, 29 were excluded as they were not RCTs, 44 articles were follow-up studies, 10 were pilot studies, 12 articles reported on more than one phase/step/trial, two were diagnostic and screening trials, 70 studies also included adults, four articles were not retrievable, and two were duplicate articles. The full text of 272 potentially relevant studies was retrieved and screened. A total of 206 RCTs were included for data extraction.

Of the included studies, 32% were from *Pediatrics*, 28% from *the Journal of Pediatrics*, 10% from the *Archives of Pediatrics and Adolescent Medicine*, 9% from the *Lancet*, 9% from the *New England Journal of Medicine*, 7% from the *American Academy of Child & Adolescent Psychiatry*, 4% from the *Journal of the American Medical Association*, 1% from *PLoS Medicine,* 0.5% from the *British Medical Journal*, and 0.5% from the *Annals of Internal Medicine*. Of the 206 RCTs, 89% were parallel in design; the remainder were crossover and factorial trials. The majority (65%) were treatment trials as opposed to prevention trials (35%). A median of two groups were studied in each trial (range 2-6; IQR 0). A variety of conditions were studied across the 206 trials and these included: type 1

diabetes, respiratory distress syndrome, patent ductus arteriosus, obesity, Kawasaki disease, bronchiolitis, cystic fibrosis, depression, asthma, and bronchopulmonary dysplasia. Only 63% of RCTs provided a sample size calculation and sample sizes ranged from 10 to 63 225 participants (median = 120, IQR = 321) (Figure 2-2). Most authors did not explicitly report actual age ranges (upper and lower bounds) of their participants but rather provided the mean age of their population. Variables extracted from the included studies are summarized in Table 2-1.

Primary Outcomes

A variety of terminology for "primary outcome" was recorded. This included primary outcome(s), primary endpoint(s), primary efficacy variable(s), main outcome measure(s), primary study variable(s), primary outcome measure(s), primary study end point(s), primary outcome variable(s), primary objective(s), primary pre-specified outcome(s), primary dependent variable(s), main outcome measurement(s), and primary efficacy parameter(s).

Of the 206 RCTs, 100 (48.5%) explicitly reported a single primary outcome, 56 (27.2%) did not identify any primary outcome, and 50 studies (24.3%) identified multiple primary outcomes. The 50 studies that reported multiple primary outcomes identified two to 20 outcomes as primary with a median of two primary outcomes (IQR 1) (Figure 2-3).

Outcome Measures

Of the 100 studies that reported a single primary outcome, 19 reported the use of an outcome scale, tool, or instrument to measure their primary outcome (Table 2-2). The other 81 studies used physiologic measures (eg. eosinophil-derived neurotoxin levels, calcium absorption, rate of decline in forced expiratory volume), diagnostic tools (eg. polysomnography, radiology), or quantitative indexes such as duration of stay in hospital to measure their primary outcome and were thus not evaluated further.

Of the 19 studies reporting the use of an outcome measure, seven (37%) reported measurement properties. All seven studies provided relevant citations to support their reports. Three (43%) of the seven studies reporting measurement properties examined measurement properties as part of their study. For the 12 studies that did not explicitly report any measurement properties, any citations provided for the outcome measures themselves were reviewed. We found that the outcome measure citations provided in 11 (92%) of the 12 studies were in fact relevant citations for measurement properties (Figure 2-4).

DISCUSSION

More than 10 years after CONSORT, one quarter of pediatric RCTs published in high impact journals fail to report any primary outcome. This is especially surprising as all of the journals included in our review have endorsed CONSORT. Furthermore, measurement properties of outcome measures are often not reported by authors although these measures are used to evaluate the trial's primary outcome. Since RCTs are "only as credible as their outcomes"[5], it is crucial that their outcomes are valid and reliable in the population in which they are being applied, and clearly reported as such.

The results of this study may be limited in part due to the methods used to search for our included studies. We recognize that assessment of reporting in only high impact journals may lead to an underestimate of the problem, however we chose this as our sample as knowledge users are more likely to be convinced of our findings if they cannot discount them due to their lack of familiarity with smaller journals (i.e. journals that they do not hold in high regard or aspire to publish in). High impact journals are assumed to have the most rigorous and stringent publication standards so if a significant problem exists in this group, then our findings likely under-estimate the extent of the problem if lower impact journals and grey literature were included. Furthermore, as we did not know the extent of the problem of reporting, our ability to perform a sample size calculation was

limited. We chose to assess a random 20% sample as we believe this represents a comprehensive and feasible sample of pediatric RCTs across disciplines.

Strengths of our approach include use of systematic review methods and reporting according to PRISMA guidelines[12]. Reviewers independently screened and extracted data from the studies using standardized forms. This systematic review also accepted a wide range of terminology for the reporting of a primary outcome. By recognizing the variety of terms used to identify a primary outcome, we avoid over-estimation and provide a clearer, fairer picture of the scope of the problem. Great heterogeneity exists in the author descriptions of primary outcomes. Of note, authors use "outcome" and "outcome measure" interchangeably. It is suggested that an outcome is a measurable variable while the outcome measure is the tool used for measuring the outcome (such as scales, questionnaires, instruments, or scoring systems) [1]. The inconsistency and heterogeneity of these terms across initiatives and organizations does not aid in clarity and it is time for trialists, editors, and guideline developers to reach consensus on acceptable terminology. Regardless of terminology, primary outcomes are not explicitly reported.

Although current literature has discovered the issues of outcomes reporting and the validation of instruments, they have been investigated to a limited degree and are restricted to individual disciplines[1-4]. A thorough synthesis of the problem has not yet been conducted. To our knowledge, our systematic review of pediatric RCTs in high impact journals is the first of its kind to look specifically at the problem of reporting and validation of primary outcomes and their measures.

RCTs are heavily relied upon by evidence-based decision makers, researchers, funding agencies, policy makers, peer reviewers, authors and journal editors. A substantial proportion of RCTs fail to report a single primary outcome and too often, measurement properties of measures are also left unreported. The validity of these trials is directly reflected by the validity of the primary outcomes and the

measures used. The results of this study can be used to improve reporting standards by facilitating the revision of reporting guidelines such that they require the clear reporting of a study's primary outcome and relevant citations for measurement properties of outcome measures. This study may also aid in the informed selection of outcomes and outcome measures by trialists and other clinical researchers. The research findings presented here have the potential to encourage higher standards for the reporting and conduct of trials such that RCT results can be used more confidently at every level of knowledge synthesis and translation.

While we have firmly established the inadequate reporting of primary outcomes in pediatric RCTs, the reporting of measurement properties needs further investigation. Authors may fail to explicitly report measurement properties or there may be a lack of formal assessment of these outcome measures therefore limiting the ability to report measurement properties. Whether the issue is one of inadequate reporting, insufficient validation or a combination of the two needs to be determined.

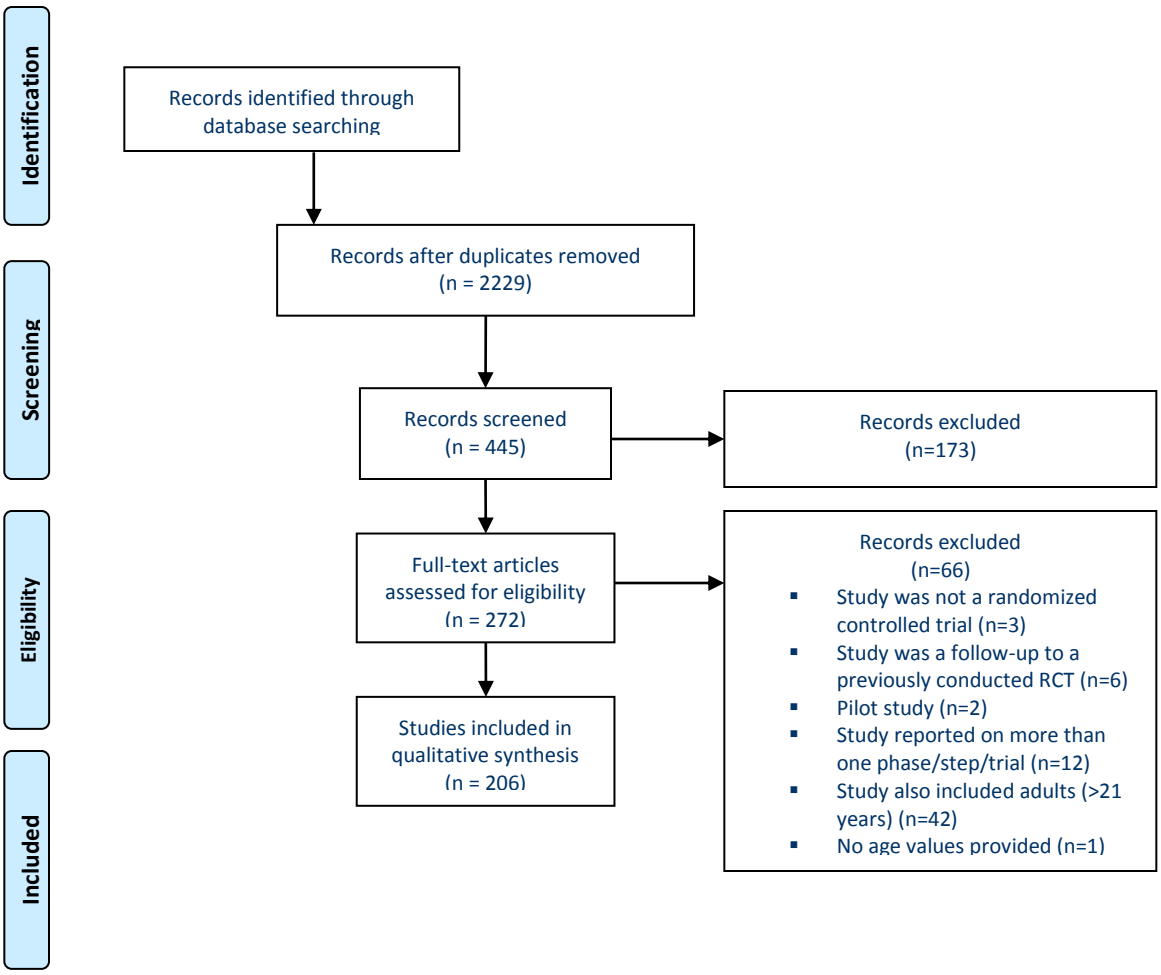

Figure 2-1 PRISMA[12] Flow Diagram of Search Results

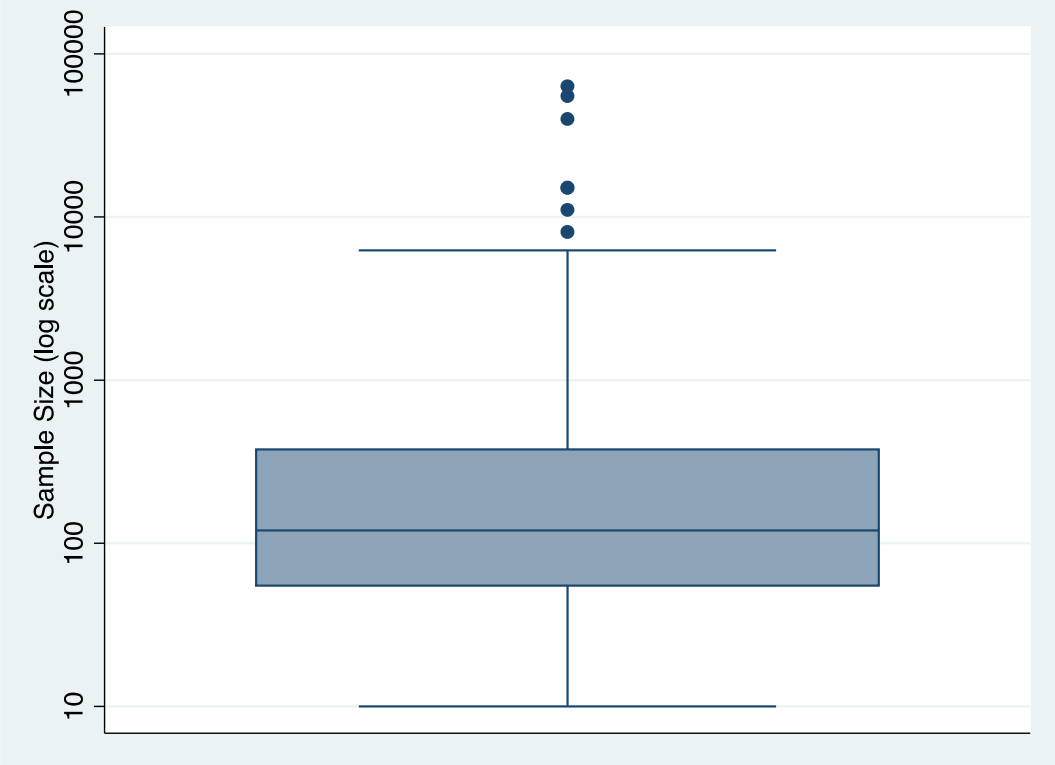

Figure 2-2 Distribution of Sample Sizes

Variable		Number of RCTs (n=206)
Publication Year	2000	18 (9%)
	2001	18 (9%)
	2002	16 (8%)
	2003	27 (13%)
	2004	19 (9%)
	2005	26 (13%)
	2006	23 (11%)
	2007	20 (10%)
	2008	17 (8%)
	2009	18 (9%)
	2010	4 (2%)
Journal	*Pediatrics*	65 (32%)
	Journal of Pediatrics	57 (28%)
	Archives of Pediatrics and Adolescent Medicine	20 (10%)
	Lancet	18 (9%)
	New England Journal of Medicine	18 (9%)
	American Academy of Child & Adolescent Psychiatry	15 (7%)
	Journal of the American Medical Association	9 (4%)
	PLoS Medicine	2 (1%)
	British Medical Journal	1 (0.5%)
	Annals of Internal Medicine	1 (0.5%)
Type of RCT	Parallel	183 (89%)
	Crossover	20 (10%)
	Factorial	3 (1%)
Type of trial	Treatment	134 (65%)
	Prevention	72 (35%)
Number of groups studied	Median	2
	Range	2-6
Sample Size Calculation	Y	131 (64%)
	N	75 (36%)
Sample Size	Median	120
	Range	10 – 63 225

Table 2-1 Summary of Included Studies

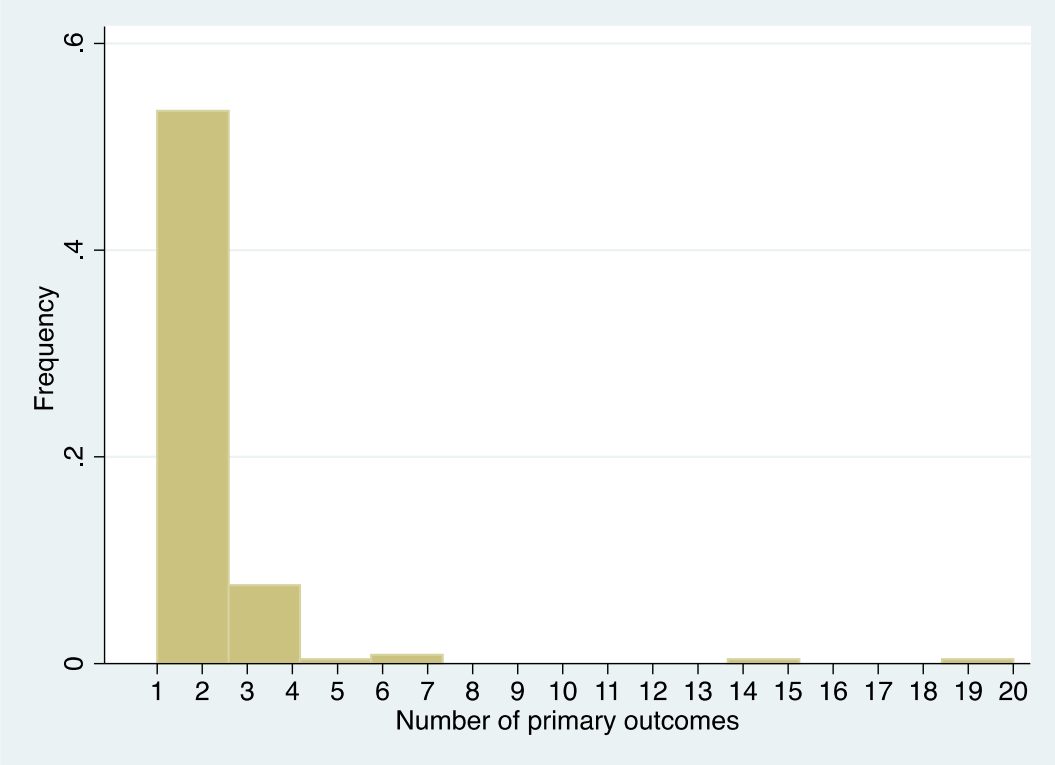

Figure 2-3 Primary Outcomes Reporting

Primary Outcome	Outcome Measure	Measurement Properties reported	Authors' Citations for Measurement Properties
Changes in the retractions and wheezing in acute bronchiolitis	Respiratory Disease Assessment Instrument (RDAI) – Respiratory Assessment Change Score[13]	Internal validity and responsiveness	10. Klassen T, Sutcliffe T, Watters L, Wells GA, Allen UD, Li MM. Dexamethasone in albuterol-treated inpatients with acute bronchiolitis: a randomized, controlled trial. J Pediatr 1997;130:191-7. 13. Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of albuterol in acute bronchiolitis. J Pediatr 1991;118:806-11. 14. Lowell DI, Lister G, Von Koss H, Mc-Carthy P. Wheezing in infants: the response to epinephrine. Pediatrics 1987; 79:939-45.
Proportion of treatment successes (i.e. need for enteral feeding in infants with resistance to feeding)	Infant Feeding Behaviours - Rater checklist (IFB – Rater checklist)[14]	Previously validated, √agreement between raters	1. Arts-Rodas D, Benoit D. Feeding problems in infancy and early childhood: identification and management. J Paediatr Child Health 1998; 3:21-7. 23. Benoit D, Green D. The Infant Feeding Behaviors - Rater Checklist: preliminary data. Poster presented at the Fortysecond Annual Meeting of the American Academy of Child and Adolescent Psychiatry, New Orleans, LA; 1995. 24. Koulis K, Arts-Rodas D, Benoit D. The Infant Feeding Behaviors - Rater checklist: comparison of coding methods. Poster presented at the forty-fourth Annual Meeting of the American Academy of Child and Adolescent Psychiatry,Toronto, Ontario; 1997.
Adequate clinical response defined by depressive symptoms	Children's Depression Rating Scale - Revised (CDRS-R)[15]	√inter-rater reliability, intra-class correlation	16. Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating ScaleYRevised. Psychopharmacol Bull. 1984;21:979Y989. 17. Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nd ed. Washington: US Government Printing Office; 1976.
Adequate clinical response defined by depressive symptoms Exacerbation rates in lithium treatment of acute mania	Clinical Global Impressions-Improvement Subscale (CGI-I)[15, 17]	√inter-rater reliability, intra-class correlation -	16. Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating ScaleYRevised. Psychopharmacol Bull. 1984;21:979Y989. 17. Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nded. Washington: US Government Printing Office; 1976. -
ADHD Symptoms	Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored (ADHDRS-IV-Teacher:Inv)[16]	validity	Faries DE, Yalcin I, Harder D, Heiligenstein JH (2001), Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 5:39–47

Exacerbation rates in lithium treatment of acute mania	Global Clinical Judgements (GCJ)[17]	-	(Campbell M, Small AM, Green WH et al. (1984), Behavioral efficacy of haloperidol and lithium carbonate: a comparison in hospitalized aggressive children with conduct disorder. *Arch Gen Psychiatry* 41:650–656 Campbell M, Adams P, Small AM et al. (1995), Lithium in hospitalized aggressive children with conduct disorder: a double-blind and placebocontrolled study. *J Am Acad Child Adolesc Psychiatry* 34:445–453 Malone RP, Delaney MA, Luebbert JF, Cater J, Campbell M (2000), A double-blind placebo-controlled study of lithium in hospitalized aggressive children and adolescents with conduct disorder. *Arch Gen Psychiatry* 57:649–654)
Pain induced by heel lance in newborns	Premature Infant Pain Profile (PIPP)[18]	validated, interrater reliability	17. Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *Clin J Pain.* 1999;15(4):297–303. 18. Jonsdottir RB, Kristjansdottir G. The sensitivity of the premature infant pain profile: PIPP to measure pain in hospitalized neonates. *J Eval Clin Pract.* 2005;11(6):598–605
Duration of acute viral upper respiratory tract infection	Canadian Acute Respiratory Illness and Flu Scale (CARIFS)[19]	-	(24 Jacobs B, Young NL, Dick PY, et al. Canadian Acute Respiratory Illness and Flu Scale (CARIFS): development of a valid measure for childhood respiratory infections. *J Clin Epidemiol* 2000; **53:**793–99.)
Gross motor function	Gross Motor Function Measure (GMFM)[20]	-	(12 Russell DJ, Rosenbaum PL, Cadman DT, Gowland C, Hardy S, Jarvis S. The Gross Motor Function Measure: a means to evaluate the effects of physical therapy. *Develop Med Child Neurol* 1989; **31:** 341–52. 13 Nordmark E, Hagglund G, Jarnlo GB. Reliability of the gross motor function measure in cerebral palsy. *Scand J Rehab Med* 1997; **29:** 25–28. 25 Trahan J, Malouin F. Changes in gross motor function measure in children with different types of cerebral palsy: an eight month follow-up study. *Pediatr Phys Ther* 1999; **11:** 12–17.)
Composite of death or severe neurodevelopmental disability[21] Composite of death, cerebral palsy, cognitive delay, deafness, or blindness[22]	Gross Motor Function Classification System (GMFCS) [21,22]		(20. Palisano RJ, Hanna SE, Rosenbaum PL, et al. Validation of a model of gross motor function for children with cerebral palsy. Phys Ther 2000;80:974-85.) (14. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997; 39:214-23.)
Composite of death or severe neurodevelopmental disability[21]	Mental Developmental Index of the Bayley Scales of Infant Development II	-	(22. Bayley N. Bayley scales of infant development. 2nd ed. San Antonio, TX: Psychological Corporation, 1993.) (15. Bayley N. Manual for the Bayley Scales of Infant Development. 2nd ed. San Antonio, TX:

Composite of death, cerebral palsy, cognitive delay, deafness, or blindness[22]	(BSID-II)[21, 22]		Psychological Corporation, 1993. 25. Hack M, Taylor G, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. Pediatrics 2005;116:333-41.)
Symptoms of obsessive-compulsive disorder (change in score from baseline)[23, 24]	Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)[23, 24]	-	(Scahill L, Riddle MA, McSwiggin-Hardin M et al. (1997), Children's Yale- Brown Obsessive Compulsive Scale: reliability and validity. *J Am Acad Child Adolesc Psychiatry* 36:844–852)
Severe deformational plagiocephaly	Oblique Diameter Difference Index (ODDI)[25]	-	(van Vlimmeren LA, Takken T, van Adrichem LN, van der Graaf Y, Helders PJ, Engelbert RH. Plagiocephalometry: a non-invasive method to quantify asymmetry of the skull; a reliability study. *Eur J Pediatr*. 2006;165(3):149-157.)
Difference in performance on tests assessing cognitive functions in children with Down syndrome	ǂCognitive Test Battery[26] - Stroop Color/Shape - Stroop Color/Word - Auditory Continuous Performance Task (ACPT) - Visual Continuous Performance Task - McCarthy Scales of Children's Abilities - Wide Range Assessment of Memory and Learning (WRAML) - Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) - Delayed match-to-sample - Match-to-sample - Go/No-go - Wide Range Assessment of Visuo-Motor Abilities (WRAVMA)		(18. Johnson CJ. Effects of color on children's naming of pictures.Percept Mot Skills. 1995;80:1091-1101. 19. Dalton AJ. Dementia in Down syndrome: methods of evaluation. In: Nadel L, Epstein CJ, eds. Down Syndrome and Alzheimer Disease. New York, NY: Wiley- Liss Inc; 1992:51-76. 20. McCarthy D. McCarthy Scales of Children's Abilities. New York, NY: Psychological Corp; 1972. 21. Sheslow D, Adams W. Wide Range Assessment of Memory and Learning. Wilmington, Del: Jastak Associates Inc; 1990. 22. Wechsler D. Manual for the Wechsler Preschool and Primary Scale of Intelligence. San Antonio, Tex: Psychological Corp; 1967. 23. Adams W, Sheslow D. Wide Range Assessment of Visuo-Motor Abilities. Wilmington, Del: Wide Range Inc; 1995.)
Change in individual test	evaluation tool designed by	(authors indicate it	-

scores to assess safety knowledge	authors[27]	has not been validated)	
Physical self-worth in obesity	Children and Youth Physical Self-Perception Profile (CY-PSPP)[28]	-	(20. Whitehead JR. A study of children's physical self-perceptions using an adapted physical self-perception profile questionnaire. *Pediatr Exerc Sci.* 1995;7:132–151 27. Biddle S, Page A, Ashford B, et al. Assessment of children's physical self-perceptions. *Int J Adolesc Youth.* 1993;4:93–109)
Anxiety of the child	Modified Yale Preoperative Anxiety Scale (m-YPAS)[29]	reliability and validity	9. Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstadter MB. The Yale Preoperative Anxiety Scale: how does it compare with a "gold standard?*" Anesth Analg.* 1997;85:783–788
Change in irritability from baseline	Aberrant Behaviour Checklist (ABC)[30]	-	(21. Aman MG, Singh NN, Stewart AW, Field CJ. The aberrant behaviour checklist: a behavior rating scale for the assessment of treatment effects.*Am J Ment Defic.* 1985;89:485–491 24. Aman MG, Singh NN. *Aberrant Behavior Checklist Manual*. East Aurora, NY: Slosson Educational Publications; 1986)
Neurobehavioral development	Neurobehavioural Assessment of the Preterm Infant (NAPI)[31]	√test-retest reliability, interrater reliability, clinical validity and sensitivity	(22. Korner AF, Kraemer HC, Reade EP, Forrest T, Dimiceli S, Thom VA. A methodological approach to developing an assessment procedure for testing the neurobehavioral maturity of preterm infants. *Child Dev.* 1987;58:1478–1487) 23. Korner AF, Constantinou J, Dimiceli S, Brown BW, Thom VA. Establishing the reliability and developmental validity of a neurobehavioral assessment for preterm infants: a methodological process. *Child Dev.* 1991;62:1200–1208 25. Korner AF, Stevenson DK, Kraemer HC, et al. Prediction of the development of low birth weight preterm infants by a new neonatal medical index. *Dev Behav Pediatr.* 1993;14:106–111

Table 2-2 Outcome Measures and Measurement Properties

ǂ battery test: comprises of 14 tests/domains selected from a variety of measures - treated as one outcome measure
( ) was not referred to by authors in text but included in the bibliographies
√ at least one of the measurement properties was examined as part of the study

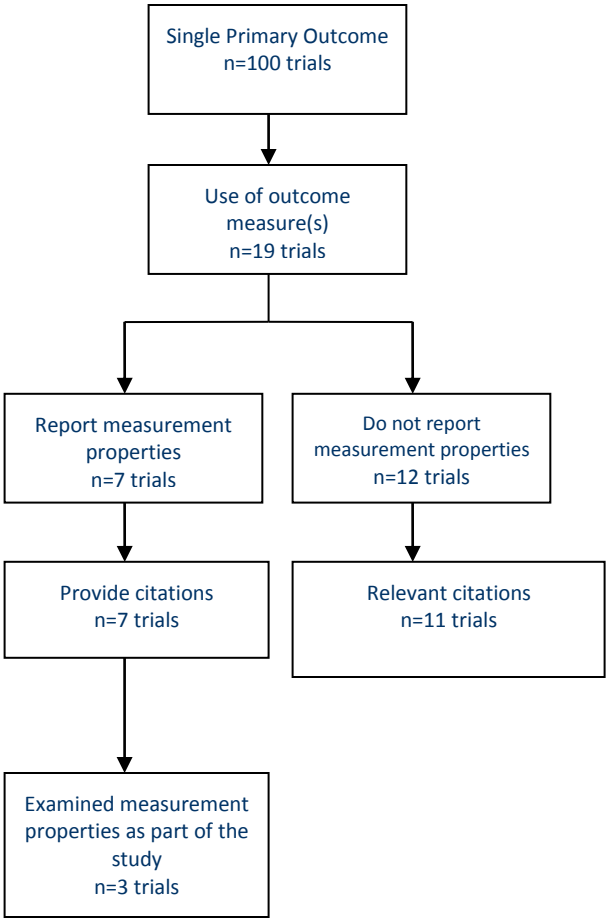

Figure 2-4 Flow Diagram of Assessment of RCTs

Appendices

Appendix 2-1: Search Strategies

Medline
1. 0890-8567.is.
2. "journal of the american academy of child & adolescent psychiatry".jn.
3. limit 2 to yr="2000 - 2010"
4. limit 3 to randomized controlled trial
5. from 4 keep 1-216
6. pediatrics.jn.
7. limit 6 to yr="2000 - 2010"
8. limit 7 to randomized controlled trial
9. from 8 keep 1-730
10. from 9 keep 1-730
11. "archives of pediatrics & adolescent medicine".jn.
12. limit 11 to yr="2000 - 2010"
13. limit 12 to randomized controlled trial
14. from 13 keep 1-213
15. "journal of pediatrics".jn.
16. limit 15 to yr="2000 - 2010"
17. limit 16 to randomized controlled trial
18. from 17 keep 1-342

1. "new england journal of medicine".jn.
2. limit 1 to yr="2000 - 2010"
3. limit 2 to randomized controlled trial
4. limit 3 to "all child (0 to 18 years)"
5. from 4 keep 1-272
6. jama.jn.
7. limit 6 to yr="2000 - 2010"
8. limit 7 to randomized controlled trial
9. limit 8 to "all child (0 to 18 years)"
10. from 9 keep 1-95
11. lancet.jn.
12. limit 11 to yr="2000 - 2010"
13. limit 12 to randomized controlled trial
14. limit 13 to "all child (0 to 18 years)"
15. from 14 keep 1-318
16. "annals of internal medicine".jn.
17. limit 16 to yr="2000 - 2010"
18. limit 17 to randomized controlled trial
19. limit 18 to "all child (0 to 18 years)"
20. from 19 keep 1-46
21. british medical journal.jn.

22. 0959-8146.is.
23. 1549-1277.is.
24. 1549-1277.il.
25. 0959-8146.il.
26. "plos medicine public library of science".jn.
27. limit 26 to yr="2000 - 2010"
28. limit 27 to randomized controlled trial
29. limit 28 to "all child (0 to 18 years)"
30. from 29 keep 1-16

Appendix 2-2: Inclusion/Exclusion Screening Form

1. Study Design

1.1 Was the study described as a randomized controlled trial (including parallel, cross-over, factorial, and N-of-1 designs)? Y/N *(If "no": EXCLUDE)*

1.2 Was this a diagnostic or screening trial? Y/N *(If "yes": EXCLUDE)*

1.3 Did this paper report on more than one phase/step/trial? Y/N *(If "yes": EXCLUDE)*

2. Study Population

2.1 Did the study include subjects less than 21 years old? Y/N *(If "no": EXCLUDE)*

2.2 Did the study also include subjects 21 years or older? Y/N *(If "yes": EXCLUDE)*

Final Decision
Was this study included? (Y/N/Unsure) *(If "Unsure": Provide Reason)*

If Disagreement between reviewers
Was this study included? Y/N

Appendix 2-3: Data Extraction Form

1. Publication

1.1 Journal name
_ *New England Journal of Medicine*
_ *Journal of the American Medical Association*
_ *Lancet*
_ *Annals of Internal Medicine*
_ *British Medical Journal*
_ *PLoS Medicine*
_ *Journal of the American Academy of Child and Adolescent Psychiatry*
_ *Pediatrics*
_ *Journal of Pediatrics*
_ *Archives of Pediatrics & Adolescent Medicine*

1.2 Publication year
_2000 _2001 _2002 _2003 _2004 _2005 _2006 _2007 _2008 _2009 _2010

2. Study design

2.1 What was the design of this RCT?
_ Parallel _crossover _ factorial _N-of-1 _other (specify) _____

2.2 Was the method of random allocation sequence generation specified? _Yes _No

2.3 What was the unit of random assignment? _Individual _cluster _other (specify) ___

2.4 If this study randomizes clusters, what is the cluster sample size?_____

2.5 Was this a treatment or prevention trial? (Note: studies of harms are allowed)
_Treatment trial _Prevention trial _Harm trial

2.6 How many groups were studied? _____

3. Population

3.1 What condition was studied? _____

3.2 Planned lower age reported? _____

3.3 Actual lower age reported? _____

3.4 Planned upper age reported? _____

3.5 Actual upper age reported? _____

3.6 If the actual age information is not clear, age as a summary score if provided?
_____ (Mean, SD, Median, IQR, other)

3.7 Was a sample size calculation reported? _Yes _No

3.8 What was the planned sample size? _____

3.9 What was the actual sample size? (i.e. number of participants randomized) _____

4. Intervention & Comparison

4.1 Was the intervention a conventional medical treatment (CMT) or a complementary and alternative medical treatment (CAM)? _____

4.2 What was the intervention compared to
_ Placebo/sham/inactive treatment
_ No treatment
_ Wait list
_ Other intervention
_ Other _____

5. Primary Outcomes
(Note: 'primary outcome' is synonymous with a variety of other terms)

5.1 Specification of at least one primary outcome (Y, N, Unclear) _____
(Specify terminology)

5.2 Was the primary outcome described as an outcome or a measure? _____

5.3 Number of primary outcomes identified? _____

(For multiple primary outcomes, stop here)

5.4 For studies with a single primary outcome

5.4.1 What was this outcome measuring? Specify broadly (pain, anxiety, safety) _____

5.4.2 Outcome measure used? _____

5.4.2.1 Psychometric properties of measure described
_Internal consistency
_Reliability/Measurement error (inter-rater reliability, test-retest reliability)
_Content validity
_Construct validity (structural validity, hypothesis testing, cross-cultural validity)
_Criterion validity

_Responsiveness
_Other _____

5.4.2.2. Were these properties examined as part of this study? _Yes_No

5.4.2.3 Was a relevant citation provided for the psychometric properties provided
[NOTE: relevant means psychometrics demonstrated for population under study in this publication]

References

[1] Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement Issues in Trials of Pediatric Acute Diarrheal Diseases: A Systematic Review. *Pediatrics* 2010; 126:e222-e231.

[2] 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(4): e96.

[3] Reid GT, Walter FM, Brisbane JM, Emery JD. Family History Questionnaires Designed for Clinical Use: A Systematic Review. *Public Health Genomics* 2009; 12:73-83.

[4] Zhang B, Schmidt B. Do we measure the right end points? A systematic review of primary outcomes in recent neonatal randomized clinical trials. *J Pediatr*. 2001;138(1): 76-80

[5] Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: Introduction. *J Rheum* 1993; 528–530.

[6] Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010;340:c365.

[7] Sinha IP, Williamson PR, Smyth RL. Outcomes in clinical trials of inhaled corticosteroids for children with asthma are narrowly focussed on short term disease activity. *PLoS One* 2009;4(7): 362-76.

[8] Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001; 285:1992-5.

[9] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010 Jul;63(7):737-45.

[10] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34-42.

[11] Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Ann Int Med 2010:152. Epub 24 March.

[12] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 2009; 6(7): e1000097.

[13] Schuh S, Coates AL, Binnie R, Allin T, Goia C, Corey M et al. Efficacy of oral dexamethasone in outpatients with acute bronchiolitis. *J Pediatr.* 2002;140(1):27-32.

[14] Benoit D, Wang EL, Zlotkin SH. Discontinuation of enterostomy tube feeding by behavioural treatment in early childhood: A randomized controlled trial. *J Pediatr.* 2000;137(4):498-503.

[15] Asarnow JR, Emslie Gr, Clarke G, Wagner KD, Spirito A, Vitiello B et al. Treatment of Selective Serotonin Reuptake Inhibitor-Resistant Depression in Adolescents: Predictors and Moderators of Treatment Response. *J Am. Acad. Child Adolesc. Psychiatry*. 2009;48(3):330-9.

[16] Weiss M, Tannock R, Kratochvil C, Dunn D, Velez-Borras J, Thomason C et al. A Randomized, Placebo-Controlled Study of Once-Daily Atomoxetine in the School Setting in Children with ADHD. *J Am. Acad. Child Adolesc. Psychiatry.* 2005;44(7):647-55.

[17] Kafantaris V, Coletti DJ, Dicker R, Padula G, Pleak RR, Alvir JMJ et al. Lithium Treatment of Acute Mania in Adolescents: A Placebo-Controlled Discontinuation Study. *J Am. Acad. Child Adolesc. Psychiatry.* 2004;43(8):984-93.

[18] Codipietro L, Ceccarelli M, Ponzone A. Breastfeeding or Oral Sucrose Solution in Term Neonates Receiving Heel Lance: A Randomized, Controlled Trial. *Pediatrics*. 2008;122(3):e716-21.

[19] Butler CC, Robling M, Prout H, Hood K, Kinnersley P. Management of suspected acute viral upper respiratory tract infection in children with intranasal sodium cromoglicate: a randomised controlled trial. *Lancet*. 2002;359:2153-8.

[20] Collet JP, Vanasse , Marois P, Amar M, Goldberg J, Lambert J et al. Hyperbaric oxygen for children with cerebral palsy: a randomised multicentre trial. *Lancet*. 2001;357:582-6.

[21] Azzopardi DV, Strohm B, Edwards AD, Dyet L, Halliday HL, Juszczak E et al. Moderate Hypothermia to Treat Perinatal Asphyxial Encephalopathy. *N Engl J Med*. 2009;361(14):1349-58.

[22] Schmidt B, Roberts RS, Davis P, Doyle LW, Barrington KJ, Ohlsson A et al. Long term effects of Caffeine Therapy for Apnea of Prematurity. *N Engl J Med.* 2009;357(19):1893-1902.

[23] Geller DA, Wagner KD, Emslie G, Murphy T, Carpenter D, Wetherhold E et al. Paroxetine Treatment in Children and Adolescents with Obsessive-Compulsive Disorder: A Randomized, Multicenter, Double-Blind, Placebo-Controlled Trial. *J Am. Acad. Child Adolesc. Psychiatry.* 2004;43(11):1387-96.

[24] Riddle M, Reeve E, Yaryura-Tobias J, Yang HM, Claghorn JL, Gaffney G et al. Fluvoxamine for Children and Adolescents with Obsessive-Compulsive Disorder: A Randomized, Controlled, Multicenter Trial. *J. Am. Acad. Child. Adolesc. Psychiatry.* 2001;40(2):222-9.

[25] van Vlimmeren LA, van der Graaf Y, Boere-Boonekamp M, L'Hoir MP, Helders PJM, Engelbert RHH. Effect of Pediatric Physical Therapy on Deformational Plagiocephaly in Children with Positional Preference. *Arch Pediatr Adolesc Med*. 2008;162(8):712-8.

[26] Lobaugh NJ, Karaskov V, Rombough V, Rovet J, Bryson S, Greenbaum R et al. Piracetam Therapy Does not Enhance Cognitive Functioning in Children with Down Syndrome. *Arch Pediatr. Adolesc. Med.* 2001;155:442-8.

[27] Luria JW, Smith GA, Chapman JI. An Evaluation of a Safety Education Program for Kindergarten and Elementary School Children. *Arch Pediatr Adolesc. Med.* 2000;154:227-31.

[28] Daley AJ, Copeland RJ, Wright NP, Roalfe A, Wales JKH. Exercise Therapy as a Treatment for Psychopathologic Conditions in Obese and Morbidly Obese Adolescents: A Randomized, Controlled Trial. *Pediatrics*. 2006;118:2126-34.

[29] Vagnoli L, Caprilli S, Robiglio A, Messeri A. Clown Doctors as a Treatment for Preoperative Anxiety in Children: A Randomized, Prospective Study. *Pediatrics*. 2005;116:e563-7.

[30] Shea S, Turgay A, Carroll A, Schulz M, Orlik H, Smith I et al. Risperidone in the Treatment of Disruptive Behavioral Symptoms in Children with Autistic and Other Pervasive Developmental Disorders. *Pediatrics.* 2004;114:e634-41.

[31] Johnston CC, Filion F, Snider L, Majnemer An, Limperopoulos C, Walker C-D et al. Routine Sucrose Analgesia During the First Week of Life in Neonates Younger than 31 Weeks' Postconceptional Age. *Pediatrics.* 2002;110(3):523-8.

[32] Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001;285:1992-5.

[33] Plint AC, Moher D, Schulz K, Altman DG, Morrison A. Does the CONSORT checklist improve the quality of reports of randomized controlled trials? A

systematic review. *Firth International Congress of Peer Review and Biomedical Publication*, September 16-18 2005.

Chapter 3
Reporting of Measurement Properties of Primary Outcome Measures in Pediatric Randomized Controlled Trials (RCTs)

Zafira Bhaloo[1,4], Lisa Hartling[2], Caroline B Terwee[3], Sunita Vohra[1,4]

[1]CARE Program, Edmonton Continuing Care Center
Unit 8B, 11111 Jasper Ave
Edmonton AB T5K 0L4
[2]Alberta Research Center for Health Evidence, Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta
Edmonton Clinic Health Academy
11405-87 Avenue
Edmonton, AB T6T 1C9
[3]VU University Medical Center, Knowledgecenter Measurement Instruments, Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
van der Boechorststraat 7
1081 BT Amsterdam
[4]Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta
Edmonton Clinic Health Academy
11405-87 Avenue

Correspondence to: S Vohra svohra@ualberta.ca

BACKGROUND

Randomized controlled trials (RCTs) are the internationally accepted gold standard for high quality evidence about treatment efficacy and effectiveness. RCT results are key determinants for service providers, researchers, policy makers and consumers in the evaluation of health care interventions. Outcome measures are tools used in RCTs that measure the primary outcome and thus generate results for evidence-based practice. The validity of the trial results is therefore a direct reflection of the validity of the outcome measures used.

When evaluating or selecting a particular outcome measure, the measurement properties are reviewed and compared. These properties should be considered in the context of the population, design and setting in which they were applied. There are three domains of measurement properties: reliability, validity, and responsiveness[1]. The reliability domain contains internal consistency, measurement error, and reliability. The validity domain includes content validity, construct validity and criterion validity. The responsiveness domain contains the responsiveness measurement property. Reporting these properties for any given measure is crucial to better understand the measure's intended purpose and whether it can discern change when it occurs. Furthermore, the properties provide insight into the measure's performance and accuracy.

Systematic reviews conducted in specified clinical areas have found that a limited number of studies state whether the outcome measures used are valid and reliable and the few that do, fail to report evidence for the measurement properties with citations[2,3]. Furthermore, outcome measures may be abundant in clinical use despite their lack of formal evaluation or validation against reference standards[4]. If measurement properties are not reported or evaluated, the quality of the measure and its appropriateness within a particular study remains unclear. If the outcome measure used within a trial is not valid or responsive this can lead to a significant risk of bias in the trial results. The interpretation of the study results is then compromised which consequently impedes knowledge

synthesis efforts and evidence-based practice. The lack of information on measurement properties significantly limits evidence-based outcome measure selection and application.

In our systematic review of primary outcomes reporting, we examined pediatric RCTs published in high impact journals and identified a small sample of seven RCTs that reported measurement properties of their primary outcome measures with relevant citations for these measurement properties. Here, we examine in-depth the eight outcome measures used within the small sample of RCTs identified, to: (1) identify studies on measurement properties of the primary outcome measures to confirm the accuracy of reporting in the pediatric RCTs, (2) assess the methodological quality of the studies on measurement properties to substantiate any conclusions about the quality of outcome measures, and (3) critically appraise the measurement properties to assess the quality of the outcome measures.

METHODS

Search Strategy

With the help of an experienced research librarian, we searched the MEDLINE electronic database using Terwee et al.'s methodological PubMed search filter[5] in order to identify all studies reporting on the measurement properties of the outcome measures in the pediatric population. This highly sensitive search filter consists of a combination of search terms designed to retrieve studies on measurement properties of outcome measures. The search filter was developed according to four phases: (1) the identification of a gold standard in which to evaluate the filter (PubMed records), (2) the selection of search terms and their subsequent combination based on their sensitivity and precision, (3) the evaluation of the search filter against the gold standard set of PubMed records (internal validity), and (4) the validation of the filter against two existing PubMed searches (external validity)[5]. The search filter is designed to be used in combination with search terms for the construct of interest, the outcome measure of interest, and the population of interest. The search terms for each of these aspects were defined with the

help of an experienced research librarian. To ensure maximum retrieval of studies, all possible names for the outcome measure, including acronyms, were used as search terms (Appendix 3-1).

Study Selection

Titles and abstracts of retrieved studies were screened and the full-text was retrieved if the original version of the outcome measure was discussed and the study was published in English. Full texts were then examined for information on measurement properties. Studies discussing translated or cross-cultural versions of outcome measures were not included as the measurement properties of original versions should be well established prior to translating or adapting for a different population. Reference lists of included studies were also searched to identify any additional relevant studies on the outcome measures and their measurement properties. These additional studies were retrieved if they had not already been included by the search filter. The studies retrieved were compared to the original citations provided in the identified sample of pediatric RCTs to determine the accuracy of reporting of measurement properties. All originally cited studies were also retrieved for evaluation if they had not been included in the search strategy or the additional reference list search.

Quality Assessment

Methodological Quality of Studies on Measurement Properties

The methodological quality of the studies on measurement properties was assessed using the COSMIN checklist with its 4-point scoring system[6,7]. The COSMIN checklist was developed based on consensus of international experts during a four-round Delphi study on definitions of measurement properties[6]. The checklist presents standards for design and statistical methodology and contains a box for each of the nine measurement properties. The subsequent scoring system was developed to provide an overall methodological quality score for each study on a measurement property[7]. In order to

assess the methodological quality of the studies, three steps are required. The first step determines which measurement properties were evaluated in the study and uses COSMIN taxonomy and definitions[1]. Next, the corresponding COSMIN measurement property boxes are marked. Each item in a box is scored on a 4-point rating scale: excellent, good, fair, or poor. The overall methodological quality score of the measurement property box is obtained by taking the lowest rating of any item in the box[7].

An "excellent" rating is given when adequate evidence is provided regarding the methodological quality aspect of the study. A "good" rating is given when relevant information is missing but the quality aspect can be assumed to be adequate. A "fair" rating is given when it is not clear whether the methodological quality is adequate. A "poor" rating is given when there is evidence that the methodological quality aspect is not adequate. Details on the 4-point rating scale are provided on www.cosmin.nl[10].

Quality of Measurement Properties

The quality of the outcome measures' measurement properties was evaluated in the studies that discussed development or evaluation of the measure and thus assessed its measurement properties. For studies that simply cited measurement properties, the citations provided were retrieved and if measurement properties were assessed within these studies, they were then evaluated.

The measurement properties were evaluated using a modified version of the Terwee quality criteria and are divided over three domains: (1) reliability (internal consistency, measurement error, reliability), (2) validity (content, construct, criterion), and (3) responsiveness (responsiveness)[1, 8, 9]. The Terwee quality criteria were originally developed for the design, methods, and outcomes of studies to enable meaningful comparison and selection of outcome measures[6]. Nine measurement properties are distinguished and the possible overall ratings for each measurement property are: positive (+), indeterminate (?), negative (-), or "no information available" (0) (Appendix 3-2).

RESULTS

The search strategy retrieved a total of 475 unique studies for all eight outcome measures, of which 311 had their full texts reviewed. After evaluating full texts and their reference lists for additional relevant studies, a total of 70 studies assessing measurement properties of the outcome measures were included. Table 3-1 presents the studies included for each outcome measure along with a comparison to the originally cited studies in the pediatric RCTs. The methodological quality of the studies for each outcome measure is presented in Table 3-2. Table 3-3 presents the assessment of the measurement properties using Terwee quality criteria.

Respiratory Disease Assessment Instrument (RDAI) – Respiratory Assessment Change Score (RACS)

The RDAI measures respiratory rate, wheezing, and retractions through clinician administration on an ordinal scale and was originally developed for children less than 24 months of age[21]. The highly sensitive Terwee search filter retrieved two studies on measurement properties, one of which was the RCT from which this measure was identified (Table 3-1). This RCT cited three studies, only one of which was retrieved by the filter. The other two did not mention any measurement properties in their abstracts and were therefore not identified by the search filter[20, 21].

Methodological Quality

None of the identified studies evaluated the RDAI's internal consistency, measurement error, construct validity, criterion validity or floor and ceiling effects. All four studies assessed were found to be of poor methodological quality due to small sample sizes (<30)[19-22], lack of assessments regarding the comprehensiveness of the measure and the relevance to the study's population[21], and lack of longitudinal design and comparator instrument definition in assessing responsiveness[21].

Quality Assessment

All four studies found strong inter-rater reliability however one study had poor inter-rater reliability for the retractions factor[21]. Content validity was assessed in one study however the completeness of the measure was not assessed or discussed. The three variables were chosen based on the frequency of their use amongst clinicians however whether the three variables together comprehensively reflect the construct measured was not discussed[21]. The same study assessed responsiveness by correlating the RDAI with "a number of respiratory and related variables" with consistent change in both however, no numerical correlations are provided and no justification for the choice of variables correlated is explained[21]. Three of the studies provide information relating to interpretability however all three provide mean and standard deviation scores for only two subgroups and do not define a minimal important change (MIC)[19, 20, 22].

Infant Feeding Behaviours – Rater Checklist (IFB – Rater Checklist)

No studies were retrieved from the search strategy for the IFB – Rater checklist. In the original pediatric RCT assessed[12], three citations were provided for the IFB – Rater checklist. The first study used the IFB – Parent Checklist and therefore was not further assessed[89]. The two other citations were poster presentations and could not be retrieved[90, 91].

Children's Depression Rating Scale – Revised (CDRS-R)

The CDRS-R assesses the severity of depression using 17 items on ordinal scales, scored through clinician conducted interviews of the child and/or parent, and was developed for children ages six to 12[40]. The highly sensitive Terwee search filter identified 18 relevant studies assessing measurement properties of the CDRS-R (Table 3-1). The one study cited in the original RCT was not identified by the search strategy as it does not mention measurement properties in its abstract but we retrieved it for further assessment[40].

Methodological Quality

All the measurement properties of the CDRS-R were assessed among the studies examined. The majority of the studies assessing the internal consistency of the CDRS-R were found to be of poor methodological quality as internal consistency was not calculated for each subscale[23-25, 30, 34, 39, 40] and factor analysis was not performed to check the unidimensionality of the scale[34, 39, 40]. Two studies were rated as good as sample sizes and unidimensionality analyses were appropriate [27, 29]. The one study that assessed measurement error had fair methodological quality due to methodological flaws regarding missing items, time interval between measurements, and stability of patients[27]. Studies assessing the reliability of the CDRS-R were of either good or fair quality (Table 3-2). Five studies either did not discuss missing items or the model or formula of the Intra-class Correlation Coefficient (ICC) calculated and therefore received good ratings[28, 31, 33, 38, 39]. The other studies assessing reliability were of fair quality as they had moderate sample sizes (30-49)[24, 36, 37, 40]. Although the risk of missing items is low as the CDRS-R is clinician-administered, the possibility still exists and therefore should be discussed. For the study assessing test-retest reliability, it was unclear whether patients and test conditions were stable during the time interval[24.]

Content validity was assessed in two studies and both were methodologically poor (Table 3-2). One study lacked critical assessments including whether all the CDRS-R items refer to relevant aspects of depression, if the items were relevant for the study population, and if the items as a whole comprehensively reflect depression[24]. The other study also failed to assess the items all together to ensure they reflect the construct to be measured[40]. The methodological quality of the four studies assessing structural validity was good as it was only unclear how many missing items were present and how these were handled[24, 25, 27, 29]. Hypotheses' testing was assessed in six studies and the majority (four) were of poor methodological quality due to small sample size (<30)[23], lack of information and description of the comparator instrument for convergent validity[34, 40], and inappropriate statistical methods (sensitivity percentages)[37]. The other two studies were determined to have fair methodological quality due to lack of clearly formulated hypotheses or minimal information provided about the comparator instrument[24, 38]. The one study assessing criterion validity had fair methodological quality as it was unclear whether the criterion

used (ICD-10) can be considered as the gold standard as no measurement properties or convincing arguments were provided[24].

All the studies assessing responsiveness were found to be of poor methodological quality due to the lack of a longitudinal design[24, 27] and clearly stated time interval[24], inappropriate statistical methods without the use of prespecified hypotheses (effect sizes)[30, 33, 36], and lack of information about the comparator instrument[34, 36].

Quality Assessment

Nine studies examined the internal consistency of the measure, eight reported good internal consistency with Cronbach's alphas of 0.70 or higher[23-25, 27, 29, 30, 34, 39] and one study received an indeterminate rating as Cronbach's alpha was not determined but item-total correlations were provided[40]. One study briefly mentioned the use of standard errors of measurement (SEM), an expression of the measurement error, however no calculation or presentation of SEMs was provided and no MIC was defined, thus earning an indeterminate rating[27]. Eleven studies assessed the inter-rater reliability of the CDRS-R and all of the studies found strong inter-rater reliability with intraclass correlation coefficients of 0.73[24] and higher. One of the studies also assessed test-retest reliability over four weeks and found a high intraclass correlation of 0.98[24].

Content validity was evaluated in two studies; however both studies received indeterminate ratings. The first study states that "all the major dimensions of depressions were represented in the CDRS-R" but does not provide any evidence that supports this claim[24]. The authors conclude that since less than half of the participants gave a score of 0 to any of the 17 items, "the content validity was appropriate to their morbid state"[24]. According to Terwee's quality criteria, the relevance of the items in the questionnaire must be assessed along with the comprehensiveness of the questionnaire[6]. The second study discusses scale construction and briefly mentions that items were dropped as they showed no correlation with depression but the authors do not provide any further information or evidence to substantiate the comprehensiveness of the measure[40]. Construct validity was assessed in four studies with structural validity, and in six studies

with hypotheses testing. The CDRS-R received positive, negative and indeterminate ratings for structural validity across the four studies. The first study received a positive rating as it analyzed factor structure and found that a six-factor structure accounted for 60.6% of variance[24]. Two studies received negative ratings as the factors explained less than 50% of the variance[25, 29]. Lastly, the fourth study examined factor loadings but explained variance was not mentioned thus earning an indeterminate rating[27]. Of the six studies assessing hypothesis testing, three received negative ratings (Table 3-3). The first study examined correlations with activity parameters and found all correlations to be less than 0.5[23]. The second study compared the CDRS-R scores with another similar measure but provided correlations only in terms of sensitivity percentages[37]. The third study correlated the CDRS-R with four other measures measuring similar constructs but found all correlations to be less than 0.5[38]. All three studies that received positive ratings tested correlations between the CDRS-R and other measures and found correlations greater than 0.5 with related measures and lower correlations with unrelated measures[24, 34, 40]. Measures compared with the CDRS-R included the Beck Depression Inventory (0.71) [24], Impact of Events scale (0.28) [24], Kiddie Schedule for Affective Disoders and Schizophrenia (0.64) [34], Clinical Global Impressions – Severity (0.87) [34], Children's Global Assessment Scale (-0.77) [34], and global ratings of clinical depression from an independent source[40]. One study assessed criterion validity using the ICD-10 as the reference standard and presented concordance rate as a percentage, however convincing information was not provided that this measure should be used as the reference standard[24].

Responsiveness assessments met acceptable criteria in two studies. In one study, the CDRS-R's sensitivity and specificity values were compared to a reference standard (ICD-10) and the area under the receiver operator characteristic curve (AUC) was 0.87[24]. The second study assessed correlations between CDRS-R and CGI-I change scores and found strong correlations (0.82 and higher) where expected[34]. Another study also provided the AUC but it was less than 0.70 (0.54)[27]. The three other studies received indeterminate ratings due to their doubtful design and methods. Three of the studies used effect sizes,

which are considered inappropriate measures of responsiveness without prespecified hypotheses, as per the COSMIN checklist guidelines[10].

Floor and ceiling effects were discussed in two studies (Table 3-3). The first study mentioned that none of the 17 items of the CDRS-R was assigned a 0, the lowest possible score, by more than half of the adolescents[24]. However, the percentage of adolescents that did score 0 remains unclear and may be greater than 15%. In the second study, more than 15% of participants scored the lowest or highest possible scores for many items. These include 42.4% scoring the lowest score to 41.4% scoring the highest possible score[29].

Interpretability was assessed in 11 studies and all the studies received an indeterminate rating. The majority of the studies presented mean and standard deviation (SD) scores for less than four relevant subgroups and did not define a MIC[23, 30, 31, 34-38]. The studies that did provide mean scores for at least four subgroups did not define a MIC[26, 32, 33].

Clinical Global Impressions – Improvement Subscale (CGI-I)

The Clinical Global Impressions – Improvement Subscale records changes in behaviour of children older than five years, based on clinican ratings on a 7-point scale, from 1 (very much improved) to 7 (very much worsened)[43]. The highly sensitive search filter retrieved four relevant studies in which measurement properties were assessed (Table 3-1). The reference originally cited in the pediatric RCT was not retrieved by the filter as it is the outcome measure's manual and the filter retrieves studies assessing measurement properties i.e. validation or comparative studies[92]. We retrieved the manual as it was consistently cited amongst all the relevant studies however it did not assess measurement properties of the outcome measure and therefore was not included in subsequent analyses[92].

Methodological Quality

Amongst the four relevant studies found, internal consistency, reliability and interpretability were assessed (Table 3-3). The study assessing internal consistency had

poor methodological quality as factor analysis was not performed or referenced to determine the unidimensionality of the scale[44]. Three of the four studies assessing reliability were of poor methodological quality (Table 3-2). Two of the studies had a small sample size (<30)[41, 43] while in the other, only percentage agreement was calculated[42]. One study did receive a good methodological quality rating for reliability as a good sample size was used (50-99) and only a description of the weighting scheme for the weighted kappa was missing[44].

Quality Assessment

The study assessing internal consistency found strong Cronbach alpha scores (0.80)[44]. Reliability of the CGI-I received positive, negative and indeterminate ratings (Table 3-3). The two positively rated studies provided adequate inter-rater reliability statistics (kappa of 0.71 or higher)[41, 43]. One study assessed inter-rater reliability but provided percentage agreements[42] which is not considered adequate according to the Terwee quality criteria[6]. The last study assessing reliability had inter- and intra-rater reliability coefficients less than 0.70 (0.37 and 0.55) and thus received a negative rating[44]. One study described interpretability but received an indeterminate rating as the mean scores were provided for only three relevant subgroups with no MIC defined[41].

Attention Deficit/Hyperactivity Disorder Rating Scale-IV: Teacher Version: Investigator administered and scored (ADHDRS-IV-Teacher:Inv)

The ADHDRS-IV-Teacher:Inv is an 18 item questionnaire that contains diagnostic criteria from the *Diagnostic and Statistical Manual of Mental Disorders*. The measure is designed to provide information regarding certain behaviours symptomatic of attention deficit/hyperactivity disorder in children from kindergarten to grade 12[45, 46]. The search filter did not identify any relevant studies that assess measurement properties of this outcome measure however studies retrieved cited two references that were subsequently retrieved[45, 46], in addition to the originally cited reference in the pediatric RCT[93]. None of the three studies retrieved were included in the search strategy results as they did not mention measurement properties and the desired measure's name in the abstract and/or

title. Of note, the reference originally cited in the pediatric RCT was for the clinician-rated version of the measure and not the teacher version and was therefore not subsequently analyzed[93].

Methodological Quality

In the two relevant studies included, internal consistency, reliability, construct validity, criterion validity, and interpretability were assessed (Table 3-3). The internal consistency assessment was found to have poor methodological quality as the sample size for the unidimensionality analysis was inadequate as per COSMIN criteria[46]. With regards to reliability, the methodological quality was assessed as fair because for the test-retest reliability, it was doubtful whether the measurements were independent, whether patients remained stable during the four week time interval, whether test conditions were similar, and the Pearson correlation coefficients were provided without evidence that systematic change had not occured[46].

The methodological quality of the study assessing structural validity was good as the only information missing was with regards to the percentage of missing items and how these were handled[45]. The study assessing hypotheses testing was rated with fair methodological quality because while it was possible to presume what was expected, the hypotheses were not formulated and the comparator measure was poorly described[46]. Lastly, the methodological quality of the study with regards to criterion validity was also rated as fair as it was unclear whether the Conners Teaching Rating Scale-39 can be considered an adequate gold standard [45].

Quality Assessment

The study assessing internal consistency was positively rated as coefficients provided for the total and subscale scores were strong (0.94, 0.96 and 0.88)[46]. The same study assessed reliability and again found strong coefficients for the test-retest reliability (0.88-0.90)[46].

Structural validity was assessed with both exploratory and confirmatory factor analyses; one-factor structure explained 64.8% of variance while two-factor structure explained

71.9% of variance[45]. Hypotheses' testing was assessed in one study and it was rated positively as 75% of the results accorded with all the hypotheses, in this case the differentiation between clinical and control groups[46]. The Conners Teaching Rating Scale-39 was used to assess criterion validity however; the authors do not provide any convincing arguments that this measure can be used as a gold standard[46]. Furthermore, the correlations themselves are not presented, it is simply stated that 20 of 30 correlations were statistically significant[46]. Interpretability could be assessed in one of the studies with means and standard deviations provided for four relevant subgroups however, no MIC was defined thus resulting in an indeterminate rating[45].

Premature Infant Pain Profile (PIPP)

The Premature Infant Pain Profile assesses acute pain in preterm and term neonates based on clinician rating of seven indicators[50]. The highly sensitive search filter retrieved 28 relevant studies assessing the PIPP's measurement properties, including the pediatric RCT originally assessed in our systematic review[56] (Table 3-1). Both citations initially referenced by the pediatric RCT were retrieved by the filter. One of them, however, is an Icelandic version of the PIPP and was therefore not included in further assessment as per the inclusion criteria previously described.

Methodological Quality

All of the measurement properties except criterion validity and responsiveness were assessed by the studies included. The sole study assessing internal consistency was rated as methodologically good as information was not provided on missing items[74]. The study assessing measurement error was rated as poor due to small sample size (<30) and the availability of only one measurement[72]. Twenty-three of the 26 studies assessing reliability were of poor methodological quality (Table 3-3). These studies calculated percentage agreements[48,49, 52, 55-70] , reported intraclass correlation coefficients when kappas would be more appropriate due to the ordinal scales assessed[50, 53, 55-58, 61, 64-71], or had inadequate sample sizes (<30)[51, 53, 55, 68, 71]. Three studies were of fair methodological

quality as an unweighted kappa was calculated[54, 74] or a moderate sample size was used (30-49)[73].

The study assessing content validity was the only study to receive an excellent methodological rating as it met all required criteria including assessment of whether items refer to relevant aspects of the construct measured, whether they comprehensively reflect this construct and whether they are relevant for the study population and the purpose for which they were applied[74]. The methodological quality of the structural validity assessment received a good rating as information on missing items was not provided[74]. Four studies assessing hypotheses testing received fair methodological quality ratings, and four received poor ratings (Table 3-3). The first four studies did not have explicitly formulated hypotheses[47, 66, 67, 70], or minimal information on measurement properties of the comparator instrument was provided[47, 70]. The other four studies either did not provide any information on measurement properties of the comparator instrument[48] or the constructs measured by the comparator instrument[69] or used inappropriate statistical methods to test hypotheses (p values)[50, 74].

Quality Assessment

One study assessed internal consistency and was given a positive rating as Cronbach's alphas were provided for each individual indicator and all were greater than 0.7[74]. Item total correlations were also provided[74]. The study assessing measurement error received an indeterminate rating as although limits of agreements (LoA) were provided for both inter- and intra-rater reliability, no MIC was defined[72]. Twenty-six of the 28 studies assessed reliability (Table 3-3). Twenty-one studies received a positive rating for reliability as inter- and/or intra-rater reliability was greater than 0.80 in all of them[50, 53, 54, 56-71, 73, 74] and was even as high as 0.99 in one study[71]. Three studies were given negative ratings as their inter-rater reliability was less than 0.70[51, 55, 73]. Three other studies received indeterminate ratings as they presented reliability statistics in percentage agreements[48, 49, 52].

One study assessed content validity and clinical and research experts reviewed the measure to ensure its relevance and comprehensiveness[74]. The same study also assessed structural validity and iterated principal component analysis found that a three factor structure explained 78.3% of the variance[74]. Hypotheses testing received a negative rating in three studies (Table 3-3). In the first study, there was low correlation (less than 0.50) between the PIPP and 2 other similar measures[47]. The second study found low correlation with the Neonatal Infant Pain Scale (NIPS)[67]. The third study also found low correlation with the NIPS as well as the Visual Analog Scale[69]. Five studies were given indeterminate ratings as the PIPP was correlated with other measures for the same construct however correlations with unrelated constructs were not provided[48, 50, 66, 70, 74]. Additionally, two of these studies did not correlate the PIPP with another measure but rather compared PIPP scores between extreme situations[50, 74] . One study found low correlation between the PIPP and the CRIES (crying, requires oxygen, increased vital signs, expression, and sleepless) measure[66].

Floor and ceiling effects were discussed in two studies and both received negative ratings (Table 3-3). In one study 20% of infants received the lowest possible score[61] and in the other, 35% of infants were assigned a score of 0[72]. Interpretability was assessed in 16 studies but each one earned an indeterminate rating as the majority of them provided mean and SD scores of less than four subgroups[47, 49, 50, 53, 62, 63, 65, 68-70, 72, 73] or failed to define a MIC[52, 54, 67, 74].

Modified Yale Preoperative Anxiety (m-YPAS)

The m-YPAS was developed for children aged two to six years, to assess anxiety, as measured by an observer, in the preoperative holding area as well as during the induction of anesthesia[77]. Seven relevant studies assessing the m-YPAS's measurement properties were retrieved by the search filter (Table 3-1), including the study originally cited by the pediatric RCT[77].

Methodological Quality

The internal consistency, reliability, content validity, construct validity, criterion validity, floor and ceiling effects and interpretability were assessed amongst the included studies examined.

The study assessing internal consistency, reliability and hypotheses testing was rated as methodologically poor in all assessments (Table 3-2). For internal consistency, factor analysis was not performed or referenced in the determination of the unidimensionality of the scale[75]. For reliability, an inadequate sample size was used (<30)[75] and for hypotheses testing, no information on the measurement properties of the Emotionality, Activity, Sociability, and Impulsivity (EASI) was provided[75].

Six of the seven studies assessing reliability were methodologically poor (Table 3-2). These studies used inadequate sample sizes (<30)[75, 76, 79], provided percentage agreements[76, 78], or provided ICCs when kappas were more appropriate due to the ordinal data[79, 80, 81]. One study received a good methodological quality rating due to a good sample size (50-99) and information lacking for the weighting scheme of the kappa calculated[77].

The study reporting content validity was methodologically poor as it was not assessed if all the m-YPAS items together reflect the anxiety of the child[77]. Both studies assessing construct validity through hypotheses testing were rated as methodologically poor as no information was provided for the EASI's measurement properties in one study[75] and p values were used to test the hypotheses in the other[77]. Lastly, the study evaluating criterion validity of the m-YPAS was determined to have good methodological quality as a good sample size was used (50-99), and it is assumable that the criterion can be considered an adequate "gold standard"[77].

Quality Assessment

The study assessing internal consistency provided Cronbach's alphas across all four categories of anxious behaviours and all met acceptable criteria[75]. Seven studies evaluated reliability of the measure and the majority received a positive rating (Table 3-

3). Inter- and intra-rater reliabilities met acceptable criteria and in some studies, the reliabilities were excellent (0.90-1.00)[75, 79, 80, 81]. Two studies received indeterminate ratings for their reliabilities as percentage of agreements were provided rather than coefficients[76, 78].

Content validity was assessed in one study that discussed scale development using the expertise of anaesthesiologists and psychologists however, the items were found to only reflect "most" of the behaviours observed[77]. Although some items were modified to include new behaviours observed, whether the modified measure can be considered a complete assessment remains unclear[77]. Construct validity was assessed in terms of hypotheses testing in two studies (Table 3-3). The first study received a negative rating as the correlations between the m-YPAS and the EASI scale were not strong (0.10-0.22) which did not accord with the hypotheses[75]. The second study did not correlate the m-YPAS with another instrument but rather compared scores at three time points[77]. Since no hypotheses were formulated as to what the expected differences between these time points were, it could not be determined if results accorded with expectations thus, the construct validity was rated as indeterminate[77]. One study assessed criterion validity by comparing the m-YPAS with the current gold standard State-Trait Anxiety Inventory for Children (STAIC) hypothesizing that the two would strongly correlate (>0.65)[77]. The authors provided citations to support their explanations of the STAIC as an appropriate gold standard and the correlation observed was strong (0.79) as hypothesized[77].

One study found that 57.4% of respondents scored at the lowest possible end of the m-YPAS and therefore received a negative rating for floor effects[76]. Two studies assessed responsiveness and both were indeterminately rated as scores for only two subgroups were provided and no MIC was defined[76, 79].

Neurobehavioural Assessment of the Preterm Infant (NAPI)

The NAPI provides an observer assessment of a newborn's competence in seven functional domains including motor development and vigour, alertness and orientation,

and irritability[82]. The Terwee highly sensitive search filter retrieved seven relevant studies assessing the NAPI's measurement properties, including one of the studies originally cited in the pediatric RCT (Table 3-1). The two other studies cited were not retrieved by the filter and we also could not retrieve the full texts of these studies.

Methodological Quality

All measurement properties were assessed amongst the studies examined with the exception of measurement error, content validity, structural validity and floor and ceiling effects (Table 3-3).

The study assessing internal consistency was found to have poor methodological quality as factor analysis was not performed or referenced and therefore the unidimensionality of the scale was not verified[85]. The six studies assessing reliability were of varied methodological quality (Table 3-2). One study was rated as having good methodological quality as the only deficiency was that the percentage of missing items was not described[82]. Two other studies received poor ratings for methodological quality as an inadequate sample size (<30) was used in one[83] and only percentage agreement was calculated in the other[87]. The three remaining studies received fair ratings (Table 3-2). For the first study, it was unclear if the measurements for test-retest reliability were independent, if patients were stable and if the test conditions were similar[84]. The second used a moderate sample size (43 infants)[85]. The third study also used a moderate sample size (43 and 55 infants), and lacked clarity with regards to the stability of patients during the test-retest time interval, the appropriateness of the time interval itself, and the similarity of test conditions at both time assessments[86].

Both studies assessing hypotheses testing for construct validity were found to be of fair methodological quality as a moderate sample size was used for analysis[88], hypotheses were not explicitly stated[83] , and minimal information on measurement properties of the Einstein Neonatal Neurobehavioural Assessment Scale (ENNAS)[83], the Score for Neonatal Acute Physiology (SNAP-PE)[88], the Neonatal Therapeutic Intervention Scoring System (NTISS)[88], and the Neurobiological Risk Score (NBRS)[88] were provided. The

criterion validity assessed was of fair methodological quality as it was unclear whether the criterion-measure (ENNAS) could be used as a substitute gold standard, missing items were not described and a moderate sample size was used[83].

Lastly, the study discussing responsiveness of the NAPI was found to be of poor methodological quality as a time interval was not described for the longitudinal design, and the statistical methods were not appropriate (effect sizes without hypotheses)[86].

Quality Assessment

Internal consistency was assessed by one study however only item correlations were provided; Cronbach's alpha was not determined[85]. Positive, negative and indeterminate ratings were accorded across nine studies assessing reliability (Table 3-3). Five studies received positive ratings as inter-rater reliability coefficients all exceed the required 0.70 minimum[82-86]. Three of the five studies also received negative ratings as although their inter-rater reliabilities were strong; their test-retest reliabilities ranges fell below the acceptable criteria (0.59-0.63) [84-86]. One study received an indeterminate rating as only percentage agreement was provided[87].

Construct validity was assessed in two studies, both of which tested hypotheses regarding the NAPI (Table 3-3). Both studies received negative ratings as the correlations found between the NAPI and the other measures used were all below 0.5[83, 88]. The measures to which the NAPI was compared included the ENNAS[83], the SNAP-PE[88], the NTISS[88], and the NBRS[88]. Criterion validity was assessed in one study using the ENNAS, for which an adequate description of the constructs it measures along with its measurement properties was provided[83]. A positive rating could not be assigned however as the authors acknowledge that no gold standard exists in measuring the competence of the neonatal central nervous system[83].

Responsiveness was discussed in one study however effect sizes were used without prespecified hypotheses which are not acceptable statistical methods according to Terwee quality criteria[86]. Interpretability could be assessed in five studies, all of which received

indeterminate ratings as mean and SDs of the scores were provided in less than four subgroups[82, 83, 85, 86, 88].

DISCUSSION

The objective of this paper was to assess the accuracy of reporting of measurement properties in a sample of pediatric RCTs. The pediatric RCTs, published in high impact journals between 2000 and 2010, report a single primary outcome and the measurement properties of the outcome measure with citations to support any claims. The Terwee methodological search filter retrieved relevant studies on measurement properties for all but two outcome measures, one of which does not appear to have any published studies evaluating its measurement properties. The methodological quality of the studies assessing measurement properties ranged from poor to good with only one instance of an excellent rating for the content validity of the PIPP. As with the quality of the measurement properties, the methodological quality of the studies varied for each measurement property assessed. All of the studies assessing responsiveness were found to be of poor methodological quality suggesting that the pediatric RCTs relying on these measures to assess change in scores may not be valid. If the methodological quality of a study assessing measurement properties is poor, the quality of the outcome measure remains unclear. Of the eight outcome measures examined, only one (CDRS-R) had an assessment for each of the COSMIN measurement properties. However, it should be noted that not all measurement properties are relevant for each outcome measure. The criterion validity for example, may not apply as gold standards are usually not available for all measures. Reliability, in particular inter- and intra-rater reliability, was the most commonly assessed measurement property, measurement error was the least commonly assessed and neither of the two assessments of measurement error received positive ratings. For use in RCTs, the reliability of an outcome measure is of greater importance than the measurement error. The content validity and interpretability reported consistently received indeterminate ratings. Of note, only three outcome measures reported assessments for responsiveness which is surprising since each outcome measure was used in a pediatric RCT, where responsiveness of a measure is critical in determining valid

trial results. The majority of the outcome measures received a variety of ratings (positive, negative, and indeterminate) across studies for a single measurement property assessed making it difficult to conclude whether the measure has adequate properties. Outcome measures are the tools used to generate trial results, if these measures are not valid or reliable for the purpose of the RCT, the results of the trials are rendered questionable.

The Terwee methodological PubMed highly sensitive search filter is the only tool we know of that comprehensively searches for studies on measurement properties of outcome measures. As stated by the authors, the development study of the filter has several strengths including the use of the gold standard PubMed sample, the validation of the filter in two settings, and the comprehensive and far-reaching search strategy that encompasses relevant terms for measurement properties[5]. There are however limitations of the search filter. The authors acknowledge the small sample size of the gold standard studies and the value of validating the filter in other validation sets[5]. The filter is also limited to validation or comparative published studies and therefore did not retrieve any of the outcome measure manuals. The outcome measure manuals may provide information on measurement properties and should therefore be considered in the retrieval of studies on measurement properties. For this particular study, the fact that the search strategy requires both the outcome measure and measurement property terminology to be in the title and/or abstract in order to be included, limited the retrieval of all relevant studies. This was evident as studies originally cited in the pediatric RCT were not retrieved by the search strategy for three of the outcome measures as the measure and measurement property terminology were not included in the title and/or abstract. In order to compensate for the search filter's limitations, we checked reference lists of all studies retrieved that discussed measurement properties and retrieved any potentially relevant studies that were cited. In the pediatric RCT using the CGI-I, for example, the manual was cited but upon examining the manual, it does not provide any information on measurement properties.We are therefore confident that the 70 studies assessed are in fact the most relevant and commonly cited studies for these outcome measures. Furthermore, studies on measurement properties should surely mention the

properties and the measure in their title and abstract and we strongly recommend that authors ensure these are included.

We recognize that the lack of independent duplication of study inclusion, measurement property quality assessment and methodological quality assessment may lead to discrepancies in the data extraction. However, based on the results of these assessments, the studies lacked multiple aspects required of each assessment (e.g. inadequate sample size and inappropriate statistical methods) and therefore minor discrepancies are unlikely to significantly change the overall results and conclusions of this study. Furthermore, objective criteria were used to evaluate the quality of the measurement properties and the studies that assess them such that the results are reproducible. The Terwee and COSMIN criteria were developed through consensus-based Delphi study and have been used extensively in the literature[6, 8-10].

We also did not include non-English language studies and studies discussing translated versions of the outcome measures. It is however reasonable to assume that prior to being translated or used in a different culture/population, the original measure must show validity and reliability. If the original measure is not found to be valid and reliable, its subsequent translations and applications will also be seen as invalid.

The authors of the original sample of pediatric RCTs accurately cited available studies on measurement properties for five of the eight outcome measures used. The citations provided by the pediatric RCTs for these five measures were retrieved and included in subsequent assessments. Although the authors could have cited additional or alternative studies, the citations provided included the original validation studies of the measures. Of note, the original validation studies of the five measures were conducted in a similar population as the RCT in which they were used with the exception of the RDAI and the CDRS-R where the measures were used in slightly older pediatric populations. The three remaining measures citations were for versions of the measure that were not used[89, 93] and included a manual for the measure that does not provide information on measurement properties[92].

While authors accurately cited studies on measurement properties for the large majority of the outcome measures, based on the results of the quality assessment of the measurement properties and the studies assessing them, further validation with methodologically strong studies is required. The original Terwee quality criteria was published in 2007[6], the COSMIN consensus on measurement property terminology was published in 2010[1] and the COSMIN 4-point rating scale was published in 2011[9]. We recognize that the pediatric RCTs identified in our systematic review predate these publications as we searched for RCTs published between 2000 and 2010. Nonetheless, the results of this study suggest that further work is urgently required as outcome measures are being used to generate RCT results despite their lack of validation and reliability.

To our knowledge, this is the first study in which citations provided for measurement properties of outcome measures in pediatric RCTs were comprehensively examined followed by the in depth assessment of the quality of the measurement properties and the studies in which they were assessed. The results of this study can be used to improve reporting guidelines. We strongly recommend that journal publication guidelines as well as methodological guidelines like CONSORT (Consolidated Standards of Reporting Trials) and SPIRIT (Standard Protocol Items for Randomized Trials) be revised such that they include the importance of reporting measurement properties and providing evidence in the form of appropriate citations to support any such reports. The results of this study should also encourage higher standards for the conduct of trials. Trialists should look to the quality assessment tools used in this study when conducting their respective studies to ensure they meet methodological rigorous standards. If the measures do not meet these standards, they should undergo formal validation prior to or during the RCT. Furthermore, the results presented here may also facilitate the informed selection of outcome measures thus instilling confidence in trial results and subsequent knowledge synthesis efforts.

CONCLUSION

Measurement properties of outcome measures used in RCTs must be reported to ensure the validity of trial results. When reporting the properties, studies on the measurement properties should be cited. Although authors may accurately cite available studies on measurement properties, the quality of the studies cited and the measurement properties assessed within these studies suggests that further validation is necessary.

Outcome Measure	Terwee Sensitive Search Filter Included Studies	Additional included studies retrieved	Studies originally cited in pediatric RCT
Respiratory Disease Assessment Instrument (RDAI) – Respiratory Assessment Change Score (RACS)[11]	Klassen T, Sutcliffe T, Watters L, Wells GA, Allen UD, Li MM. Dexamethasone in albuterol-treated inpatients with acute bronchiolitis: a randomized, controlled trial. *J Pediatr* 1997;130:191-7. Schuh S, Coates AL, Binnie R, Allin T, Goia C, Corey M et al. Efficacy of oral dexamethasone in outpatients with acute bronchiolitis. *J Pediatr.* 2002;140(1):27-32.	Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of albuterol in acute bronchiolitis. *J Pediatr* 1991;118:806-11. Lowell DI, Lister G, Von Koss H, Mc-Carthy P. Wheezing in infants: the response to epinephrine. *Pediatrics* 1987; 79:939-45.	Klassen T, Sutcliffe T, Watters L, Wells GA, Allen UD, Li MM. Dexamethasone in albuterol-treated inpatients with acute bronchiolitis: a randomized, controlled trial. *J Pediatr* 1997;130:191-7. Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of albuterol in acute bronchiolitis. *J Pediatr* 1991;118:806-11. Lowell DI, Lister G, Von Koss H, Mc-Carthy P. Wheezing in infants: the response to epinephrine. *Pediatrics* 1987; 79:939-45.
Infant Feeding Behaviours – Rater Checklist (IFB)[12]	-	-	Arts-Rodas D, Benoit D. Feeding problems in infancy and early childhood: identification and management. *J Paediatr Child Health* 1998; 3:21-7. Benoit D, Green D. The Infant Feeding Behaviors - Rater Checklist: preliminary data. Poster presented at the Fortysecond Annual Meeting of the American Academy of Child and Adolescent Psychiatry, New Orleans, LA; 1995. Koulis K, Arts-Rodas D, Benoit D. The Infant Feeding Behaviors - Rater checklist: comparison of coding methods. Poster presented at the forty-fourth Annual Meeting of the American Academy of Child and Adolescent Psychiatry,Toronto, Ontario; 1997.
Children's Depression Rating Scale – Revised (CDRS-R)[13]	Aronen ET, Teicher MH, Geenens D, Curtin S, Glod CA, Pahlavan K. Motor activity and severity of depression in hospitalized prepubertal children. *J. Am. Acad. Child Adolesc. Psychiatry* 1996; 35(6):752-763. Basker MM, Russell PSS, Russell S, Moses PD. Validation of the children's depression rating scale-revised for adolescents in primary-care pediatric use in India. *Indian Journal of Medical Sciences*. 2010; 64(2):72-80. Bernstein IH, Rush J, Trivedi MH, Hughes CW, Macleod L, Witte BP, Jain S, Mayes TL, Emslie GJ. Psychometric properties of the quick inventory of depressive symptomatology in adolescents. *Int. J. Methods Psychiatr. Res.* 2010; 19(4): 185–194. Brent D, Emslie G, Clarke G, Wagner KD, Asarnow JR,	Poznanski, E. O., Cook, S. C , & Carroll, B. J. A depression rating scale for children. Pediatrics, 1979. 64, 442-450.	Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating Scale-Revised. Psychopharmacol Bull. 1984;21:979-989.

	Keller M…Zelazny J. Switching to another SSRI or to venlafaxine with or without cognitive behavioural therapy for adolescents with SSRI-resistant depression: the TORDIA randomized controlled trial. *JAMA*. 2008; 299(8):901-913. Frazier TW, Demeter CA, Youngstrom EA, Calabrese JR, Stansbrey RJ, McNamara NK, Findling RL. Evaluation and comparison of psychometric instruments for pediatric bipolar spectrum disorders in four age groups. *J. Child and Adolesc. Psychopharm.* 2007; 17(6): 853-866. Fristad MA, Verducci JS, Walters K, Young ME. Impact of multifamily psychoeducational psychotherapy in treating children aged 8 to 12 years with mood disorders. *Arch Gen Psychiatry*. 2009; 66(9):1013-1021. Guo Y, Nilsson ME, Heiligenstein J, Wilson MG, Emslie G. An exploratory factor analysis of the children's depression rating scale-revised. *J. Child and Adolesc. Psychopharm.* 2006; 16(4):482-491. Jain S, Carmody TJ, Trivedi MH, Hughes C, Bernstein IH, Morris DW, Emslie GJ, Rush AJ. A psychometric evaluation of the CDRS and MADRS in assessing depressive symptoms in children. *J. Am. Acad. Child Adolesc. Psychiatry*. 2007; 46(9): 1204-1212. Kennard BD, Silva SG, Mayes TL, Rohde P, Hughes JL, Vitiello B…March JS. Assessment of safety and long-term outcomes of initial treatment with placebo in TADS. *Am J Psychiatry*. 2009; 166:337-344. King CA, Klaus N, Kramer A, Venkataraman S, Quinlan P, Gillespie B. The youth-nominated support team – version II for suicidal adolescents: a randomized controlled intervention trial. *Journal of Consulting and Clinical Psychology*. 2009; 77(5): 880-893. Treatment for Adolescents with Depression Study (TADS) Team. Fluoxetiine, cognitive-behavioral therapy, and their combination for adolescents with depression. *JAMA*. 2004; 292:807-820. Mayes TL, Bernstein IH, Haley CL, Kennard BD, Emslie GJ. Psychometric properties of the children's depression		

	rating scale-revised in adolescents. *J Child and Adolesc. Psychopharm*. 2010; 20(6): 513-516. Mokros HB, Poznanski E, Grossman JA, Freeman LN. A comparison of child and parent ratings of depression for normal and clinically referred children. *J. Child Psychol. Psychat.* 1987; 28(4):613-627. Pavuluri MN, Henry DB, Carbray JA, Naylor MW, Janicak PG. Divalproex sodium for pediatric mixed mania: a 6-month prospective trial. *Bipolar Disorders*. 2005; 7:266-273. Shain BN, King CA, Naylor M, Alessi N. Chronic depression and hospital course in adolescents. *J. Am. Acad. Child Adolesc. Psychiatry.* 1991; 30(3): 428-433. Stein GL, Curry JF, Hersh J, Breland-Noble A, Silva SG, Reinecke MA, Jacobs R. Ethnic differences among adolescents beginning treatment for depression. *Cultural Diversity and Ethnic Minority Psychology*. 2010; 16(2): 152-158. Wood BL, Lim J, Miller BD, Cheah PA, Simmens S, Stern T, Waxmonsky J, Ballow M. Family emotional climate, depression, emotional triggering of asthma, and disease severity in pediatric asthma: examination of pathways of effect. *Journal of Pediatric Psychology*. 2007; 32(5): 542-551.		
Clinical Global Impressions-Improvement Subscale (CGI-I)[13,14]	Davanzo P, Gunderson B, Belin T, Mintz J, Pataki C, Ott D…Strober M. Mood stabilizers in hospitalized children with bipolar disorder: A retrospective review. *Psychiatry and Clinical Neurosciences*. 2003; 57:504-510. King BH, Hollander E, Sikich L, McCracken JT, Scahill L, Bregman JD...Ritz L. Lack of efficacy of citalopram in children with autism spectrum disorders and high levels of repetitive behaviour. Masi G, Cosenza A, Millepiedi S, Muratori F, Pari C, Salvadori F. Aripiprazole monotherapy in children and young adolescents with pervasive developmental disorders. *CNS Drugs*. 2009; 23(6): 511-521. Steiner H, Petersen ML, Saxena K, Ford S, Matthews Z.	Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nded. Washington: US Government Printing Office; 1976.	Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nded. Washington: US Government Printing Office; 1976.

	Divalproex sodium for the treatment of conduct disorder: a randomized controlled clinical trial. *J Clin Psychiatry*. 2003; 64:1183-1191.		
Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored (ADHDRS-IV-Teacher:Inv)[15]	-	DuPaul GJ, Power TJ, Anastopoulos AD, Reid R, McGoey KE, Ikeda MJ. Teacher ratings of attention deficit hyperactivity disorder symptoms: factor structure and normative data. *Psychological Assessment*. 1997; 9(4): 436-444. DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD rating scale-IV: checklists, norms, and clinical interpretation. *Journal of Psychoeducational Assessment*. 1998; 24:172-178. Faries DE, Yalcin I, Harder D, Heiligenstein JH (2001), Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 5:39–47	Faries DE, Yalcin I, Harder D, Heiligenstein JH (2001), Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 5:39–47
Premature Infant Pain Profile (PIPP)[16]	Ahn Y, Jun Y. Measurement of pain-like response to various NICU stimulants for high-risk infants. *Early Human Development*. 2007; 83: 255-262. Arias MCC, Guinsburg R. Differences between uni- and multidimensional scales for assessing pain in term newborn infants at the bedside. *Clinics*. 2012; 67(10): 1165-1170. Badr LK, Abdallah B, Hawari M, Sidani S, Kassar M, Pascale N, Julianna B. Determinants of premature infant pain responses to heel sticks. *Pediatric Nursing*. 2010; 36(3): 129-136. Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *The Clinical Journal of Pain*. 1999; 15(4):297-303. Bellieni CV, Maffei M, Ancora G, Cordelli D, Mastrocola M, Faldella G, Ferretti E, Buonocore G. Is the ABC pain scale reliable for premature babies. *Acta Paediatrica*. 2007; 96:1008-1010. Boyle EM, Khan-Orakzai Z, Watkinson M, Wright E, Ainsworth JR, McIntosh N. Sucrose and non-nutritive		Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *Clin J Pain.* 1999;15(4):297–303. Jonsdottir RB, Kristjansdottir G. The sensitivity of the premature infant pain profile: PIPP to measure pain in hospitalized neonates. *J Eval Clin Pract.* 2005;11(6):598–605

	sucking for the relief of pain in screening for retinopathy of prematurity: a randomised controlled trial. *Arch Dis Child Fetal Neonatal Ed*. 2006; 91:F166-F168. Bueno M, Stevens B, de Camargo PP, Toma E, Lucia V, Krebs J, Kimura AF. Breast milk and glucose for pain relief in preterm infants: a noninferiority randomized controlled trial. *Pediatrics*. 2012; 129:664. Chermont AG, Falcao LFM, de Souza Silva EHL, Balda RCX, Guinsburg R. Skin-to-skin contact and/or oral 25% dextrose for procedural pain relief for tern newborn infants. *Pediatrics* 2009; 124(6):e1101-1107 Cignacco E, Denhaerynck K, Nelle M, Buhrer C, Engberg S. Variability in pain response to a non-pharmacological intervention across repeated routine pain exposure in preterm infants: a feasibility study. *Acta Paediatrica*. 2009; 98:842-846. Codipietro L, Ceccarelli M, Ponzone A. Breastfeeding or oral sucrose solution in term neonates receiving heel lance: a randomized, controlled trial. *Pediatrics* 2008; 122:e716-721. Franck LS, Ridout D, Howard R, Peters J, Honour JW. A comparison of pain measures in newborn infants after cardiac surgery. *Pain* 2011; 152:1758-1765. Gradin M, Eriksson M, Holmqvist G, Holstein A, Schollin J. Pain reduction at venipuncture in newborns: oral glucose compared with local anesthetic cream. *Pediatrics* 2002; 110(6): 1053-1057. Gradin M, Finnstrom O, Schollin J. Feeding and oral glucose – additive effects on pain reduction in newborns. *Early Human Development*. 2004; 77:57-65. Johnston CC, Campbell-Yeo M, Filion F. Paternal vs. maternal kangaroo care for procedural pain in preterm neonates. *Arch Pediatr Adolesc Med*. 2011; 165(9):792-796. Johnston CC, Stevens BJ, Franck LS, Jack A, Stremler R, Platt R. Factors explaining lack of response to heel stick in preterm newborns. *JOGNN* 1999; 28: 587-594.		

	Lemyre B, Hogan DL, Gaboury I, Sherlock R, Blanchard C, Moher D. How effective is tetracaine 4% gel, before a venipuncture, in reducing procedural pain in infants: a randomized double-blind placebo controlled trial. *BMC Pediatrics* 2007; 7(7). Lemyre B, Sherlock R, Hogan D, Gaboury I, Blanchard C, Moher D. How effective is tetracaine 4% gel, before a peripherally inserted central catheter, in reducing procedural pain in infants: a randomized double-blind placebo controlled trial. *BMC Medicine* 2006; 4(11). Liaw JJ, Yang L, Wang KWK, Chen CM, Chang YC, Yin T. Non-nutritive sucking and facilitated tucking relieve preterm infant pain during heel-stick procedures: a prospective, randomised controlled crossover trial. *International Journal of Nursing Studies* 2012; 49:300-309. Liaw JJ, Yang L, Chou HL, Yin T, Chao SC, Lee TY. Psychometric analysis of a Taiwan-version pain assessment scale for preterm infants. *Journal of Clinical Nursing* 2011; 21:89-100. McNair C, Ballantyne M, Dionne K, Stephens D, Stevens B. Postoperative pain assessment in the neonatal intensive care unit. *Arch Dis Child Fetal Neonatal Ed* 2004; 89:F537-F541. Morelius E, Hellstrom-Westas L, Carlen C, Norman E, Nelson N. Is a nappy change stressful to neonates? *Early Human Development* 2006; 82:669-676. Ozawa M, Kanda K, Hirata M, Kusakawa I, Suzuki C. Influence of repeated painful procedures on prefrontal cortical pain responses in newborns. *Acta Paediatrica* 2011; 100: 198-203. Simons SHP, van Dijk M, van Lingen RA, Roofthooft D, Duivenvoorden HJ, Jongeneel N...Tibboel D. Routine morphine infusion in preterm newborns who received ventilatory support: a randomized controlled trial. JAMA 2003; 290(18): 2419-2427. Simonse E, Mulder PGH, van Beek RHT. Analgesic effect of breast milk versus sucrose for analgesia during		

	heel lance in late preterm infants. *Pediatrics* 2012; 129(4):657-663. Slater R, Cantarella A, Franck L, Meek J, Fitzgerald M. How well do clinical pain assessment tools reflect pain in infants? *PLoS Medicine* 2008; 5(6):e129. Slater R, Cornelissen L, Fabrizi L, Patten D, Yoxen J, Worley A, Boyd S, Meek J, Fitzgerald M. Oral sucrose as an analgesic drug for procedural pain in newborn infants: a randomised controlled trial. *Lancet* 2010; 376:1225-32. South MMT, Strauss RA, South AP, Boggess JF, Thorp JM. The use of non-nutritive sucking to decrease the physiologic pain response during neonatal circumcision: a randomized controlled trial. *American Journal of Obstetrics and Gynecology* 2005; 193:537-43. Stevens B, Johnston C, Petryshen P, Taddio A. Premature infant paint profile: development and initial validation. *The Clinical Journal of Pain* 1996; 12(1):13-22.		
Modified Yale Preoperative Anxiety Scale (m-YPAS)[17]	Finley GA, Stewart SH, Buffett-Jerrott S, Wright KD, Millington D. High levels of impulsivity may contraindicate midazolam premedication in children. *Can J Anesth* 2006; 53(1):73-78. Kain ZN, MacLaren J, McClain BC, Saadat H, Wang SM, Mayes LC, Anderson GM. Effects of age and emotionality on the effectiveness of midazolam administered preoperatively to children. *Anesthesiology* 2007; 107:545-52. Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstader MB. The yale preoperative anxiety scale: how does it compare with a "gold standard"? *Anesth Analg* 1997; 85:783-8. MacLaren JE, Thompson C, Weinberg M, Fortier MA, Morrison DE, Perret D, Kain ZN. Prediction of preoperative anxiety in children: who is most accurate? *Anesth Analg* 2009; 108:1777-82. Mifflin KA, Hackmann T, Chorney JM. Streamed video clips to reduce anxiety in children during inhaled induction of anesthesia. *Anest Analg* 2012; 115(5):1162—		Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstadter MB. The Yale Preoperative Anxiety Scale: how does it compare with a "gold standard?*" Anesth Analg.* 1997;85:783–788

	7. Sadhasivam S, Cohen LL, Hosu L, Gorman KL, Wang Y, Nick TG…Gunter JB. Real-time assessment of perioperative behaviors in children and parents: development and validation of the perioperative adult child behavioural interaction scale. *Anesth Analg* 2010; 110:1109-15. Sadhasivam S, Cohen LL, Szabova A, Varughese A, Kurth CD, Willging…Gunter J. Real-time assessment of perioperative behaviors and prediction of perioperative outcomes. *Anesth Analg* 2009; 108:822-6.		
Neurobehavioural Assessment of the Preterm Infant (NAPI)[18]	Glazebrook C, Marlow N, Israel C, Croudace T, Johnson S, White IR, Whitelaw A. Randomised trial of a parenting intervention during neonatal intensive care. *Arch Dis Child Fetal Neonatal Ed* 2007; 92:F438-F443. Hyman C, Snider LM, Majnemer A, Mazer B. Concurrent validity of the Neurobehavioural Assessment for pre-term infants (NAPI) at term age. *Pediatric Rehabilitation* 2005; 8(3): 225-234. Johnston CC, Filion F, Snider L, Majnemer A, Limperopoulos C, Walker CD…Boyer K. Routine sucrose analgesia during the first week of life in neonates younger than 31 weeks' postconceptional age. *Pediatrics* 2002; 110(3):523-528. Korner AF. Reliable individual differences in preterm infants' excitation management. *Child Development* 1996; 67:1793-1805. Korner AF, Constantinou J, Dimiceli S, Brown Jr. BW, Thom VA. Establishing the reliability and developmental validity of a neurobehavioral assessment for preterm infants: a methodological process. *Child Development* 1991; 62:1200-1208. Morrow CJ, Field TM, Scafidi FA, Roberts J, Eisen L, Hogan AE, Bandstra ES. Transcutaneous oxygen tension in preterm neonates during neonatal behavioural assessments and heelsticks. *Developmental and Behavioral Pediatrics*. 1990; 11(6):312-316.		Korner AF, Kraemer HC, Reade EP, Forrest T, Dimiceli S, Thom VA. A methodological approach to developing an assessment procedure for testing the neurobehavioral maturity of preterm infants. *Child Dev.* 1987;58:1478–1487 Korner AF, Constantinou J, Dimiceli S, Brown BW, Thom VA. Establishing the reliability and developmental validity of a neurobehavioral assessment for preterm infants: a methodological process. *Child Dev.* 1991;62:1200–1208 Korner AF, Stevenson DK, Kraemer HC, et al. Prediction of the development of low birth weight preterm infants by a new neonatal medical index. *Dev Behav Pediatr.* 1993;14:106–111

	Snider L, Tremblay S, Limperopoulos C, Majnemer A, Filion F, Johnston C. Construct validity of the neurobehavioral assessment of preterm infants. *Physical & Occupational Therapy in Pediatrics* 2005; 25(3):81-95.		

Table 3-1 Studies retrieved from the Terwee Methodological Highly Sensitive Search Filter

Outcome Measure	Reliability			Validity				Responsiveness
	Internal Consistency	Measurement error	Reliability	Content	Construct S	Construct HT	Criterion	Responsiveness
Respiratory Distress Instrument (RDAI): the Respiratory Assessment Change Score (RACS)			Poor [19, 20, 21 22]	Poor [21]				Poor [21]
Children's Depression Rating Scale – Revised (CDRS-R)	Poor [23, 24, 25, 30, 32, 34. 39, 40] Good [27, 29]	Fair [27]	Fair [24, 26, 36 37, 40] Good [28, 31, 33 38, 39]	Poor [24, 40]	Good [24, 25, 27, 29]	Poor [23, 34 37, 40] Fair [24, 38]	Fair [24]	Poor [24, 27, 30, 33, 34, 36]
Clinical Global Impressions – Improvement Subscale (CGI-I)	Poor [44]		Poor [41, 42, 43] Fair [44]					
Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored ADHDRS-IV-Teacher:Inv	Poor [46]		Fair [46]		Good [45]	Fair [46]	Fair [46]	
Premature Infant Pain Profile (PIPP)	Good [74]	Poor [72]	Poor [48, 49, 50 51, 52, 53 55, 56, 57 58, 59, 60 61, 62, 63 64, 65, 66 67, 68, 69 70, 71] Fair [54, 73, 74]	Excellent [74]	Good [74]	Fair [47, 66, 67, 70] Poor [48, 50 69, 74]		
Modified Yale Preoperative Anxiety (m-YPAS)	Poor [75]		Poor [75, 76, 78 79, 80, 81] Good [77]	Poor [77]		Poor [75, 77]	Fair [77]	
Neurobehavioural	Poor [85]		Good [82]			Fair [83, 88]	Poor [83]	Poor [86]

Assessment of the Preterm Infant (NAPI)	Poor [83, 87] Fair [84, 85, 86]		

[..] = reference number

Table 3- 2 Methodological Quality of Studies using COSMIN Scoring System

Outcome Measure	Reliability			Validity				Responsiveness	Floor or ceiling effect	Interpretability
	Internal Consistency	Measurement error	Reliability	Content	Construct S	Construct HT	Criterion	Responsiveness		
Respiratory Distress Instrument (RDAI): the Respiratory Assessment Change Score (RACS)	0	0	+ [19, 20, 21, 22] ? [21]	? [21]	0	0	0	? [21]	0	? [19, 20, 22]
Infant Feeding Behaviours – Rater Checklist (IFB)	0	0	0	0	0	0	0	0	0	0
Children's Depression Rating Scale – Revised (CDRS-R)	+ [23, 24, 25, 27 29, 30, 34, 39] ? [40]	? [27]	+ [24, 26, 28, 31 32, 33, 36, 37, 38, 39 40]	? [24, 40]	+ [24] - [25, 29] ? [27]	- [23, 37, 38] + [24, 34, 40]	- [24]	+[24, 34] - [27] ? [30, 33, 36]	? [24] - [29]	? [23, 26, 30, 31, 32, 33, 34, 35, 36, 37, 38]
Clinical Global Impressions – Improvement Subscale (CGI-I)	+ [44]	0	+ [41, 43] ? [42] - [44]	0	0	0	0	0	0	? [41]
Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored ADHDRS-IV-Teacher:Inv	+ [46]	0	+ [46]	0	+ [45]	+ [46]	? [46]	0	0	? [45]
Premature Infant Pain Profile	+ [74]	? [72]	? [48, 49, 52] + [50, 53, 54,	+ [74]	+ [74]	- [47, 67, 69]	0	0	- [61, 72]	? [47, 49, 50. 52, 53, 54, 62, 63, 65, 67, 68,

(PIPP)			56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 73, 74] - [51, 55, 73]			? [48, 50, 66, 70, 74]			69, 70, 72, 73, 74]
Modified Yale Preoperative Anxiety (m-YPAS)	+ [75]	0	+ [75, 77, 79 80, 81] ? [76, 78]	? [77]	0	- [75] ? [77]	+ [77]	0	- [76]	? [76, 79]
Neurobehavioural Assessment of the Preterm Infant (NAPI)	? [85]	0	+ [82, 83, 84 85, 86] - [84, 85, 86] ? [87]	0	0	- [83, 88]	? [83]	? [86]	0	? [82, 83, 85, 86, 88]

[..] = reference number

Table 3-3 Quality Assessment of Measurement Properties using modified version of Terwee's Quality Criteria

Appendix 3-1 – Search Strategy

1. (instrumentation or methods).sh.
2. (Validation Studies or Comparative Study).pt.
3. exp Psychometrics/
4. psychometr*.ti,ab.
5. (clinimetr* or clinometr*).tw.
6. exp "Outcome Assessment (Health Care)"/
7. outcome assessment.ti,ab.
8. outcome measure*.tw.
9. exp Observer Variation/
10. observer variation.ti,ab.
11. exp Health Status Indicators/
12. exp "Reproducibility of Results"/
13. reproducib*.ti,ab.
14. exp Discriminant Analysis/
15. (reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or "internal consistency").ti,ab.
16. (cronbach* and (alpha or alphas)).ti,ab.
17. (item and (correlation* or selection* or reduction*)).ti,ab.
18. (agreement or precision or imprecision or "precise values" or test-retest).ti,ab.
19. (test and retest).ti,ab.
20. (reliab* and (test or retest)).ti,ab.
21. (stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intratester or intra-tester or interobserver or inter-observer or intraobserver or intra-observer or intertechnician or inter-technician or intratechnician or intra-technician or interexaminer or inter-examiner or intraexaminer or intra-examiner or interassay or inter-assay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-individual or interparticipant or inter-participant or intraparticipant or intra-participant or kappa or kappa's or kappas or repeatab*).ti,ab.
22. ((replicab* or repeated) and (measure or measures or findings or result or results or test or tests)).ti,ab.
23. (generaliza* or generalisa* or concordance).ti,ab.
24. (intraclass and correlation*).ti,ab.
25. (discriminative or "known group" or factor analysis or factor analyses or dimension* or subscale*).ti,ab.
26. (multitrait and scaling and (analysis or analyses)).ti,ab.
27. (item discriminant or interscale correlation* or error or errors or "individual variability").ti,ab.
28. (variability and (analysis or values)).ti,ab.
29. (uncertainty and (measurement or measuring)).ti,ab.
30. ("standard error of measurement" or sensitiv* or responsive*).ti,ab.
31. ((minimal or minimally or clinical or clinically) and (important or significant or detectable) and (change or difference)).ti,ab.
32. (small* and (real or detectable) and (change or difference)).ti,ab.

33. (meaningful change or "ceiling effect" or "floor effect" or "Item response model" or IRT or Rasch or "Differential item functioning" or DIF or "computer adaptive testing" or "item bank" or "cross-cultural equivalence").ti,ab.
34. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33
35. (child* or pediatric* or infan* or neonat* or newborn* or teen* or youth*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
36. 34 and 35

Outcome Measure	Search Terms
Respiratory Disease Assessment Instrument (RDAI) – Respiratory Assessment Change Score (RACS)	37. Respiratory Disease Assessment Instrument.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. RDAI.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. Respiratory Assessment Change Score.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. RACS.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. 37 or 38 or 39 or 40 42. 36 and 41
Infant Feeding Behaviours – Rater Checklist (IFB)	37. The Infant Feeding Behaviors - Rater Checklist.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. The Infant Feeding Behaviors.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. The Infant Feeding Behaviors Checklist.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. IFB - Rater checklist.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. IFB.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 42. 38 or 41 43. 36 and 42
Children's Depression Rating Scale – Revised (CDRS-R)	37. Children's Depression Rating Scale - Revised.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. CDRS-R.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38

	40. 36 and 39
Clinical Global Impressions – Improvement Scale (CGI-I)	37. Clinical Global Impressions-Improvement Subscale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. CGI-I.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier 39. 37 or 38 40. 36 and 39
Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored ADHDRS-IV-Teacher:Inv	37. (Attention-Deficit Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. ADHDRS-IV-Teacher:Inv.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. Attention-Deficit Hyperactivity Disorder Rating Scale-IV.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. Attention-Deficit Hyperactivity Disorder Rating Scale-IV-Teacher Version.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. ADHD RS.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 42. ADHD RS-IV.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 43. ADHD RS-IV Teacher Version.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 44. 37 or 38 or 39 or 40 or 41 or 42 or 43 45. 36 and 44
Premature Infant Pain Profile (PIPP)	37. Premature Infant Pain Profile.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. Premature Infant Pain Profile Score.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. Premature Infant Pain Profile Scale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. PIPP.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. 37 or 38 or 39 or 40 42. 36 and 41
Modified Yale Preoperative Anxiety Scale (m-YPAS)	37. The Modified Yale Preoperative Anxiety Scale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. m-YPAS.mp. [mp=title, abstract, original title, name of substance word, subject

	heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39
Neurobehavioural Assessment of Preterm Infant (NAPI)	37. Neurobehavioral Assessment of the Preterm Infant.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. NAPI.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39

Appendix 3-2: Quality criteria for measurement properties (Based on Terwee et al 2007)

Property (definitions are based on COSMIN taxonomy)			Rating	Quality Criteria
Reliability: The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions	Internal consistency: The degree of the interrelatedness among the items		+ ? -	Sub)scale unidimensional AND Cronbach's alpha(s) ≥ 0.70 Dimensionality not known OR Cronbach's alpha not determined (Sub)scale not unidimensional OR Cronbach's alpha(s) < 0.70
	Measurement error: The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured		+ ? -	MIC > SDC OR MIC outside the LOA MIC not defined MIC ≤ SDC OR MIC equals or inside LOA
	Reliability: The proportion of the total variance in the measurements which is due to 'true' differences between patients		+ ? -	ICC/weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80 Neither ICC/weighted Kappa, nor Pearson's r determined ICC/weighted Kappa < 0.70 OR Pearson's r < 0.80
Validity: The degree to which an instrument measures the construct(s) it purports to measure	Content validity: The degree to which the content of an instrument is an adequate reflection of the construct to be measured		+ ? -	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete No target population involvement The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
	Construct validity: The degree to which the scores of an instrument are consistent with hypotheses	Structural: The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be	+ ? -	Factors should explain at least 50% of the variance Explained variance not mentioned Factors explain < 50% of the variance

		measured		
		Hypothesis testing:	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses) AND correlation with related constructs is higher than with unrelated constructs;
			?	Solely correlations determined with unrelated constructs;
			-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
	Criterion validity: The degree to which the scores of an instrument are an adequate reflection of a 'gold standard'		+	Convincing arguments that gold standard is ''gold'' AND correlation with gold standard >0.70;
			?	No convincing arguments that gold standard is ''gold'' OR doubtful design or method;
			-	Correlation with gold standard <0.70, despite adequate design and method;
Responsiveness: The ability of an instrument to detect change over time in the construct to be measured			+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70) AND correlation with related constructs is higher than with unrelated constructs

	? - 	Solely correlations determined with unrelated constructs Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlation with related constructs is lower than with unrelated constructs
Floor and ceiling effect: The number of respondents who achieved the lowest or highest possible score	+ ? -	<15% of the respondents achieved the highest or lowest possible scores; Doubtful design or method; >15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods;
Interpretability : The degree to which one can assign qualitative meaning to quantitative scores	+ ? 	Mean and SD scores presented of at least four relevant subgroups of patients and MIC or MID defined; Doubtful design or method OR less than four subgroups OR no MIC or MID defined;

MIC = minimal important change, ; MID = minimal important difference; SDC = smallest detectable change,
LOA = limits of agreement, ICC = intraclass correlation coefficient,
AUC = area under the curve
+ = positive rating, ? = indeterminate rating, - = negative rating, 0 = no information

Appendix 3-3 COSMIN Checklist with 4-point rating scale (example)

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

References

[1] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737-745.

[2] Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement Issues in Trials of Pediatric Acute Diarrheal Diseases: A Systematic Review. *Pediatrics* 2010;126;e222-e231.

[3] Squires JE, Estabrooks CA, O'Rourke HM, Gustavsson P, Newburn-Cook CV, Wallin L. A systematic review of the psychometric properties of self-report research utilization measures used in healthcare. *Science* 2011;6:83.

[4] Reid GT, Walter FM, Brisbane JM, Emery JD. Family History Questionnaires Designed for Clinical Use: A Systematic Review. *Public Health Genomics* 2009; 12:73-83.

[5] 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;18:1115-1123.

[6] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Qual Life Res* 2010; 19: 539-549.

[7] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2011; 21:651-657.

[8] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, and de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34-42.

[9] Schellingerhout JM, Verhagen AP Heymans MW, , de Vet HC, Koes BW, and Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: A systematic review. *BMC Med Res Methodol* 2011; 11: 87.

[10] Cosmin | cosmin Retrieved 5/13/2013, 2013, from http://cosmin.nl/.

[11] Schuh S, Coates AL, Binnie R, Allin T, Goia C, Corey M et al. Efficacy of oral dexamethasone in outpatients with acute bronchiolitis. *J Pediatr*. 2002;140(1):27-32.

[12] Benoit D, Wang EL, Zlotkin SH. Discontinuation of enterostomy tube feeding by behavioural treatment in early childhood: A randomized controlled trial. *J Pediatr*. 2000;137(4):498-503.

[13] Asarnow JR, Emslie Gr, Clarke G, Wagner KD, Spirito A, Vitiello B et al. Treatment of Selective Serotonin Reuptake Inhibitor-Resistant Depression in Adolescents: Predictors and Moderators of Treatment Response. *J Am. Acad. Child Adolesc. Psychiatry*. 2009;48(3):330-9.

[14] Kafantaris V, Coletti DJ, Dicker R, Padula G, Pleak RR, Alvir JMJ et al. Lithium Treatment of Acute Mania in Adolescents: A Placebo-Controlled Discontinuation Study. *J Am. Acad. Child Adolesc. Psychiatry.* 2004;43(8):984-93.

[15] Weiss M, Tannock R, Kratochvil C, Dunn D, Velez-Borras J, Thomason C et al. A Randomized, Placebo-Controlled Study of Once-Daily Atomoxetine in the School Setting in Children with ADHD. *J Am. Acad. Child Adolesc. Psychiatry.* 2005;44(7):647-55.

[16] Codipietro L, Ceccarelli M, Ponzone A. Breastfeeding or Oral Sucrose Solution in Term Neonates Receiving Heel Lance: A Randomized, Controlled Trial. *Pediatrics*. 2008;122(3):e716-21.

[17] Vagnoli L, Caprilli S, Robiglio A, Messeri A. Clown Doctors as a Treatment for Preoperative Anxiety in Children: A Randomized, Prospective Study. *Pediatrics*. 2005;116:e563-7.

[18] Johnston CC, Filion F, Snider L, Majnemer An, Limperopoulos C, Walker C-D et al. Routine Sucrose Analgesia During the First Week of Life in Neonates Younger than 31 Weeks' Postconceptional Age. *Pediatrics.* 2002;110(3):523-8.

[19] Klassen T, Sutcliffe T, Watters L, Wells GA, Allen UD, Li MM. Dexamethasone in albuterol-treated inpatients with acute bronchiolitis: a randomized, controlled trial. *J Pediatr* 1997;130:191-7.

[20] Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of albuterol in acute bronchiolitis. *J Pediatr* 1991;118:806-11.

[21] Lowell DI, Lister G, Von Koss H, Mc-Carthy P. Wheezing in infants: the response to epinephrine. *Pediatrics* 1987; 79:939-45.

[22] Schuh S, Coates AL, Binnie R, Allin T, Goia C, Corey M et al. Efficacy of oral dexamethasone in outpatients with acute bronchiolitis. *J Pediatr.* 2002;140(1):27-32.

[23] Aronen ET, Teicher MH, Geenens D, Curtin S, Glod CA, Pahlavan K. Motor activity and severity of depression in hospitalized prepubertal children. *J. Am. Acad. Child Adolesc. Psychiatry* 1996; 35(6):752-763.

[24] Basker MM, Russell PSS, Russell S, Moses PD. Validation of the children's depression rating scale-revised for adolescents in primary-care pediatric use in India. *Indian Journal of Medical Sciences*. 2010; 64(2):72-80.

[25] Bernstein IH, Rush J, Trivedi MH, Hughes CW, Macleod L, Witte BP, Jain S, Mayes TL, Emslie GJ. Psychometric properties of the quick inventory of depressive symptomatology in adolescents. *Int. J. Methods Psychiatr. Res.* 2010; 19(4): 185–194.

[26] Brent D, Emslie G, Clarke G, Wagner KD, Asarnow JR, Keller M…Zelazny J. Switching to another SSRI or to venlafaxine with or without cognitive behavioural therapy for adolescents with SSRI-resistant depression: the TORDIA randomized controlled trial. *JAMA*. 2008; 299(8):901-913.

[27] Frazier TW, Demeter CA, Youngstrom EA, Calabrese JR, Stansbrey RJ, McNamara NK, Findling RL. Evaluation and comparison of psychometric instruments for pediatric bipolar spectrum disorders in four age groups. *J. Child and Adolesc. Psychopharm.* 2007; 17(6): 853-866.

[28] Fristad MA, Verducci JS, Walters K, Young ME. Impact of multifamily psychoeducational psychotherapy in treating children aged 8 to 12 years with mood disorders. *Arch Gen Psychiatry*. 2009; 66(9):1013-1021.

[29] Guo Y, Nilsson ME, Heiligenstein J, Wilson MG, Emslie G. An exploratory factor analysis of the children's depression rating scale-revised. *J. Child and Adolesc. Psychopharm.* 2006; 16(4):482-491.

[30] Jain S, Carmody TJ, Trivedi MH, Hughes C, Bernstein IH, Morris DW, Emslie GJ, Rush AJ. A psychometric evaluation of the CDRS and MADRS in assessing depressive symptoms in children. *J. Am. Acad. Child Adolesc. Psychiatry*. 2007; 46(9): 1204-1212.

[31] Kennard BD, Silva SG, Mayes TL, Rohde P, Hughes JL, Vitiello B…March JS. Assessment of safety and long-term outcomes of initial treatment with placebo in TADS. *Am J Psychiatry*. 2009; 166:337-344.

[32] King CA, Klaus N, Kramer A, Venkataraman S, Quinlan P, Gillespie B. The youth-nominated support team – version II for suicidal adolescents: a randomized controlled intervention trial. *Journal of Consulting and Clinical Psychology*. 2009; 77(5): 880-893.

[33] Treatment for Adolescents with Depression Study (TADS) Team. Fluoxetiine, cognitive-behavioral therapy, and their combination for adolescents with depression. *JAMA*. 2004; 292:807-820.

[34] Mayes TL, Bernstein IH, Haley CL, Kennard BD, Emslie GJ. Psychometric properties of the children's depression rating scale-revised in adolescents. *J Child and Adolesc. Psychopharm*. 2010; 20(6): 513-516.

[35] Mokros HB, Poznanski E, Grossman JA, Freeman LN. A comparison of child and parent ratings of depression for normal and clinically referred children. *J. Child Psychol. Psychat.* 1987; 28(4):613-627.

[36] Pavuluri MN, Henry DB, Carbray JA, Naylor MW, Janicak PG. Divalproex sodium for pediatric mixed mania: a 6-month prospective trial. *Bipolar Disorders*. 2005; 7:266-273.

[37] Shain BN, King CA, Naylor M, Alessi N. Chronic depression and hospital course in adolescents. *J. Am. Acad. Child Adolesc. Psychiatry*. 1991; 30(3): 428-433.

[38] Stein GL, Curry JF, Hersh J, Breland-Noble A, Silva SG, Reinecke MA, Jacobs R. Ethnic differences among adolescents beginning treatment for depression. *Cultural Diversity and Ethnic Minority Psychology*. 2010; 16(2): 152-158.

[39] Wood BL, Lim J, Miller BD, Cheah PA, Simmens S, Stern T, Waxmonsky J, Ballow M. Family emotional climate, depression, emotional triggering of asthma, and disease severity in pediatric asthma: examination of pathways of effect. *Journal of Pediatric Psychology*. 2007; 32(5): 542-551.

[40] Poznanski, E. O., Cook, S. C , & Carroll, B. J. A depression rating scale for children. *Pediatrics* 1979. 64, 442-450.

[41] Davanzo P, Gunderson B, Belin T, Mintz J, Pataki C, Ott D…Strober M. Mood stabilizers in hospitalized children with bipolar disorder: A retrospective review. *Psychiatry and Clinical Neurosciences*. 2003; 57:504-510.

[42] King BH, Hollander E, Sikich L, McCracken JT, Scahill L, Bregman JD...Ritz L. Lack of efficacy of citalopram in children with autism spectrum disorders and high levels of repetitive behaviour. *Arch Gen Psychiatry*. 2009; 66(6):583-590.

[43] Masi G, Cosenza A, Millepiedi S, Muratori F, Pari C, Salvadori F. Aripiprazole monotherapy in children and young adolescents with pervasive developmental disorders. *CNS Drugs*. 2009; 23(6): 511-521.

[44] Steiner H, Petersen ML, Saxena K, Ford S, Matthews Z. Divalproex sodium for the treatment of conduct disorder: a randomized controlled clinical trial. *J Clin Psychiatry* 2003; 64:1183-1191.

[45] DuPaul GJ, Power TJ, Anastopoulos AD, Reid R, McGoey KE, Ikeda MJ. Teacher ratings of attention deficit hyperactivity disorder symptoms: factor structure and normative data. *Psychological Assessment* 1997; 9(4): 436-444.

[46] DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD rating scale-IV: checklists, norms, and clinical interpretation. *Journal of Psychoeducational Assessment*. 1998; 24:172-178.

[47] Ahn Y, Jun Y. Measurement of pain-like response to various NICU stimulants for high-risk infants. *Early Human Development*. 2007; 83: 255-262.

[48] Arias MCC, Guinsburg R. Differences between uni- and multidimensional scales for assessing pain in term newborn infants at the bedside. *Clinics*. 2012; 67(10): 1165-1170.

[49] Badr LK, Abdallah B, Hawari M, Sidani S, Kassar M, Pascale N, Julianna B. Determinants of premature infant pain responses to heel sticks. *Pediatric Nursing*. 2010; 36(3): 129-136.

[50] Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *The Clinical Journal of Pain*. 1999; 15(4):297-303.

[51] Bellieni CV, Maffei M, Ancora G, Cordelli D, Mastrocola M, Faldella G, Ferretti E, Buonocore G. Is the ABC pain scale reliable for premature babies. *Acta Paediatrica*. 2007; 96:1008-1010.

[52] Boyle EM, Khan-Orakzai Z, Watkinson M, Wright E, Ainsworth JR, McIntosh N. Sucrose and non-nutritive sucking for the relief of pain in screening for retinopathy of prematurity: a randomised controlled trial. *Arch Dis Child Fetal Neonatal Ed*. 2006; 91:F166-F168.

[53] Bueno M, Stevens B, de Camargo PP, Toma E, Lucia V, Krebs J, Kimura AF. Breast milk and glucose for pain relief in preterm infants: a noninferiority randomized controlled trial. *Pediatrics*. 2012; 129:664.

[54] Chermont AG, Falcao LFM, de Souza Silva EHL, Balda RCX, Guinsburg R. Skin-to-skin contact and/or oral 25% dextrose for procedural pain relief for tern newborn infants. *Pediatrics* 2009; 124(6):e1101-1107

[55] Cignacco E, Denhaerynck K, Nelle M, Buhrer C, Engberg S. Variability in pain response to a non-pharmacological intervention across repeated routine pain exposure in preterm infants: a feasibility study. *Acta Paediatrica*. 2009; 98:842-846.

[56] Codipietro L, Ceccarelli M, Ponzone A. Breastfeeding or oral sucrose solution in term neonates receiving heel lance: a randomized, controlled trial. *Pediatrics* 2008; 122:e716-721.

[57] Franck LS, Ridout D, Howard R, Peters J, Honour JW. A comparison of pain measures in newborn infants after cardiac surgery. *Pain* 2011; 152:1758-1765.

[58] Gradin M, Eriksson M, Holmqvist G, Holstein A, Schollin J. Pain reduction at venipuncture in newborns: oral glucose compared with local anesthetic cream. *Pediatrics* 2002; 110(6): 1053-1057.

[59] Gradin M, Finnstrom O, Schollin J. Feeding and oral glucose – additive effects on pain reduction in newborns. *Early Human Development*. 2004; 77:57-65.

[60] Johnston CC, Campbell-Yeo M, Filion F. Paternal vs. maternal kangaroo care for procedural pain in preterm neonates. *Arch Pediatr Adolesc Med*. 2011; 165(9):792-796.

[61] Johnston CC, Stevens BJ, Franck LS, Jack A, Stremler R, Platt R. Factors explaining lack of response to heel stick in preterm newborns. *JOGNN* 1999; 28: 587-594.

[62] Lemyre B, Hogan DL, Gaboury I, Sherlock R, Blanchard C, Moher D. How effective is tetracaine 4% gel, before a venipuncture, in reducing procedural pain in infants: a randomized double-blind placebo controlled trial. *BMC Pediatrics* 2007; 7(7).

[63] Lemyre B, Sherlock R, Hogan D, Gaboury I, Blanchard C, Moher D. How effective is tetracaine 4% gel, before a peripherally inserted central catheter, in reducing procedural pain in infants: a randomized double-blind placebo controlled trial. *BMC Medicine* 2006; 4(11).

[64] Liaw JJ, Yang L, Wang KWK, Chen CM, Chang YC, Yin T. Non-nutritive sucking and facilitated tucking relieve preterm infant pain during heel-stick procedures: a prospective, randomised controlled crossover trial. *International Journal of Nursing Studies* 2012; 49:300-309.

[65] Liaw JJ, Yang L, Chou HL, Yin T, Chao SC, Lee TY. Psychometric analysis of a Taiwan-version pain assessment scale for preterm infants. *Journal of Clinical Nursing* 2011; 21:89-100.

[66] McNair C, Ballantyne M, Dionne K, Stephens D, Stevens B. Postoperative pain assessment in the neonatal intensive care unit. *Arch Dis Child Fetal Neonatal Ed* 2004; 89:F537-F541.

[67] Morelius E, Hellstrom-Westas L, Carlen C, Norman E, Nelson N. Is a nappy change stressful to neonates? *Early Human Development* 2006; 82:669-676.

[68] Ozawa M, Kanda K, Hirata M, Kusakawa I, Suzuki C. Influence of repeated painful procedures on prefrontal cortical pain responses in newborns. *Acta Paediatrica* 2011; 100: 198-203.

[69] Simons SHP, van Dijk M, van Lingen RA, Roofthooft D, Duivenvoorden HJ, Jongeneel N...Tibboel D. Routine morphine infusion in preterm newborns who received ventilatory support: a randomized controlled trial. *JAMA* 2003; 290(18): 2419-2427.

[70] Simonse E, Mulder PGH, van Beek RHT. Analgesic effect of breast milk versus sucrose for analgesia during heel lance in late preterm infants. *Pediatrics* 2012; 129(4):657-663.

[71] Slater R, Cantarella A, Franck L, Meek J, Fitzgerald M. How well do clinical pain assessment tools reflect pain in infants? *PLoS Medicine* 2008; 5(6):e129.

[72] Slater R, Cornelissen L, Fabrizi L, Patten D, Yoxen J, Worley A, Boyd S, Meek J, Fitzgerald M. Oral sucrose as an analgesic drug for procedural pain in newborn infants: a randomised controlled trial. *Lancet* 2010; 376:1225-32.

[73] South MMT, Strauss RA, South AP, Boggess JF, Thorp JM. The use of non-nutritive sucking to decrease the physiologic pain response during neonatal circumcision: a randomized controlled trial. *American Journal of Obstetrics and Gynecology* 2005; 193:537-43.

[74] Stevens B, Johnston C, Petryshen P, Taddio A. Premature infant paint profile: development and initial validation. *The Clinical Journal of Pain* 1996; 12(1):13-22.

[75] Finley GA, Stewart SH, Buffett-Jerrott S, Wright KD, Millington D. High levels of impulsivity may contraindicate midazolam premedication in children. *Can J Anesth* 2006; 53(1):73-78.

[76] Kain ZN, MacLaren J, McClain BC, Saadat H, Wang SM, Mayes LC, Anderson GM. Effects of age and emotionality on the effectiveness of midazolam administered preoperatively to children. *Anesthesiology* 2007; 107:545-52.

[77] Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstader MB. The yale preoperative anxiety scale: how does it compare with a “gold standard”? *Anesth Analg* 1997; 85:783-8.

[78] MacLaren JE, Thompson C, Weinberg M, Fortier MA, Morrison DE, Perret D, Kain ZN. Prediction of preoperative anxiety in children: who is most accurate? *Anesth Analg* 2009; 108:1777-82.

[79] Mifflin KA, Hackmann T, Chorney JM. Streamed video clips to reduce anxiety in children during inhaled induction of anesthesia. *Anest Analg* 2012; 115(5):1162—7.

[80] Sadhasivam S, Cohen LL, Hosu L, Gorman KL, Wang Y, Nick TG…Gunter JB. Real-time assessment of perioperative behaviors in children and parents: development and validation of the perioperative adult child behavioural interaction scale. *Anesth Analg* 2010; 110:1109-15.

[81] Sadhasivam S, Cohen LL, Szabova A, Varughese A, Kurth CD, Willging…Gunter J. Real-time assessment of perioperative behaviors and prediction of perioperative outcomes. *Anesth Analg* 2009; 108:822-6.

[82] Glazebrook C, Marlow N, Israel C, Croudace T, Johnson S, White IR, Whitelaw A. Randomised trial of a parenting intervention during neonatal intensive care. *Arch Dis Child Fetal Neonatal Ed* 2007; 92:F438-F443.

[83] Hyman C, Snider LM, Majnemer A, Mazer B. Concurrent validity of the Neurobehavioural Assessment for pre-term infants (NAPI) at term age. *Pediatric Rehabilitation* 2005; 8(3): 225-234.

[84] Johnston CC, Filion F, Snider L, Majnemer A, Limperopoulos C, Walker CD…Boyer K. Routine sucrose analgesia during the first week of life in neonates younger than 31 weeks' postconceptional age. *Pediatrics* 2002; 110(3):523-528.

[85] Korner AF. Reliable individual differences in preterm infants' excitation management. *Child Development* 1996; 67:1793-1805.

[86] Korner AF, Constantinou J, Dimiceli S, Brown Jr. BW, Thom VA. Establishing the reliability and developmental validity of a neurobehavioral assessment for preterm infants: a methodological process. *Child Development* 1991; 62:1200-1208.

[87] Morrow CJ, Field TM, Scafidi FA, Roberts J, Eisen L, Hogan AE, Bandstra ES. Transcutaneous oxygen tension in preterm neonates during neonatal behavioural assessments and heelsticks. *Developmental and Behavioral Pediatrics*. 1990; 11(6):312-316.

[88] Snider L, Tremblay S, Limperopoulos C, Majnemer A, Filion F, Johnston C. Construct validity of the neurobehavioral assessment of preterm infants. *Physical & Occupational Therapy in Pediatrics* 2005; 25(3):81-95.

[89] Arts-Rodas D, Benoit D. Feeding problems in infancy and early childhood: identification and management. *Paediatr Child Health* 1998; 3(1):21-27.

[90] Benoit D, Green D. The Infant Feeding Behaviors - Rater Checklist: preliminary data. Poster presented at the Forty-second Annual Meeting of the American Academy of Child and Adolescent Psychiatry, New Orleans, LA; 1995.

[91] Koulis K, Arts-Rodas D, Benoit D. The Infant Feeding Behaviors - Rater checklist: comparison of coding methods. Poster presented at the forty-fourth Annual

Meeting of the American Academy of Child and Adolescent Psychiatry,Toronto, Ontario; 1997.

[92] Guy W. *ECDEU Assessment Manual for Psychopharmacology, Revised.* Rockville, MD: National Institute of Mental Health; 1976.

[93] Faries DE, Yalcin I, Harder D, Heiligenstein JH (2001), Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 5:39–47

Chapter 4
Validation of Primary Outcome Measures in Pediatric Randomized Controlled Trials (RCTs)

Zafira Bhaloo[1,4], Lisa Hartling[2], Caroline B Terwee[3], Sunita Vohra[1,4]

[1]CARE Program, Edmonton Continuing Care Center
Unit 8B, 11111 Jasper Ave
Edmonton AB T5K 0L4
[2]Alberta Research Center for Health Evidence, Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta
Edmonton Clinic Health Academy
11405-87 Avenue
Edmonton, AB T6T 1C9
[3]VU University Medical Center, Knowledgecenter Measurement Instruments, Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
van der Boechorststraat 7
1081 BT Amsterdam
[4]Department of Pediatrics, Faculty of Medicine & Dentistry, University of Alberta
Edmonton Clinic Health Academy
11405-87 Avenue

Correspondence to: S Vohra svohra@ualberta.ca

BACKGROUND

In the hierarchy of research design, randomized controlled trials (RCTs) are found at the top, superseded only by knowledge synthesis efforts such as systematic reviews. The high quality evidence generated by RCTs is gathered in knowledge synthesis efforts and informs decision makers at various levels. The evidence produced by the RCTs depends largely on the tools used to measure the variable of interest or the primary outcome. An inappropriate measure can thus compromise the results of an RCT. It is therefore crucial that these tools or outcome measures are valid, reliable and carefully selected by trialists and researchers.

In order to make an evidence-based decision when selecting outcome measures, evidence supporting the validity and reliability of the measures should be readily available and cautiously reviewed. This evidence can be found in the form of measurement properties. Measurement properties fall within three domains: reliability, validity, and responsiveness[1]. Reporting measurement properties is an important step of any RCT that uses an outcome measure as these properties outline the outcome measure's purpose, how well it performs and with what accuracy and, critical for trials, its ability to perceive change when it occurs. When such evidence is not provided for a particular measure, trialists and researchers are unable to make an informed choice as to which measure is the most appropriate for the trial's purpose. This then leads to compromised confidence in the trial's results.

The lack of reports on the validity and reliability of measures used in RCTs has been briefly discussed in systematic reviews of specified clinical areas[2,3]. Furthermore, the abundance of measures in clinical use that have not been formally evaluated has been raised as a concern[4]. The importance of formally validating instruments has been discussed in terms of the consequent impediment on the conduct of meaningful trials and knowledge synthesis.

In our systematic review of primary outcomes reporting, we examined pediatric RCTs published in high impact journals and identified a small sample of 12 RCTs that did not report measurement properties of their primary outcome measures. One of the outcome measures used within these 12 RCTs, the Clinical Global Impressions – Improvement Subscale (CGI-I), was also used in one of seven trials that did report its measurement properties. The CGI-I was examined in-depth when assessing the accuracy of reporting of measurement properties in that trial. Measurement properties for the CGI-I were available and their quality along with the methodological quality of the studies in which they were evaluated were assessed. The CGI-I was therefore not re-assessed here as the same methods are used to identify and evaluate the measurement properties. In this paper, we further examine the 11 outcome measures used in the sample of RCTs identified, to: (1) identify studies on measurement properties of the primary outcome measures to determine whether any measurement properties are available to be reported, (2) assess the methodological quality of the studies on measurement properties identified in order to corroborate any conclusions about the quality of outcome measures, and (3) assess the quality of the outcome measures by critically appraising their measurement properties.

METHODS

Search Strategy

The Terwee methodological PubMed search filter[5] was used, with the help of an experienced research librarian, in the MEDLINE electronic database, to identify studies on measurement properties of the outcome measures used in the pediatric population. This highly sensitive search filter, developed and validated in four phases, is designed to retrieve studies on measurement properties of outcome measures[5]. The filter consists of a combination of search terms, carefully selected and amalgamated according to their sensitivity and precision, to be used in combination with additional search terms outlining the construct of interest, the population of interest, and the outcome measure of interest. These additional search terms were defined with the help of an experienced research librarian. All possible names for each outcome measure, including acronyms, were

included as search terms to ensure the retrieval of all potential studies assessing measurement properties (Appendix 4-1).

Study Selection

Titles and abstracts of all retrieved studies were screened. If, within the titles and abstracts of studies, the original version of the outcome measure was mentioned, the full text was retrieved. Only studies published in English were included. Studies discussing translated or alternate versions of the outcome measures were not included as it is important that the measurement properties of the original versions of measures are well established prior to their translation or adaptation in a different population. Full texts of studies meeting the inclusion criteria were retrieved and examined for information on measurement properties. Additional studies that were not initially included in the search filter's results were also retrieved by searching reference lists of the included studies. In our sample of pediatric RCTs, measurement properties of the measures were not reported but citations were provided for the measures themselves. All studies that were originally cited in our sample of pediatric RCTs were also retrieved if they had not already been included.

Quality Assessment

Methodological Quality of Studies on Measurement Properties

The 4-point scoring system of the COSMIN checklist was used to assess the methodological quality of the studies on measurement properties[8,9]. The COSMIN checklist is the result of a four-round Delphi study engaging international experts on the definitions of measurement properties[8]. The checklist presents measurement property specific items that represent standards for the design and statistical methodology for evaluating measurement properties. A box is provided for each of the nine measurement properties. In order to allocate an overall methodological quality score for each measurement property, the scoring system was developed[9]. To complete the COSMIN

checklist with 4-point rating scale, three steps must be followed. The first requires the identification of the measurement properties that were evaluated in the study. Next, the boxes corresponding to the properties identified in the first step are completed by rating each item on a 4-point scale: excellent, good, fair, or poor. Finally, the lowest rating of any item in the box is determined to be the overall methodological quality score.

If adequate evidence is provided regarding the methodological quality aspect of the study, the property receives an "excellent" rating. When the quality aspect can be assumed to be adequate although relevant information is missing, a "good" rating is given. If the adequacy of the methodological quality remains unclear, the property receives a "fair" rating. When evidence is provided that the methodological quality is inadequate, a "poor" rating is allocated to the property. The 4-point rating scale is explained in detail on www.cosmin.nl[10].

Quality of Measurement Properties

For studies that assessed measurement properties of the measures, the quality of the properties was evaluated using a modified version of the Terwee quality criteria[6,7]. If studies cited measurement properties, these citations were retrieved and reviewed. If measurement properties were assessed within these citations they too were evaluated.

The modified version of the Terwee quality criteria consists of nine measurement properties over three domains: 1) reliability (internal consistency, measurement error, reliability), (2) validity (content, construct, criterion), and (3) responsiveness (responsiveness)[1,6,7]. Possible ratings for each measurement property are: positive (+), indeterminate (?), negative (-), or "no information available" (0) (Appendix 4-2).

RESULTS

The Terwee sensitive search filter retrieved 1328 unique studies for the 11 outcome measures. After examining the full texts and reference lists of 684 studies, a total of 62

studies assessing measurement properties of the measures of interest were included. All studies included as well as the studies originally cited in our pediatric RCT sample, are presented in Table 4-1 by outcome measure. Table 4-2 presents the methodological quality of the studies for each outcome measure. The Terwee quality criteria measurement property assessment can be found in Table 4-3.

Global Clinical Judgements (GCJ)

The GCJ provides a clinician rating of the global improvement or worsening of a subject at the end of treatment[10]. The authors of the pediatric RCT reporting the use of the GCJ provided three citations for the measure without any reports of measurement properties. Two of these citations were retrieved by the Terwee filter however the full text of one could not be recovered. The third study originally cited was not identified by the filter and upon attempting its retrieval, the full text could not be found. However, the title and abstract of this study did not mention the GCJ or any measurement properties. The Terwee filter also retrieved one other study that was not originally cited by the pediatric RCT but the full text could not be found. None of the studies retrieved discussed the measurement properties of the GCJ and therefore none were included in the subsequent quality assessments.

Canadian Acute Respiratory Infection and "Flu Scale" (CARIFS)

The CARIFS is an 18-item, three domain measure of the severity of acute respiratory infection validated for children up to 12 years of age and is completed by parents[11]. Each item consists of a 4-point ordinal scale, with the sum of all items forming the total score[11]. The RCT from which this measure was identified did not report any measurement properties for the CARIFS however the authors did provide a reference for the outcome measure. This citation along with one other relevant citation was retrieved by the Terwee search filter (Table 4-1).

Methodological Quality

In the two studies identified, all measurement properties except measurement error, structural validity, criterion validity and floor and ceiling effects were assessed (Table 4-3).

The internal consistency evaluations were determined to be of poor methodological quality as factor analysis was not performed or referenced in either study to determine the unidimensionality of the measure and the first study did not provide Cronbach's alpha for each subscale[11, 12]. The study assessing reliability showed fair methodological quality as it was not clear how missing items were handled[11].

The first study showed excellent methodological quality for content validity as all required assessments relating to the relevance and comprehensiveness of the items were conducted[11]. The second study did not assess whether the items comprehensively reflected the construct and was therefore rated poorly[12]. Both studies received poor ratings for the methodological quality surrounding hypotheses testing as no information on the measurement properties of the comparator instrument was provided in either and the first study failed to provide a description of the comparator measure[11, 12].

The responsiveness of the CARIFS evaluated in both studies was also of poor methodological quality due to the lack of description for the comparator instrument and its measurement properties as well as the use of effect sizes without prespecified hypotheses in the second study[11, 12].

Quality Assessment

Both studies assessed the internal consistency of the measure using Cronbach's alpha and item-total correlations however, neither assessed nor referenced the unidimensionality of the scales[11, 12]. One study assessed the intra-rater reliability and found it to be adequate (0.808)[11]. Of note, content validity was assessed in both studies and received positive ratings as the target population was involved in determining the relevance and comprehensiveness of the CARIFS and its items[11, 12]. Construct validity in the form of hypotheses testing was also assessed in both studies using clinician and parent assessment

of the visual analog scale (VAS) however all correlations were weak (<0.50)[11, 12]. The responsiveness of the CARIFS was evaluated in both studies but the second study reported effect sizes without prespecified hypotheses and thus could not be well interpreted[12]. In the first study, the changes in the CARIFS score were compared with the changes in the parent's rating of the VAS and the nurse's clinical assessment and stronger correlations were determined with the related constructs than the unrelated[11]. Lastly, interpretability could be evaluated in one study however no minimal important change (MIC) was defined[11].

Gross Motor Function Measure (GMFM)

The GMFM was designed for pediatric physical therapists to measure gross motor function in children with cerebral palsy. The original version consists of 88 items across five domains with each item scored on a 4-point rating scale. Item scores are then summed into a percentage score to give an overall GMFM score[13]. Following the design and validation of the GMFM, further analyses were conducted and an "interval-level" GMFM-66 measure was validated[13]. We searched for studies on measurement properties for the entire GMFM-88 measure as the pediatric RCT from which this measure was identified used the original total GMFM-88. The Terwee filter identified 31 relevant studies, one of which was one of the three citations originally provided in the pediatric RCT (Table 4-1). The two other citations provided in the pediatric RCT were retrieved and one was included. The other study's full text could not be found.

Methodological Quality

The GMFM had all measurement properties evaluated with the exception of content, structural and criterion validity (Table 4-3).

The study assessing internal consistency was of poor methodological quality as factor analysis was not performed or referenced, and internal consistency was not calculated for subscales[29]. The methodological quality of the evaluation of measurement error was poor due to a small sample size of 10[17]. Four of the 23 studies assessing reliability had fair

methodological quality due to the use of Spearman correlation coefficients[14], moderate sample size[24, 30] , and lack of specification of a time interval and stability of patients and test conditions for test-retest reliability[25]. The other studies all received poor methodological quality ratings mostly due to small sample size[15, 18, 19, 21, 22, 23, 26, 27, 30, 34, 37, 39, 40, 41, 44] and/or reporting of percentage agreements[15, 16, 23, 39]. Three studies received good ratings as only information on missing items was not provided[33, 35, 36] and several assumptions could be made with regards to the conditions of the test-retest reliability[35].

The studies evaluating the GMFM's construct validity through hypotheses testing received mostly poor ratings as minimal information was provided about the comparator instruments[18, 31] and small sample sizes were used for analysis (<30) [22, 31]. One study received a fair rating as hypotheses were not formulated and more information was needed about the comparator instrument[29].

The majority of the studies evaluating responsiveness were poorly rated due to small sample size[32, 37, 45], inappropriate statistical measures of responsiveness without prespecified hypotheses[28, 32, 38, 42] , and lack of information on the comparator instrument[42, 43, 44]. Two studies received fair ratings for providing vague hypotheses[20, 35].

Quality Assessment

One study assessed the internal consistency of the measure and received an indeterminate rating as although the Cronbach's alpha was high (0.99), the unidimensionality of the scale was not assessed or referenced[29]. Measurement error was evaluated in one study and received a negative rating as the minimal important change (MIC) falls inside the limits of agreement (LOA) calculated[17]. The reliability of the GMFM was the most commonly evaluated property (n=23 studies), and received positive ratings for the majority of the studies (n=19) (Table 4-3). Most studies found high intra- and/or inter-rater reliability (>0.80) [14, 15, 18, 19, 21, 22, 24, 26, 27, 33, 34, 36, 37, 40, 41, 44] for the GMFM and adequate test-retest reliability (>0.70)[15, 25, 30, 35, 37]. Four studies received indeterminate ratings for their reliability as percentage agreements were provided which are not considered adequate[6].

Construct validity in the form of hypotheses testing was assessed in four studies. All four studies received positive ratings as when compared to other measures there were strong correlations (>0.50) with related constructs and weaker correlations with unrelated constructs[18, 22, 31] and when discriminating between known groups, all the results were in accordance with hypotheses[29]. The measures with which the GMFM was compared include the Quantitative Muscle Testing (QMT) [31] and the Pediatric Evaluation of Disability Inventory (PEDI) [18].

The responsiveness of the GMFM was evaluated in 10 studies and none received a negative rating (Table 4-3). Six studies received indeterminate ratings as inappropriate measures of responsiveness were used such as effect sizes[28], standard response means[28, 42], p values[38] or mean change scores[32, 37, 45] without prespecified hypotheses. The four studies that received positive ratings provided correlations with other measures on related and unrelated constructs, discussed how these accorded with a priori hypotheses[20, 35, 43] and provided area under the curve (AUC) [45]. Floor and ceiling effects were also discussed in three studies, two of which found significant ceiling effects for the GMFM (>15% achieved highest possible score) [29, 43]. One study found that none of the subjects had GMFM scores close to the minimum or maximum[28]. Interpretability could be evaluated in one study that provided means and standard deviations (SDs) of four subgroups and also provided a MIC to enable interpretation[17]. The six other studies had no defined MIC and three of them had means and SDs for less than four groups[25, 28, 29].

Gross Motor Function Classification System (GMFCS)

The GMFCS, a 5-level classification system, assesses current gross motor function for children upto 12 years of age with cerebral palsy and is rated by physical therapists[59]. The Terwee search filter retrieved 13 relevant studies on the GMFCS's measurement properties (Table 4-1). Two studies were originally cited in the pediatric RCT from which the GMFCS was identified; neither was identified by the filter as the measure's full name was not mentioned in the title or abstract. Both studies were retrieved, one was found to be relevant and is included in the subsequent assessments.

Methodological Quality

Reliability, content validity, construct validity with hypotheses testing and criterion validity were four measurement properties evaluated for the GMFCS (Table 4-3).

The studies evaluating reliability of the GMFCS had varied methodological quality (Table 4-2). The majority, five studies, had fair methodological quality due to moderate sample size[48, 59], use of an unweighted kappa[59] , and for test-retest reliability, minimal information on time interval and test conditions[53, 56, 57]. Five studies had good methodological quality as many aspects of test conditions and measurements could be assumed to be adequate[51,52, 58] and sample sizes ranged from good[51, 52, 55, 58] to adequate[54].

The study assessing content validity was of excellent methodological quality as a nominal group process followed by a Delphi survey consensus assessed the relevance and comprehensiveness of the measure for the study population[59]. The one study evaluating criterion validity was found to be of fair methodological quality as it had a moderate sample size and an unweighted kappa was calculated[48]. For hypotheses testing, four of the five studies evaluating the construct validity were of poor methodological quality as measurement properties of the comparator instruments were not provided[49-51] and for one study correlations were not provided[48]. One of the studies assessing construct validity received a fair rating as the hypotheses could have been more clearly stated along with a description of how missing items were handled[46].

Quality Assessment

All three possible ratings were provided for reliability. Ten studies received a positive rating as the intra- or inter-rater reliability was >0.75[47, 48, 51-56, 58]. Three of these studies also found strong test-retest reliability (>0.79)[56-58]. One study received a negative rating as the inter-rater reliability for children under 2 years of age was low (0.55)[59].

Content validity was assessed in one study among experts including physical and occupational therapists as well as pediatricians with expertise in cerebral palsy[59]. A nominal group process along with Delphi consensus in a series of four development

stages were used to ensure relevance and comprehensiveness of the GMFCS for the target population[59]. Hypotheses' testing to assess construct validity was evaluated in five studies, four of which received positive ratings as they all showed strong correlations with similar constructs and weaker correlations with less related parameters/constructs[46, 49, 50, 51]. The GMFCS was commonly correlated with the Manual Ability Classification System (MACS) to show concurrent validity[49-51] and with the Communication Function Classification System (CFCS) to show weaker correlations[50]. One study received an indeterminate rating as the authors claimed to show construct validity with a known discriminant group however no correlations were provided[48]. This same study also assessed criterion validity as it compared the GMFCS with the Bayley Scales of Infant Development, the Peabody Developmental Motor Scales and the Vineland Adaptive Behaviour Scales and found stronger correlations with related constructs, however no convincing arguments were provided that these measures could serve as gold standards[48].

Mental Developmental Index of the Bayley Scales of Infant Development II (BSID-II)

The Mental Development Index (MDI) of the BSID-II is a measure of cognitive function in high-risk and preterm infants one to 42 months of age[60]. The Terwee search filter retrieved one relevant study assessing the measurement properties of the Mental Development Index. The pediatric RCT from which we identified the measure cited two references, one being the reference retrieved by the filter and the other was the measure's manual (Table 4-1). The Terwee filter does not include manuals in its publication types and therefore would not retrieve this reference.

Methodological Quality

The methodological quality of the study assessing construct validity was of poor quality as no information on the measurement properties of the Kaufmann Assessment Battery for Children Mental Processing Composite (KABC-MPC) was provided[60].

Quality Assessment

One study was assessed for one measurement property – the construct validity of the measure (Table 4-3). The authors compared the MDI with the KABC-MPC but found low correlations (0.37)[60].

Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)

The CY-BOCS is a clinician-administered measure of obsession and compulsion severity and consists of 10 items on a five-point Likert scale[68]. Nine relevant studies were retrieved by the Terwee filter including the study originally cited in the pediatric RCT that stated the use of this measure (Table 4-1).

Methodological Quality

Five measurement properties were evaluated among the nine relevant studies including internal consistency, reliability, construct validity, and responsiveness (Table 4-3).

The studies assessing internal consistency ranged from poor to good methodological quality (Table 4-2). Five studies received poor ratings as factor analysis was not performed or referenced[62, 67] and internal consistency was not calculated for each subscale[65, 66, 68]**.** One study received a good rating as the item response theory (IRT) model was applied however its method of estimation was not adequately described[64]. The last study also earned a good rating as the sample size was good for both the internal consistency analysis and the unidimensionality analysis[69]. The studies assessing reliability were of poor quality due to small sample size (<30) [61, 66, 67] and fair quality due to moderate sample sizes[62, 68].

Construct validity ranged from fair to good methodological quality (Table 4-2). The two studies assessing structural validity received good ratings as the method of estimation for the IRT model was not described[64] and the sample size was adequate[69]. For hypotheses testing, three studies received fair ratings as hypotheses were vague but could be assumed[62, 68, 69]. The two other studies assessing construct validity received poor ratings as no information was provided for the comparator instruments[63, 67].

Lastly, both studies assessing responsiveness of the measure were of poor methodological quality as no information was provided on the comparator instrument[69] and inappropriate statistical methods were used or missing altogether[67, 69].

Quality Assessment

Internal consistency received all three possible ratings. Five studies received indeterminate ratings as although the Cronbach's alphas provided were adequate, the studies did not perform or reference factor analysis to confirm the unidimensionality of the scale[62, 65-68]. One study conducted factor analysis and provided strong internal consistencies for the 2 subscales as well as for the total scale (>0.90) thus earning a positive rating[64]. The last study received a negative rating as the internal consistency of one of the subscales was quite low (0.47) [69]. All studies evaluating reliability received positive ratings. Two studies assessed inter-rater reliability and found strong correlations[61, 67] and the other two studies found strong test-retest reliability over six weeks[62, 68].

Construct validity was evaluated in terms of structural validity and hypotheses testing. For structural validity, one study received an indeterminate rating as confirmatory factor analysis was performed but the explained variance was not reported[64]. The other study received a positive rating as factor analysis was performed and the two factor model explained 70% of the variance[69]. Five studies used hypotheses testing to confirm construct validity. Four of the five earned positive ratings as convergent and divergent validity were strongly demonstrated using the National Institutes of Mental Health-Global Rating Scales[62] and Global Obsessive-Compulsive Scale[63], the Conners Parent Rating Scale-Revised[62], the Hamilton Depression Rating Scale[63], and the Children's Depression Inventory[63]. Strong correlations were found with related constructs and weak correlations with unrelated constructs. One study received a negative rating for hypotheses testing as correlations with the measures used were all less than 0.5, when stronger correlations were expected[69].

Two studies assessed responsiveness of the CY-BOCS and both received indeterminate ratings. One study used inappropriate statistical methods (p values)[62] and the second study used scores from previously conducted RCTs and compared these with an external measure however no correlations were provided[67].

Oblique Diameter Difference Index (ODDI)

The ODDI is one of the most clinically important measures reflecting the severity of deformational plagiocephaly in children from 0 to 24 months of age and is obtained when using plagiocephalometry[70]. One relevant study was obtained using the Terwee filter and this was the same study cited by the pediatric RCT from which we identified the measure (Table 4-1).

Methodological Quality

The methodological quality of the reliability assessment was good as a good sample size was used (50-99) and intra-class correlation coefficients (ICCs) were calculated[70]. The measurement error assessment was fair as information regarding the time interval and test conditions for the measurement error was not provided[70].

Quality Assessment

One study evaluating the reliability of the ODDI was retrieved and the inter- and intra-rater reliability was evaluated along with measurement error[70]. The ICCs for the intra- and inter-rater reliability were high (0.96 and 0.92 respectively) [70]. LoA were calculated and provided however no MIC was defined to allow for interpretation of the measurement error[70].

Cognitive Test Battery

The authors of the pediatric RCT identified in our systematic review used a cognitive test battery, consisting of 14 tests, as their primary outcome measure[71]. This battery aimed to assess attention, learning and memory, perceptual abilities, executive processing, and fine

motor/visuomotor skills[71]. The 14 tests were individual tasks from subscales of measures including the McCarthy Scales of Children's Abilities, the Wide Range Assessment of Memory and Learning and Visuo-Motor Abilities, and the Wechsler Preschool and Primary Scale of Intelligence-Revised. No measurement properties were reported for the individual tests, the measures or for the battery as a whole.

Validity and reliability assessments of a measure incorporate its subscales and the items found within these subscales. The assessments usually apply to the measure as a whole and evaluate how the items and their subscales together contribute to the validity of the measure. The validity of a measure therefore, does not necessarily translate to its subscales and each item as independent entities. Since the authors chose to use individual items from external measures as tasks, the measurement properties of their cognitive test battery should be assessed. As this is the only study known to use this cognitive test battery and studies on measurement properties of individual items are not usually conducted, the battery was not included in the Terwee search filter and could not be included in quality assessments.

Evaluation Tool

One of the pediatric RCTs in the sample identified, developed an evaluation tool specifically for their trial using questions from various other documents[72]. The authors state that the instrument may represent a limitation of their study as it has not been previously used and validated[72]. The tool could therefore not be included in subsequent analyses.

Children and Youth Physical Self-Perception Profile (CY-PSPP)

The CY-PSPP evaluates self-perceptions through self-worth factors distributed across six scales, is validated for use in seventh and eighth graders and is administered by the teachers[73]. The Terwee filter identified one relevant study assessing the measure's validity however did not retrieve the two studies originally cited in the RCT from which

the measure was identified (Table 4-1). The full texts of the two studies cited could not be retrieved.

Methodological Quality

The assessment of structural validity had fair methodological quality as there were minor methodological flaws in the design of the study including lack of information on the rotation method used in the factor analysis and how missing items were handled[73].

Quality Assessment

The study retrieved assessed the structural validity of the CY-PSPP using confirmatory factor analysis[73]. Factor loadings of the individual items were determined however explained variance was not mentioned thus justifying an indeterminate rating[73] (Table 4-3).

The Aberrant Behavior Checklist (ABC)

The ABC is a 58-item behaviour rating scale, completed by parents/caregivers, with five subscales that assess problem behaviours in children three years of age and older[74]. Two relevant studies assessing the ABC's measurement properties were retrieved by the Terwee filter (Table 4-1). The pediatric RCT from which the ABC was identified provided two citations for the measure. One citation is the measure's manual, which the filter would not retrieve, and the other study's full text could not be retrieved.

Methodological Quality

The internal consistency, inter-rater reliability, construct validity, and criterion validity of the ABC were evaluated in the studies retrieved (Table 4-3).

The methodological quality of the measurement properties assessed was either fair or poor (Table 4-2). The assessment of internal consistency had poor methodological quality as the sample size used in the factor analysis was inadequate considering the measure has 58 items, this also resulted in a poor methodological quality for the structural validity

assessment[75]. The reliability assessment had fair methodological quality as Pearson correlation coefficients were calculated rather than the ICC[75].

For construct validity, the hypotheses testing had fair methodological quality as hypotheses were not formulated but expectations could be assumed and how missing items were handled was not described[74]. The criterion validity assessment had poor methodological quality as no correlations were calculated[75].

Quality Assessment

Strong internal consistency was reported for the entire measure as well as for the subscales (>0.80) and analyses confirmed the unidimensionality of the ABC[75]. Interrater reliability was not adequate (<0.70) and therefore received a negative rating[75].

The structural validity of the measure was assessed but the five factor model explained only 31.5% of the variance thus earning a negative rating[75]. Hypotheses testing was assessed through convergent validity of the ABC with the Behaviour Problems Inventory (BPI) however the correlation was <0.5 (0.37)[74]. The other study sought to establish criterion validity of the ABC by examining differences in scores between different psychiatric diagnosed groups[75]. No correlations between the groups were calculated, only means were provided[75].

DISCUSSION

This paper aimed to determine whether the lack of reporting of measurement properties in a sample of pediatric RCTs is justified or whether measurement properties have been assessed for the measures and thus, should be reported. Studies evaluating measurement properties were identified for eight of the 11 outcome measures used in the sample of pediatric RCTs. One of the measures used, the GCJ, does not appear to have any assessments of its measurement properties. The two other measures were tools constructed by authors for their RCTs, were unique to the pediatric RCTs in which they

were used, and the measurement properties were not evaluated. Assessments of these measures were therefore not applicable.

None of the eight measures for which measurement properties were evaluated had assessments for all nine properties. However, not all measurement properties are appropriate for every outcome measure. Reliability was the most commonly assessed property while floor and ceiling effects were only evaluated for one measure. Internal consistency received indeterminate ratings for a large proportion of the studies as unidimensionality of the measures was not assessed or referenced. Measurement error was assessed for two measures but neither received positive ratings. Hypotheses testing was the most commonly tested form of construct validity amongst all the included studies. Content validity was only assessed for two of the measures; the CARIFS received a positive rating. Criterion validity was also assessed for two of the eight measures and both received indeterminate ratings. The GMFM received the largest amount of positive ratings for reliability as well as for responsiveness. Responsiveness was only evaluated for three measures, a concerning observation as each measure was used in an RCT to detect change when an intervention was applied.

Where multiple studies evaluated the same measurement property, a variety of ratings were allocated. This renders the selection of measures challenging as it remains unclear if the measures are adequate and applicable. Methodological quality of the studies on measurement properties was commonly assessed as poor or fair. The lack of adequate methodological quality makes the quality of the measurement properties evaluated indeterminate. One excellent rating was allocated for the content validity of the CARIFS. Each measure for which internal consistency was assessed received a poor rating for a proportion of its studies as was the case for a fair rating of reliability. Of particular concern, the three measures for which responsiveness was evaluated all received a poor methodological quality rating. The pediatric RCTs that depended on the performance of these measures to assess change within their trials may not be valid based on the responsiveness quality assessments presented. Trials that rely on outcome measures to

generate their results must ensure that these measures are both valid and reliable so as to enable the confident dissemination of valid results.

The large majority of the studies included were retrieved through the application of the Terwee methodological PubMed highly sensitive search filter. The search filter was developed according to four phases which include the identification of gold standards, the precise and comprehensive selection of search terms and the assessment of both internal and external validity[5]. The Terwee filter does present with some limitations including a small sample size for the comparative gold standard and the restricted settings in which the filter was validated[5]. Based on the experience of applying this filter for this study, we also found that because the search strategy retrieves studies that specifically mention the outcome measure in combination with measurement properties in their titles and abstracts and excludes outcome measure manuals, citations provided in the pediatric RCTs were sometimes not identified. While studies on measurement properties should mention the measure and its measurement properties, reporting standards do not require that these be mentioned specifically within the titles and abstracts and therefore many studies may be published without meeting these criteria. In order to overcome this limitation, reference lists of all studies retrieved were searched for any additional potentially relevant studies and full texts were retrieved when available. Studies retrieved were also compared with studies cited in the sample of pediatric RCTSs reporting the use of the measures and if any were not identified, these too were retrieved. The comprehensive search for potentially relevant studies ensures that the studies on measurement properties included in this study are the most relevant for the outcome measures identified.

While independent duplication of assessments for the quality criteria and the methodological quality is preferable, based on the results of the individual quality assessments, minor incongruities in the quality ratings accorded would not significantly alter the conclusions. This can be concluded as multiple required aspects were lacking or inadequate for each assessment. The Terwee and COSMIN criteria used for the quality assessments are also objective and have been rigorously developed through consensus-based Delphi study[6-9].

Another possible limitation lies in the fact that only studies published in English and discussing the original version of the measures were included. However, it is necessary for the original versions of measures to be valid and reliable prior to their cross-cultural adaptations or translations. If the version from which a measure is translated cannot be considered valid and reliable, the subsequent translation is also negatively affected.

Authors failed to report measurement properties for nine of the 12 outcome measures, including the CGI-I for which assessments have been previously conducted, used in their pediatric RCTs. Although authors provided relevant citations for the measures in populations similar to their study populations, the measurement properties evaluated in these citations should be reported to facilitate the confident interpretation of the trial results and to inform the selection and applicability of measures. The three measures for which no studies were retrieved do not appear to have evaluated measurement properties, validation is therefore recommended. Although measurement properties were evaluated for nine of these measures, the quality assessments indicate that further validation with methodologically rigorous studies is required. In particular, the responsiveness of all the measures needs to be carefully assessed in compliance with design and statistical methods identified in the quality criteria.

We recognize that some of the standards and criteria applied to these studies were made available after the publication of the pediatric RCTs. The pediatric RCTs were published between 2000 and 2010. The Terwee quality criteria were made available in 2007[6], the COSMIN consensus on measurement property terminology was published in 2010[8] and the 4-point rating scale[9] used was made available in 2011. Regardless of the publication timelines, the results of this study indicate that the quality of measurement properties and the studies in which they were evaluated do not meet reasonable standards. Further validation and the subsequent reporting of the validation results are urgently required as pediatric RCTs continue to generate results used in evidence-based practice.

The results of this study can be used to improve current reporting guidelines for trials, like CONSORT (Consolidated Standards of Reporting Trials) and SPIRIT (Standard

Protocol Items for Randomized Trials), such that the reporting of measurement properties and the citations that support these reports is emphasized. Trialists and researchers should use the tools identified in this study to ensure higher standards for the conduct of their trials. With further validation of measures and the subsequent improvement in the reporting of their measurement properties, informed selection of measures will be possible. Trial results can then be used more confidently in knowledge synthesis efforts and evidence-based practice.

CONCLUSION

An outcome measure's measurement properties provide important information on the measure's intended purpose, performance, accuracy and ability to detect a true change. These measurement properties must therefore be reported to ensure that trial results are valid. Although authors use outcome measures and provide citations for the measures, they do not report measurement properties that have been evaluated. While measurement properties have been evaluated, the quality of these properties and the studies in which they are assessed suggest the need for further validation. This study highlights the need for further validation in methodologically rigourous studies and the subsequent reporting of the results of validation for outcome measures used in RCTs.

Outcome Measure	Terwee Sensitive Search Filter Included Studies	Additional included studies retrieved	Studies originally cited in pediatric RCT for the outcome measure
Global Clinical Judgements (GCJ)	Campbell M, Adams P, Small AM et al. (1995), Lithium in hospitalized aggressive children with conduct disorder: a double-blind and placebocontrolled study. *J Am Acad Child Adolesc Psychiatry* 34:445–453 Cueva JE, Overall JE, Small AM, Armeteros JL, Perry R, Campbell M. Carbamazepin in aggressive children with conduct disorder: a double-blind and placebo-controlled study. Malone RP, Delaney MA, Luebbert JF, Cater J, Campbell M (2000), A double-blind placebo-controlled study of lithium in hospitalized aggressive children and adolescents with conduct disorder. *Arch Gen Psychiatry* 57:649–654	Campbell M, Small AM, Green WH et al. (1984), Behavioral efficacy of haloperidol and lithium carbonate: a comparison in hospitalized aggressive children with conduct disorder. *Arch Gen Psychiatry* 41:650–656	Campbell M, Small AM, Green WH et al. (1984), Behavioral efficacy of haloperidol and lithium carbonate: a comparison in hospitalized aggressive children with conduct disorder. *Arch Gen Psychiatry* 41:650–656 Campbell M, Adams P, Small AM et al. (1995), Lithium in hospitalized aggressive children with conduct disorder: a double-blind and placebocontrolled study. *J Am Acad Child Adolesc Psychiatry* 34:445–453 Malone RP, Delaney MA, Luebbert JF, Cater J, Campbell M (2000), A double-blind placebo-controlled study of lithium in hospitalized aggressive children and adolescents with conduct disorder. *Arch Gen Psychiatry* 57:649–654
Canadian Acute Respiratory Infection and "Flu Scale" (CARIFS)	Jacobs B, Young NL, Dick PT, Ipp MM, Dutkowski R, Davies HD…Wang EEL. Canadian Acute Respiratory Illness and Flu Scale (CARIFS): Development of a valid measure for childhood respiratory infections. *J Clin Epidemiology*. 2000; 53: 793-799. Shepperd S, Perera R, Bates S, Jenkinson C, Hood K, Harnden A, Mant D. A children's acute respiratory illness scale (CARIFS) predicted functional severity and family burden. *J Clin Epidemiology*. 2004; 57: 809-814.		Jacobs B, Young NL, Dick PY, et al. Canadian Acute Respiratory Illness and Flu Scale (CARIFS): development of a valid measure for childhood respiratory infections. *J Clin Epidemiol* 2000; **53:**793–99.
Gross Motor Function Measure (GMFM)	Beckung E, Carlsson G, Carlsdotter S, Uvebrant P. The natural history of gross motor development in children with cerebral palsy aged 1 to 15 years. *Developmental Medicine & Child Neurology*. 2007; 49: 751-756. Bjornson K, Graubert C, McLaughlin J, et al. Test-Retest Reliability of the Gross Motor Function Measure in Children with Cerebral Palsy. *Pediatrics*. 1998; 18:51-6. Bjornson KF, Schmale GA, Adamczyk-Foster A, McLaughlin J. The Effect of Dynamic Ankle Foot Orthoses on Function in Children with Cereebral Palsy. *J Pediatr Orthop* 2006; 26:773-776. Bower E, McLellan DL, Arney J, Campbell MJ. A randomised controlled trial of different intensities of physiotherapy and different goal-setting procedures in 44 children with cerebral palsy. *Developmental Medicine & Child Neurology*. 1996; 38:226-237.	Trahan J, Malouin F. Changes in gross motor function measure in children with different types of cerebral palsy: an eight month follow-up study. *Pediatr Phys Ther* 1999; **11:** 12–17.	Russell DJ, Rosenbaum PL, Cadman DT, Gowland C, Hardy S, Jarvis S. The Gross Motor Function Measure: a means to evaluate the effects of physical therapy. *Develop Med Child Neurol* 1989; **31:** 341–52. Nordmark E, Hagglund G, Jarnlo GB. Reliability of the gross motor function measure in cerebral palsy. *Scand J Rehab Med* 1997; **29:** 25–28. Trahan J, Malouin F. Changes in gross motor function measure in children with different types of cerebral palsy: an eight month follow-up study. *Pediatr Phys Ther* 1999; **11:** 12–17.

	Cassady RL, Nichols-Larsen DS. The effect of hippotherapy on ten children with cerebral palsy. *Pediatr Phys Ther* 2004; 16:165-172. Champagne D, Dugas C. Improving gross motor function and postural control with hippotherapy in children with Down syndrome: case reports. *Physiotherapy Theory and Practice*. 2010; 26(8):564-571. Damiano DL, Gilgannon MD, Abel MF. Responsiveness and uniqueness of the pediatric outcomes data collection instrument compared to the gross motor function measure for measuring orthopaedic and neurosurgical outcomes in cerebral palsy. *J Pediatr Orthop*. 2005; 25:641-645. Footer CB. The effects of therapeutic taping on gross motor function in children with cerebral palsy. *Pediatr Phys Ther*. 2006; 18:245-252. Goh HT, Thompson M, Huang WB, Schafer S. Relationships among measures of knee musculoskeletal impairments, gross motor function, and walking efficiency in children with cerebral palsy. *Pediatr Phys Ther*. 2006; 18:253-261. Hamill D, Washington K, White OR. The effect of hippotherapy on postural control in sitting for children with cerebral palsy. *Physical & Occupational Therapy in Pediatrics*. 2007; 27(4):23-42. Iannacone ST, Hynan LS, AmSMART Group. Reliability of 4 outcome measures in pediatric spinal muscular atrophy. *Arch Neurol.* 2003; 60:1130-1136. Kaufmann P, McDermott MP, Darras BT, Finkel R, Kang P, Oskoui M, Constantinescu A, et al. Observational study of spinal muscular atrophy type 2 and 3. *Arch Neurol.* 2011; 68(6):779-786. Lacey DJ, Stolfi A, Pilati LE. Effects of hyperbaric oxygen on motor function in children with cerebral palsy. *Ann Neurol*. 2012; 72:695-703. LaForme AC, Effgen SK, Page J, Shasby S. Effect of sensorimotor groups on gross motor acquisition for young children with down syndrome. *Pediatr Phys Ther*. 2009;		

	21:158-166. Josenby AL, Jarnlo GB, Gummesson C, Nordmark E. Longitudinal construct validity of the GMFM-88 total score and goal total score and the GMFM-66 score in a 5-year follow up study. *Phys Ther*. 2009; 89:342-350. McCarthy ML, Silberstein CE, Atkins EA, Harryman SE, Sponseller PD, Hadley-Miller NA. Comparing reliability and validity of pediatric instruments for measuring health and well-being of children with spastic cerebral palsy. *Developmental Medicine & Child Neurology*. 2002; 44:468-476. McLaughlin JF, Bjornson KF, Astley SJ, Graubert C, Hays RM, Roberts TS, Price R, et al. Selective dorsal rhizotomy: efficacy and safety in an investigator-masked randomized clinical trial. *Developmental Medicine & Child Neurology*. 1998; 40:220-232. Nelson L, Owens H, Hynan LS, Iannacone ST, AmSMART Group. The gross motor function measure is a valid and sensitive outcome measure for spinal muscular atrophy. *Neuromuscular Disorders* 2006; 16:374-380. Nordmark E, Jarnlo GB, Hagglund G. Comparison of the gross motor function measure and pediatric evaluation of disability inventory in assessing motor function in children undergoing selective dorsal rhizotomy. *Developmental Medicine & Child Neurology*. 2000; 42:245-252. Palisano RJ, Walter SD, Russell DJ, Rosenbaum PL, Gemus M, Galuppi BE, Cunningham L. Gross motor function of children with down syndrome: creation of motor growth curves. *Arch Phys Med Rehabil*. 2001; 82:494-500. Ruck-Gibis J, Plotkin H, Hanley J, Wood-Dauphinee S. Reliability of the gross motor function measure for children with osteogenesis imperfecta. *Pediatr Phys Ther* 2001; 13:10-17. Russell DJ, Rosenbaum PL, Cadman DT, Gowland C, Hardy S, Jarvis S. The gross motor function measure: a means to evaluate the effects of physical therapy.		

	Developmental & Child Neurology. 1989; 31: 341-352. Russell DJ, Rosenbaum PL, Lane M, Gowland C, Goldsmith CH, Boyce WF, Plews N. Training users in the gross motor function measure: methodological and practical issues. *Phys Ther*. 1994; 74:630-636. Russell D, Palisano R, Walter S, Rosenbaum P, Gemus M, Gowland C, Galuppi B, et al. Evaluating motor function in chidlren with Down syndrome: validity of the GMFM. *Developmental Medicine & Child Neurology*. 1998; 40: 693-701. Russell DJ, Gorter JW. Assessing functional differences in gross motor skills in children with cerebral palsy who use an ambulatory aid or orthoses: can the GMFM-88 help? *Developmental & Child Neurology*. 2005; 47: 462-467. Schreiber J. Increased intensity of physical therapy for a child with gross motor developmental delay: a case report. *Physical & Occupational Therapy in Pediatrics*. 2004; 24(4):63-78. Sterba JA, Rogers BT, France AP, Vokes DA. Horseback riding in children with cerebral palsy: effect on gross motor function. *Developmental Medicine & Child Neurology*. 2002; 44: 301-308. Sterba JA. Adaptive downhill skiing in children with cerebral palsy: effect on gross motor function. *Pediatr Phys Ther*. 2006; 18:289-296. Thomas-Stonell N, Johnson P, Rumney P, Wright V, Oddson B. An evaluation of the responsiveness of a comprehensive set of outcome measures for children and adolescents with traumatic brain injuries. *Pediatric Rehabilitation.* 2006; 9(1):14-23. Vos-Vromans DCWM, Ketelaar M, Gorter JW. Responsiveness of evaluative measures for children with cerebral palsy: the gross motor function measure and the pediatric evaluation of disability inventory. *Disability and Rehabilitation*. 2005; 27(20):1245-1252. Wang HY, Yang H. Evaluating the responsiveness of 2		

	versions of the gross motor function measure for children with cerebral palsy. *Arch Phys Med Rehabil*. 2006; 87:51-56.		
Gross Motor Function Classification System (GMFCS)	Beckung E, Hagberg G. Correlation between ICIDH handicap code and gross motor function classification system in children with cerebral palsy. *Developmental Medicine & Child Neurology*. 2000; 42:669-673. Benedict RE, Patz J, Maenner MJ, Arneson CL, Yeargin-Allsopp M, Doernberg NS, et al. Feasibility and reliability of classifying gross motor function among children with cerebral palsy using population-based record surveillance. *Paediatric and Perinatal Epidemiology*. 2010; 25:88-96. Bodkin AW, Robinson C, Perales FP. Reliability and validity of the gross motor function classification system for cerebral palsy. *Pediatr Phys Ther*. 2003; 15:247-252. Gunel MK, Mutlu A, Tarsuslu T, Livanelioglu A. Relationship among the manual ability classification system (MACS), the gross motor function classification system (GMFCS), and the functional status (WeeFIM) in children with spastic cerebral palsy. *Eur J Pediatr*. 2009; 168: 477-485. Hidecker MJC, Ho NT, Dodge N, Hurvitz EA, Slaughter J, Workinger MS, Kent RD, et al. Inter-relationships of functional status in cerebral palsy: analyzing gross motor function, manual ability, and communication function classification systems in children. *Developmental Medicine & Child Neurology*. 2012; 54:737-742. Imms C, Carlin J, Eliasson AC. Stability of caregiver-reported manual ability and gross motor function classifications of cerebral palsy. *Developmental Medicine & Child Neurology*. 2010; 52:153-159. Jahnsen R, Aamodt G, Rosenbaum P. Gross motor classification system used in adults with cerebral palsy: agreement of self-reported versus professional rating. *Developmental Medicine & Child Neurology*. 2006; 48:734-738. McCormick A, Brien M, Plourde J, Wood E, Rosenbaum P, McLean J. Stability of the gross motor function classification system in adults with cerebral palsy.	Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997; 39:214-23.	Palisano RJ, Hanna SE, Rosenbaum PL, et al. Validation of a model of gross motor function for children with cerebral palsy. Phys Ther 2000;80:974-85. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997; 39:214-23.

	Developmental Medicine & Child Neurology. 2007; 49:265-269. McDowell BC, Kerr C, Parkes J. Interobserver agreement of the gross motor function classification system in an ambulant population of children with cerebral palsy. *Developmental Medicine & Child Neurology*. 2007; 49:528-533. Morris C, Galuppi BE, Rosenbaum PL. Reliability of family report for the gross motor function classification system. *Developmental Medicine & Child Neurology*. 2004; 46: 455-460. Morris C, Kurinczuk JJ, Fitzpatrick R, Rosenbaum PL. Who best to make assessments? Professionals' and families' classifications of gross motor function in cerebral palsy are highly consistent. *Arch Dis Child*. 2006; 91:675-679. Palisano RJ, Cameron D, Rosenbaum PL, Walter SD, Russell D. Stability of the gross motor function classification system. *Developmental Medicine & Child Neurology*. 2006; 48:424-428. Wood E, Rosenbaum P. The gross motor function classification system for cerebral palsy: a study of reliability and stability over time. *Developmental Medicine & Child Neurology*. 2000; 42:292-296.		
Mental Developmental Index of the Bayley Scales of Infant Development II (BSID-II)	Hack M, Taylor G, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. Pediatrics 2005;116:333-41		Bayley N. Manual for the Bayley Scales of Infant Development. 2nd ed. San Antonio, TX: Psychological Corporation, 1993. Hack M, Taylor G, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. Pediatrics 2005;116:333-41
Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)	Franklin ME, Sapyta J, Freeman JB, Khanna M, Compton S, Almirall D, Moore P, et al. Cognitive behaviour therapy augmentation of pharmacotherapy in pediatric obsessive-compulsive disorder. *JAMA.* 2001; 306(11):1224-1232. Freeman J, Flessner CA. The children's yale-brown obsessive compulsive scale: reliability and validity for use among 5 to 8 year olds with obsessive-compulsive		Scahill L, Riddle MA, McSwiggin-Hardin M et al. (1997), Children's Yale- Brown Obsessive Compulsive Scale: reliability and validity. *J Am Acad Child Adolesc Psychiatry* 36:844–852

	disorder. *J Abnorm Child Psychol*. 2011; 39:877-883. Hanna GL. Demographic and clinical features of obsessive-compulsive disorder in children and adolescents. *J Am Acad Child Adoelsc Psychiatry*. 1995; 34(1):19-27. McKay D, Piacentini J, Greisberg S, Graae F, Jaffer M, Miller J, Neziroglu F, et al. The children's yale-brown obsessive-compulsive scale: item structure in an outpatient setting. *Psychological Assessment*. 2003; 15(4): 578-581. Peris TS, Bergman L, Langley A, Chang S, McCracken JT, Piacentini J. Correlates of accommodation of pediatric obsessive-compulsive disorder: parent, child, and family characteristics. *J Am Acad Child Adolesc Psychitatry*. 2008; 47(10):1173-1181. Peris TS, Sugar CA, Bergman L, Chang S, Langley A, Piacentini J. Family factors predict treatment outcome for pediatric obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology*. 2012; 80(2):225-263. Scahill L, Riddle MA, McSwiggin-Hardin M, Sharon I, King, RA, Goodman WK, Cicchetti D, et al. Children's yale-brown obsessive compulsive scale: reliability and validity. *J Am Acad Child Adoles Psychiatry*. 1997; 36(6):844-852. Storch EA, Murphy TK, Geffken GR, Soto O, Sajid M, Allen P, Roberti JW, Killiany EM, Goodman WK. Psychometric evaluation of the children's yale-brown obsessive-compulsive scale. *Psychiatry Research*. 2004; 129:91-98. Storch EA, Murphy TK, Geffken GR, Bagner DM, Soto O, Sajid M, Allen P, et al. Factor analystic study of the children's yale-brown obsessive-compulsive scale. *J Am Acad Child Adolesc Psychiatry*. 2005; 34(2):312-319.		
Oblique Diameter Difference Index (ODDI)	van Vlimmeren LA, Takken T, van Adrichem LN, van der Graaf Y, Helders PJ, Engelbert RH. Plagiocephalometry: a non-invasive method to quantify asymmetry of the skull; a reliability study. *Eur J Pediatr*.		van Vlimmeren LA, Takken T, van Adrichem LN, van der Graaf Y, Helders PJ, Engelbert RH. Plagiocephalometry: a non-invasive method to quantify asymmetry of the skull; a reliability study. *Eur J Pediatr*. 2006;165(3):149-157

	2006;165(3):149-157		
Children and Youth Physical Self-Perception Profile (CYPSPP)	Eklund RC, Whitehead JR, Welk GJ. Validity of the children and youth physical self-perception profile: a confirmatory factor analysis. *Research Quaterly for Exercise and Sport*. 1987; 68(3):249-256.		Whitehead JR. A study of children's physical self-perceptions using an adapted physical self-perception profile questionnaire. *Pediatr Exerc Sci.* 1995;7:132–151 Biddle S, Page A, Ashford B, et al. Assessment of children's physical self-perceptions. *Int J Adolesc Youth.* 1993;4:93–109
Aberrant Behaviour Checklist (ABC)	Hill J, Powlitch S, Furniss F. Convergent validity of the aberrant behaviour checklist and behaviour problems inventory with people with complex needs. *Research in Developmental Disabilities*. 2008; 29:45-60. Rojahn J. The aberrant behaviour checklist with children and adolescents with dual diagnosis. *Journal of Autism and Developmental Disorders*. 1991; 21(1):17-28.		Aman MG, Singh NN, Stewart AW, Field CJ. The aberrant behaviour checklist: a behavior rating scale for the assessment of treatment effects.*Am J Ment Defic.* 1985;89:485–491 Aman MG, Singh NN. *Aberrant Behavior Checklist Manual*. East Aurora, NY: Slosson Educational Publications; 1986

Table 4-1 Studies retrieved from the Terwee Methodological Highly Sensitive Search Filter

Outcome Measure	Reliability			Validity			Responsiveness
	Internal Consistency	Measurement error	Reliability	Content	Construct S HT	Criterion	Responsiveness
Canadian Acute Respiratory Illness and Flu Scale (CARIFS)	Poor [11, 12]		Fair [11]	Excellent [11] Poor [12]	Poor [11, 12]		Poor [11, 12]
Gross Motor Function Measure (GMFM)	Poor [29]	Poor [17]	Good [33, 35 36] Fair [14, 24, 25, 30] Poor [15, 16 18, 19 21, 22 23, 26 27, 30 34, 37, 39, 40 41, 44]		Poor [18, 22 31] Fair [29]		Fair [20, 35] Poor [28, 32, 37, 38, 42, 43, 44, 45]
Gross Motor Classification System			Fair [48, 53 56, 57 59] Good [47, 51, 52, 54 55] Poor [58]	Poor [59]	Fair [46] Poor [48, 49 50, 51]	Fair [48]	
Mental Developmental Index of the Bayley Scales of Infant Development II (BSID-II)					Poor [60]		
Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)	Poor [62, 65 66, 67 68] Fair [64]		Poor [61, 66 67] Fair [62, 68]		Fair [64] Good [69] Fair [62, 68 69] Poor [63, 67]		Poor [62, 67]

	Good [69]		
Oblique Diameter Difference Index (ODDI)	Fair [70] Good [70]		
Children and Youth Physical Self-Perception Profile (CY-PSPP)		Fair [73]	
Aberrant Behavior Checklist (ABC)	Poor [75] Fair [75]	Poor [75] Fair [74] Poor [75]	

[..] = reference number

Table 4-2 Methodological Quality of Studies using COSMIN Scoring System

Outcome Measure	Reliability			Validity				Responsiveness	Floor or ceiling effect	Interpretability
	Internal Consistency	Measurement error	Reliability	Content	Construct S	Construct HT	Criterion	Responsiveness		
Global Clinical Judgements (GCJ)	0	0	0	0	0	0	0	0	0	0
Canadian Acute Respiratory Infection and "Flu Scale" (CARIFS)	? [11,12]	0	+ [11]	+ [11,12]	0	- [11, 12]	0	+ [11] ? [12]	0	? [11]
Gross Motor Function Measure (GMFM)	? [29]	- [17]	+ [14, 15, 18 19, 21, 22 24, 25, 26 27, 30, 33 34, 35, 36 37, 40, 41 44] ? [15, 16, 23 30, 39]	0	0	+ [18, 22, 29, 31]	0	+ [20, 35, 43, 44] ? [28, 32, 37, 38, 42, 45]	+ [28] - [29, 43]	+ [17] ? [25, 28, 29, 31, 38, 41]
Gross Motor Function Classification System (GMFCS)	0	0	+ [47, 48, 51 52, 53, 54 55, 56, 57 58] - [59]	+ [59]	0	+ [46, 49 50, 51] ? [48]	? [48]	0	0	0
Mental Development Index of the Bayley Scales of Infant Development II (MDI-BSID-II)	0	0	0	0	0	- [60]	0	0	0	0
Children's Yale-Brown Obsessive-Compulsive Scale	? [62, 65 66, 67 68]	0	+ [61, 62 66, 67 68]	0	? [64] + [69]	+ [62, 63 67, 68] - [69]	0	? [62, 67]	0	0

(CY-BOCS)	+ [64] - [69]								
Oblique Diameter Difference Index (ODDI)	0	? [70]	+ [70]	0	0	0	0	0	0	0
Children and Youth Physical Self-Perception Profile (CY-PSPP)	0	0	0	0	? [73]	0	0	0	0	0
Aberrant Behaviour Checklist (ABC)	+ [75]	0	- [75]	0	- [75]	- [74]	? [75]	0	0	0

[..] = reference number

Table 4-3 Quality Assessment of Measurement Properties using modified version of Terwee's Quality Criteria

Appendix 4-1 – Search Strategy

1. (instrumentation or methods).sh.
2. (Validation Studies or Comparative Study).pt.
3. exp Psychometrics/
4. psychometr*.ti,ab.
5. (clinimetr* or clinometr*).tw.
6. exp "Outcome Assessment (Health Care)"/
7. outcome assessment.ti,ab.
8. outcome measure*.tw.
9. exp Observer Variation/
10. observer variation.ti,ab.
11. exp Health Status Indicators/
12. exp "Reproducibility of Results"/
13. reproducib*.ti,ab.
14. exp Discriminant Analysis/
15. (reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or "internal consistency").ti,ab.
16. (cronbach* and (alpha or alphas)).ti,ab.
17. (item and (correlation* or selection* or reduction*)).ti,ab.
18. (agreement or precision or imprecision or "precise values" or test-retest).ti,ab.
19. (test and retest).ti,ab.
20. (reliab* and (test or retest)).ti,ab.
21. (stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intratester or intra-tester or interobserver or inter-observer or intraobserver or intra-observer or intertechnician or inter-technician or intratechnician or intra-technician or interexaminer or inter-examiner or intraexaminer or intra-examiner or interassay or inter-assay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-individual or interparticipant or inter-participant or intraparticipant or intra-participant or kappa or kappa's or kappas or repeatab*).ti,ab.
22. ((replicab* or repeated) and (measure or measures or findings or result or results or test or tests)).ti,ab.
23. (generaliza* or generalisa* or concordance).ti,ab.
24. (intraclass and correlation*).ti,ab.
25. (discriminative or "known group" or factor analysis or factor analyses or dimension* or subscale*).ti,ab.
26. (multitrait and scaling and (analysis or analyses)).ti,ab.
27. (item discriminant or interscale correlation* or error or errors or "individual variability").ti,ab.
28. (variability and (analysis or values)).ti,ab.
29. (uncertainty and (measurement or measuring)).ti,ab.
30. ("standard error of measurement" or sensitiv* or responsive*).ti,ab.
31. ((minimal or minimally or clinical or clinically) and (important or significant or detectable) and (change or difference)).ti,ab.
32. (small* and (real or detectable) and (change or difference)).ti,ab.

33. (meaningful change or "ceiling effect" or "floor effect" or "Item response model" or IRT or Rasch or "Differential item functioning" or DIF or "computer adaptive testing" or "item bank" or "cross-cultural equivalence").ti,ab.
34. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33
35. (child* or pediatric* or infan* or neonat* or newborn* or teen* or youth*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
36. 34 and 35

Outcome Measure	Search Terms
Global Clinical Judgements (GCJ)	37. Global Clinical Judgments.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. Global Clinical Judgments Rating.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. GCJ.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. Global Clinical Judgments Scale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. Global Clinical Judgments Consensus Scale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 42. Global Clinical Consensus Rating.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 43. Global Clinical Consensus Scale.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 44. consensus CGJ rating.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 45. CGJ rating.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 46. 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 47. 36 and 46
Canadian Acute Respiratory Illness and Flu Scale (CARIFS)	37. (Canadian Acute Respiratory Illness and Flu Scale).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. (Canadian Acute Respiratory Illness and Flu Scale Score).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. CARIFS.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. CARIFS score.mp. [mp=title, abstract, original title, name of substance word, subject

	heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. 37 or 38 or 39 or 40 42. 36 and 41
Gross Motor Function Measure (GMFM)	37. Gross Motor Function Measure.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. GMFM.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39
Gross Motor Function Classification System (GMFCS)	37. Gross Motor Function Classification System.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. GMFCS.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39
Mental Developmental Index of the Bayley Scales of Infant Development II (BSID-II)	37. Mental Development Index on the Bayley Scales of Infant Development II.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. Bayley Scales of Infant Development II.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. BSID-II.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. Mental Development Index.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. Mental Development Index - Bayley Scales of Infant Development II.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 42. 37 or 38 or 39 or 40 or 41 44. 36 and 42
Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)	37. (The Children and Youth Physical Self-Perception Profile).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. CY-PSPP.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39
Oblique Diameter Difference Index (ODDI)	37. Oblique Diameter Difference Index.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. Oblique Diameter Difference Index Score.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol

	supplementary concept, rare disease supplementary concept, unique identifier] 39. ODDI.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. ODDI score.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 41. 37 or 38 or 39 or 40 42. 36 and 41
Children and Youth Physical Self-Perception Profile (CY-PSPP)	37. (The Children and Youth Physical Self-Perception Profile).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. CY-PSPP.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. 37 or 38 40. 36 and 39
Aberrant Behaviour Checklist (ABC)	37. Aberrant Behavior Checklist.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 38. ABC.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 39. ABC Checklist.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier] 40. 37 or 38 or 39 41. 36 and 40

Appendix 4-2: Quality criteria for measurement properties (Based on Terwee et al 2007)

Property (definitions are based on COSMIN taxonomy)			Rating	Quality Criteria
Reliability: The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions	Internal consistency: The degree of the interrelatedness among the items		+ ? -	Sub)scale unidimensional AND Cronbach's alpha(s) ≥ 0.70 Dimensionality not known OR Cronbach's alpha not determined (Sub)scale not unidimensional OR Cronbach's alpha(s) < 0.70
	Measurement error: The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured		+ ? -	MIC > SDC OR MIC outside the LOA MIC not defined MIC ≤ SDC OR MIC equals or inside LOA
	Reliability: The proportion of the total variance in the measurements which is due to 'true' differences between patients		+ ? -	ICC/weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80 Neither ICC/weighted Kappa, nor Pearson's r determined ICC/weighted Kappa < 0.70 OR Pearson's r < 0.80
Validity: The degree to which an instrument measures the construct(s) it purports to measure	Content validity: The degree to which the content of an instrument is an adequate reflection of the construct to be measured		+ ? -	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be completew No target population involvement The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
	Construct validity: The degree to which the scores of an instrument are consistent with hypotheses	Structural: The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be	+ ? -	Factors should explain at least 50% of the variance Explained variance not mentioned Factors explain < 50% of the variance

		measured		
		Hypothesis testing:	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses) AND correlation with related constructs is higher than with unrelated constructs;
			?	Solely correlations determined with unrelated constructs;
			-	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
	Criterion validity: The degree to which the scores of an instrument are an adequate reflection of a 'gold standard'		+	Convincing arguments that gold standard is ''gold'' AND correlation with gold standard >0.70;
			?	No convincing arguments that gold standard is ''gold'' OR doubtful design or method;
			-	Correlation with gold standard <0.70, despite adequate design and method;
Responsiveness: The ability of an instrument to detect change over time in the construct to be measured			+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70) AND correlation with related constructs is higher than with unrelated constructs

	? - 	Solely correlations determined with unrelated constructs Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlation with related constructs is lower than with unrelated constructs
Floor and ceiling effect: The number of respondents who achieved the lowest or highest possible score	+ ? -	<15% of the respondents achieved the highest or lowest possible scores; Doubtful design or method; >15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods;
Interpretability : The degree to which one can assign qualitative meaning to quantitative scores	+ ? 	Mean and SD scores presented of at least four relevant subgroups of patients and MIC or MID defined; Doubtful design or method OR less than four subgroups OR no MIC or MID defined;

MIC = minimal important change, ; MID = minimal important difference; SDC = smallest detectable change,
LOA = limits of agreement, ICC = intraclass correlation coefficient,
AUC = area under the curve
+ = positive rating, ? = indeterminate rating, - = negative rating, 0 = no information

Appendix 4-3 COSMIN Checklist with 4-point rating scale (example)

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	**Good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

References

[1] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737-745.

[2] Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement Issues in Trials of Pediatric Acute Diarrheal Diseases: A Systematic Review. *Pediatrics* 2010;126;e222-e231.

[3] Squires JE, Estabrooks CA, O'Rourke HM, Gustavsson P, Newburn-Cook CV, Wallin L. A systematic review of the psychometric properties of self-report research utilization measures used in healthcare. *Science* 2011;6:83.

[4] Reid GT, Walter FM, Brisbane JM, Emery JD. Family History Questionnaires Designed for Clinical Use: A Systematic Review. *Public Health Genomics* 2009; 12:73-83.

[5] 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;18:1115-1123.

[6] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, and de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34-42.

[7] Schellingerhout JM, Verhagen AP Heymans MW, , de Vet HC, Koes BW, and Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: A systematic review. *BMC Med Res Methodol* 2011; 11: 87.

[8] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, and de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Qual Life Res* 2010; 19: 539-549.

[9] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2011; 21:651-657.

[10] Kafantaris V, Coletti DJ, Dicker R, Padula G, Pleak RR, Alvir JMJ et al. Lithium Treatment of Acute Mania in Adolescents: A Placebo-Controlled Discontinuation Study. *J Am. Acad. Child Adolesc. Psychiatry.* 2004;43(8):984-93.

[11] Jacobs B, Young NL, Dick PY, et al. Canadian Acute Respiratory Illness and Flu Scale (CARIFS): development of a valid measure for childhood respiratory infections. *J Clin Epidemiol* 2000; 53:793–99.

[12] Shepperd S, Perera R, Bates S, Jenkinson C, Hood K, Harnden A, Mant D. A children's acute respiratory illness scale (CARIFS) predicted functional severity and family burden. *J Clin Epidemiology*. 2004; 57: 809-814.

[13] Avery LM, Russell DJ, Raina PS, Walter SD, Rosenbaum PL. Rasch Analysis of the gross motor function measure: validating the assumptions of the Rasch model to create an interval-level measure. *Arch Phys Med Rehabil*. 2003; 84:697-705.

[14] Beckung E, Carlsson G, Carlsdotter S, Uvebrant P. The natural history of gross motor development in children with cerebral palsy aged 1 to 15 years. *Developmental Medicine & Child Neurology*. 2007; 49: 751-756.

[15] Bjornson K, Graubert C, McLaughlin J, et al. Test-Retest Reliability of the Gross Motor Function Measure in Children with Cerebral Palsy. *Pediatrics*. 1998; 18:51-6.

[16] Bjornson KF, Schmale GA, Adamczyk-Foster A, McLaughlin J. The Effect of Dynamic Ankle Foot Orthoses on Function in Children with Cereebral Palsy. *J Pediatr Orthop* 2006; 26:773-776.

[17] Bower E, McLellan DL, Arney J, Campbell MJ. A randomised controlled trial of different intensities of physiotherapy and different goal-setting procedures in 44 children with cerebral palsy. *Developmental Medicine & Child Neurology* 1996; 38:226-237.

[18] Cassady RL, Nichols-Larsen DS. The effect of hippotherapy on ten children with cerebral palsy. *Pediatr Phys Ther* 2004; 16:165-172.

[19] Champagne D, Dugas C. Improving gross motor function and postural control with hippotherapy in children with Down syndrome: case reports. *Physiotherapy Theory and Practice*. 2010; 26(8):564-571.

[20] Damiano DL, Gilgannon MD, Abel MF. Responsiveness and uniqueness of the pediatric outcomes data collection instrument compared to the gross motor function measure for measuring orthopaedic and neurosurgical outcomes in cerebral palsy. *J Pediatr Orthop*. 2005; 25:641-645.

[21] Footer CB. The effects of therapeutic taping on gross motor function in children with cerebral palsy. *Pediatr Phys Ther*. 2006; 18:245-252.

[22] Goh HT, Thompson M, Huang WB, Schafer S. Relationships among measures of knee musculoskeletal impairments, gross motor function, and walking efficiency in children with cerebral palsy. *Pediatr Phys Ther*. 2006; 18:253-261.

[23] Hamill D, Washington K, White OR. The effect of hippotherapy on postural control in sitting for children with cerebral palsy. *Physical & Occupational Therapy in Pediatrics*. 2007; 27(4):23-42.

[24] Iannacone ST, Hynan LS, AmSMART Group. Reliability of 4 outcome measures in pediatric spinal muscular atrophy. *Arch Neurol.* 2003; 60:1130-1136.

[25] Kaufmann P, McDermott MP, Darras BT, Finkel R, Kang P, Oskoui M, Constantinescu A, et al. Observational study of spinal muscular atrophy type 2 and 3. *Arch Neurol.* 2011; 68(6):779-786.

[26] Lacey DJ, Stolfi A, Pilati LE. Effects of hyperbaric oxygen on motor function in children with cerebral palsy. *Ann Neurol*. 2012; 72:695-703.

[27] LaForme AC, Effgen SK, Page J, Shasby S. Effect of sensorimotor groups on gross motor acquisition for young children with down syndrome. *Pediatr Phys Ther*. 2009; 21:158-166.

[28] Josenby AL, Jarnlo GB, Gummesson C, Nordmark E. Longitudinal construct validity of the GMFM-88 total score and goal total score and the GMFM-66 score in a 5-year follow up study. *Phys Ther*. 2009; 89:342-350.

[29] McCarthy ML, Silberstein CE, Atkins EA, Harryman SE, Sponseller PD, Hadley-Miller NA. Comparing reliability and validity of pediatric instruments for measuring health and well-being of children with spastic cerebral palsy. *Developmental Medicine & Child Neurology.* 2002; 44:468-476.

[30] McLaughlin JF, Bjornson KF, Astley SJ, Graubert C, Hays RM, Roberts TS, Price R, et al. Selective dorsal rhizotomy: efficacy and safety in an investigator-masked randomized clinical trial. *Developmental Medicine & Child Neurology*. 1998; 40:220-232.

[31] Nelson L, Owens H, Hynan LS, Iannacone ST, AmSMART Group. The gross motor function measure is a valid and sensitive outcome measure for spinal muscular atrophy. *Neuromuscular Disorders* 2006; 16:374-380.

[32] Nordmark E, Jarnlo GB, Hagglund G. Comparison of the gross motor function measure and pediatric evaluation of disability inventory in assessing motor function in children undergoing selective dorsal rhizotomy. *Developmental Medicine & Child Neurology*. 2000; 42:245-252.

[33] Palisano RJ, Walter SD, Russell DJ, Rosenbaum PL, Gemus M, Galuppi BE, Cunningham L. Gross motor function of children with down syndrome: creation of motor growth curves. *Arch Phys Med Rehabil*. 2001; 82:494-500.

[34] Ruck-Gibis J, Plotkin H, Hanley J, Wood-Dauphinee S. Reliability of the gross motor function measure for children with osteogenesis imperfecta. *Pediatr Phys Ther* 2001; 13:10-17.

[35] Russell DJ, Rosenbaum PL, Cadman DT, Gowland C, Hardy S, Jarvis S. The gross motor function measure: a means to evaluate the effects of physical therapy. *Developmental & Child Neurology*. 1989; 31: 341-352.

[36] Russell DJ, Rosenbaum PL, Lane M, Gowland C, Goldsmith CH, Boyce WF, Plews N. Training users in the gross motor function measure: methodological and practical issues. *Phys Ther*. 1994; 74:630-636.

[37] Russell D, Palisano R, Walter S, Rosenbaum P, Gemus M, Gowland C, Galuppi B, et al. Evaluating motor function in chidlren with Down syndrome: validity of the GMFM. *Developmental Medicine & Child Neurology*. 1998; 40: 693-701.

[38] Russell DJ, Gorter JW. Assessing functional differences in gross motor skills in children with cerebral palsy who use an ambulatory aid or orthoses: can the GMFM-88 help? *Developmental & Child Neurology*. 2005; 47: 462-467.

[39] Schreiber J. Increased intensity of physical therapy for a child with gross motor developmental delay: a case report. *Physical & Occupational Therapy in Pediatrics*. 2004; 24(4):63-78.

[40] Sterba JA, Rogers BT, France AP, Vokes DA. Horseback riding in children with cerebral palsy: effect on gross motor function. *Developmental Medicine & Child Neurology*. 2002; 44: 301-308.

[41] Sterba JA. Adaptive downhill skiing in children with cerebral palsy: effect on gross motor function. *Pediatr Phys Ther*. 2006; 18:289-296.

[42] Thomas-Stonell N, Johnson P, Rumney P, Wright V, Oddson B. An evaluation of the responsiveness of a comprehensive set of outcome measures for children and adolescents with traumatic brain injuries. *Pediatric Rehabilitation.* 2006; 9(1):14-23.

[43] Vos-Vromans DCWM, Ketelaar M, Gorter JW. Responsiveness of evaluative measures for children with cerebral palsy: the gross motor function measure and the pediatric evaluation of disability inventory. *Disability and Rehabilitation*. 2005; 27(20):1245-1252.

[44] Wang HY, Yang H. Evaluating the responsiveness of 2 versions of the gross motor function measure for children with cerebral palsy. *Arch Phys Med Rehabil*. 2006; 87:51-56.

[45] Trahan J, Malouin F. Changes in gross motor function measure in children with different types of cerebral palsy: an eight month follow-up study. *Pediatr Phys Ther* 1999; 11: 12–17.

[46] Beckung E, Hagberg G. Correlation between ICIDH handicap code and gross motor function classification system in children with cerebral palsy. *Developmental Medicine & Child Neurology*. 2000; 42:669-673.

[47] Benedict RE, Patz J, Maenner MJ, Arneson CL, Yeargin-Allsopp M, Doernberg NS, et al. Feasibility and reliability of classifying gross motor function among children with cerebral palsy using population-based record surveillance. *Paediatric and Perinatal Epidemiology*. 2010; 25:88-96.

[48] Bodkin AW, Robinson C, Perales FP. Reliability and validity of the gross motor function classification system for cerebral palsy. *Pediatr Phys Ther*. 2003; 15:247-252.

[49] Gunel MK, Mutlu A, Tarsuslu T, Livanelioglu A. Relationship among the manual ability classification system (MACS), the gross motor function classification system (GMFCS), and the functional status (WeeFIM) in children with spastic cerebral palsy. *Eur J Pediatr*. 2009; 168: 477-485.

[50] Hidecker MJC, Ho NT, Dodge N, Hurvitz EA, Slaughter J, Workinger MS, Kent RD, et al. Inter-relationships of functional status in cerebral palsy: analyzing gross motor function, manual ability, and communication function classification systems in children. *Developmental Medicine & Child Neurology*. 2012; 54:737-742.

[51] Imms C, Carlin J, Eliasson AC. Stability of caregiver-reported manual ability and gross motor function classifications of cerebral palsy. *Developmental Medicine & Child Neurology*. 2010; 52:153-159.

[52] Jahnsen R, Aamodt G, Rosenbaum P. Gross motor classification system used in adults with cerebral palsy: agreement of self-reported versus professional rating. *Developmental Medicine & Child Neurology*. 2006; 48:734-738.

[53] McCormick A, Brien M, Plourde J, Wood E, Rosenbaum P, McLean J. Stability of the gross motor function classification system in adults with cerebral palsy. *Developmental Medicine & Child Neurology*. 2007; 49:265-269.

[54] McDowell BC, Kerr C, Parkes J. Interobserver agreement of the gross motor function classification system in an ambulant population of children with cerebral palsy. *Developmental Medicine & Child Neurology*. 2007; 49:528-533.

[55] Morris C, Galuppi BE, Rosenbaum PL. Reliability of family report for the gross motor function classification system. *Developmental Medicine & Child Neurology*. 2004; 46: 455-460.

[56] Morris C, Kurinczuk JJ, Fitzpatrick R, Rosenbaum PL. Who best to make assessments? Professionals' and families' classifications of gross motor function in cerebral palsy are highly consistent. *Arch Dis Child*. 2006; 91:675-679.

[57] Palisano RJ, Cameron D, Rosenbaum PL, Walter SD, Russell D. Stability of the gross motor function classification system. *Developmental Medicine & Child Neurology*. 2006; 48:424-428.

[58] Wood E, Rosenbaum P. The gross motor function classification system for cerebral palsy: a study of reliability and stability over time. *Developmental Medicine & Child Neurology*. 2000; 42:292-296.

[59] Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev Med Child Neurol* 1997; 39:214-23.

[60] Hack M, Taylor G, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. *Pediatrics* 2005;116:333-41

[61] Franklin ME, Sapyta J, Freeman JB, Khanna M, Compton S, Almirall D, Moore P, et al. Cognitive behaviour therapy augmentation of pharmacotherapy in pediatric obsessive-compulsive disorder. *JAMA* 2001; 306(11):1224-1232.

[62] Freeman J, Flessner CA. The children's yale-brown obsessive compulsive scale: reliability and validity for use among 5 to 8 year olds with obsessive-compulsive disorder. *J Abnorm Child Psychol*. 2011; 39:877-883.

[63] Hanna GL. Demographic and clinical features of obsessive-compulsive disorder in children and adolescents. *J Am Acad Child Adoelsc Psychiatry*. 1995; 34(1):19-27.

[64] McKay D, Piacentini J, Greisberg S, Graae F, Jaffer M, Miller J, Neziroglu F, et al. The children's yale-brown obsessive-compulsive scale: item structure in an outpatient setting. *Psychological Assessment*. 2003; 15(4): 578-581.

[65] Peris TS, Bergman L, Langley A, Chang S, McCracken JT, Piacentini J. Correlates of accommodation of pediatric obsessive-compulsive disorder: parent, child, and

family characteristics. *J Am Acad Child Adolesc Psychitatry*. 2008; 47(10):1173-1181.

[66] Peris TS, Sugar CA, Bergman L, Chang S, Langley A, Piacentini J. Family factors predict treatment outcome for pediatric obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology*. 2012; 80(2):225-263.

[67] Scahill L, Riddle MA, McSwiggin-Hardin M, Sharon I, King, RA, Goodman WK, Cicchetti D, et al. Children's yale-brown obsessive compulsive scale: reliability and validity. *J Am Acad Child Adoles Psychiatry*. 1997; 36(6):844-852.

[68] Storch EA, Murphy TK, Geffken GR, Soto O, Sajid M, Allen P, Roberti JW, Killiany EM, Goodman WK. Psychometric evaluation of the children's yale-brown obsessive-compulsive scale. *Psychiatry Research*. 2004; 129:91-98.

[69] Storch EA, Murphy TK, Geffken GR, Bagner DM, Soto O, Sajid M, Allen P, et al. Factor analystic study of the children's yale-brown obsessive-compulsive scale. *J Am Acad Child Adolesc Psychiatry*. 2005; 34(2):312-319.

[70] van Vlimmeren LA, Takken T, van Adrichem LN, van der Graaf Y, Helders PJ, Engelbert RH. Plagiocephalometry: a non-invasive method to quantify asymmetry of the skull; a reliability study. *Eur J Pediatr*. 2006;165(3):149-157.

[71] Lobaugh NJ, Karaskov V, Rombough V, Rovet J, Bryson S, Greenbaum R et al. Piracetam Therapy Does not Enhance Cognitive Functioning in Children with Down Syndrome. *Arch Pediatr. Adolesc. Med.* 2001;155:442-8.

[72] Luria JW, Smith GA, Chapman JI. An Evaluation of a Safety Education Program for Kindergarten and Elementary School Children. *Arch Pediatr Adolesc. Med.* 2000;154:227-31.

[73] Eklund RC, Whitehead JR, Welk GJ. Validity of the children and youth physical self-perception profile: a confirmatory factor analysis. *Research Quaterly for Exercise and Sport*. 1987; 68(3):249-256.

[74] Hill J, Powlitch S, Furniss F. Convergent validity of the aberrant behaviour checklist and behaviour problems inventory with people with complex needs. *Research in Developmental Disabilities*. 2008; 29:45-60.

[75] Rojahn J. The aberrant behaviour checklist with children and adolescents with dual diagnosis. *Journal of Autism and Developmental Disorders*. 1991; 21(1):17-28.

Chapter 5: Conclusion

Reporting and validation of primary outcomes and their measures: a dual issue

Evidence-based decision making relies on a knowledge synthesis chain. The first link of this chain consists of results generated in randomized controlled trials (RCTs). RCTs sit at the top of the research hierarchy and are internationally accepted as the gold standard for evidence about treatment effectiveness. Trial results are preferentially used by health care providers, researchers, policy-makers, and various other decision-makers. The importance placed upon RCTs and their subsequent widespread use necessitates that their reporting be of the best possible standard. Although many trials continue to be published annually in high impact journals, we and others have identified concerns regarding the reporting and validity of the trial's primary outcome and primary outcome measures[1-4].

An RCT's primary outcome is the pre-specified variable that is measured in the trial and is considered to be of greatest importance and relevance[5]. The sample size calculation for a trial is often based upon the primary outcome. The primary outcome measure is the tool used to measure the primary outcome and therefore generates trial results. These results are thus dependent on the validity of the measures. The careful selection of valid and reliable measures is an important step in the conduct of any RCT.

Thesis Objectives and Results

We conducted a series of studies assessing the reporting of primary outcomes (objective 1) and the subsequent reporting of measurement properties of the primary outcome measures. Specifically, we assessed the accuracy of reporting of measurement properties (objective 2) and whether the paucity of reporting was justified due to a lack of validation of the outcome measures (objective 3). In order to achieve the outlined objectives, we conducted a systematic review and used a variety of objective tools that were developed within Delphi consensus panels. These tools included a methodological PubMed search filter designed to retrieve studies on measurement properties[7], the COSMIN checklist with its 4-point rating scale[8] to assess the methodological quality of studies on

measurement properties, and a modified version of Terwee quality criteria[9] to assess the quality of measurement properties.

In our systematic review assessing primary outcomes reporting in pediatric RCTs published in high impact journals between 2000 and 2010, we discovered that only half of included RCTs reported a single primary outcome (48.5%), and more than a quarter of trials did not identify any primary outcome (27.2%). The remaining trials (24.3%) identified more than one primary outcome with the majority reporting two primary outcomes but this number increased to an alarming 20 primary outcomes in one trial. Inadequate reporting of primary outcomes was identified in these pediatric trials published in top medical journals. This is cause for great concern as the credibility of trials relies on their outcomes [6]. It is crucial that their outcomes are valid and reliable, and clearly reported as such.

We further divided the trials reporting single primary outcomes identified in our systematic review into those that used a primary outcome measure and reported its measurement properties (n=7) and those that did not report measurement properties (n=12). We assessed the accuracy of reporting of the seven trials that provided information on measurement properties. We found that the authors accurately reported relevant studies on measurement properties for seven of the eight primary outcome measures identified but that the quality of the measurement properties and the studies in which they were evaluated highlighted a need for further validation. We then assessed the 12 trials that did not report any measurement properties. We found that authors failed to report measurement properties for nine of 12 measures when they had in fact been evaluated. Although measurement properties had been evaluated, the quality of these properties along with the quality of the studies assessing the properties was insufficient and the need for further validation was once again highlighted.

Limitations

Limitations of the studies conducted include the use of a search strategy that only retrieves studies mentioning the outcome measure of interest and terms for measurement

properties in their titles and abstracts. The sensitive search filter is also limited to validation or comparative published studies and therefore did not retrieve any outcome measure manuals. To compensate for this, we searched reference lists of retrieved studies and compared retrieved studies to citations provided in the pediatric RCTs, to retrieve any additional potentially relevant studies that were not identified by the filter.

Another limitation lies in the absence of independent duplication of the quality assessments. However, due to the lack of multiple aspects required for a positive assessment and the use of objective tools that facilitate reproducibility of the results, minor discrepancies or errors would not alter the overall conclusion of inadequate reporting and insufficient validation.

Lastly, we restricted our studies by the English language and to the original versions of the outcome measures. This is however a reasonable restriction as the validity and reliability of the original version of the measure should be established prior to any cross-cultural adaptations of the measure.

Implications for Practice

Based on the results of the studies conducted, the issue amongst pediatric RCTs published in high impact journals is twofold. There is inadequate reporting of primary outcomes and the measurement properties of their primary outcome measures but there is also insufficient validation such that authors do not have high quality evidence of measurement properties that they can confidently report.

The implications of these results include collaboration with stakeholders already vested in improving the reporting and validation of outcomes and their measures to increase awareness of this dual issue. These knowledge users represent an ideal forum for knowledge translation. The findings of these studies can be disseminated through their vast networks that include researchers, clinicians, trialists, funding agencies, journalists, and educators. The results generated provide empiric data for these knowledge users to revise their current policies and guidelines such that the reporting and validation of

outcomes and their measures becomes a standardized step in their evidence-based decision processes. Established reporting guidelines like the Consolidated Standards of Reporting Trials (CONSORT) checklist, can be revised to include items that require adequate reporting of primary outcomes and their measures. Journals that endorse CONSORT, including all the high impact journals assessed in our systematic review, would thus require these elements for publication and editors and peer reviewers would therefore seek to identify the reporting of these aspects in any submitted work.

In order to facilitate the reporting of outcomes and the informed selection of outcome measures, knowledge users involved in the pre-conception phases of RCTs can incorporate these findings into their review criteria. Funding agencies for example, can require clear reporting of a pre-defined primary outcome along with explanations supporting the selection of a particular outcome measure when researchers seek funding from their agency. It can be argued that conducting an RCT without defining a primary outcome or using a measure that is valid and reliable is not in keeping with ethical practice. Ethics review boards could therefore also include the need to clearly define a primary outcome and the appropriateness of an outcome measure in the ethics approval of an RCT.

Furthermore, interactive outcomes-focused workshops can be organized with knowledge users so members of their groups will be better equipped to make informed selections of outcome measures and will also be able to critically evaluate and interpret results of studies that use outcome measures.

Collaboration, dissemination and subsequent training of knowledge users will result in informed selection of outcome measures and evidence-based decision making. These results have the potential to enhance the rigour and subsequent relevance of primary research.

Implications for Research

The implications of these results for future research include further validation of outcome measures in methodologically rigorous studies such that trialists, researchers and authors can make informed selections of measures. The criteria used in our studies to evaluate the quality of properties and their studies can serve as guides for trialists who are designing validation studies to ensure they are adhering to design, methodology, statistical and reporting standards. Other evidence-based decision makers can also use these criteria to critically evaluate and interpret results of studies they use for their evidence-based practice. Confidence in the measure used in a trial will translate into confident interpretation of the trial results by all knowledge users.

Our systematic review on primary outcomes reporting found widespread variety in the terminology used to identify a primary outcome including, primary endpoint, primary efficacy variable, primary objective, and primary dependent variable(s), to name a few. Consensus on the appropriate terminology for these outcomes is necessary as this will facilitate their improved reporting and identification. This consensus can be reached using rigorous study designs such as a Delphi study to involve international expertise on this issue.

Finally, with further evaluation of measurement properties of outcome measures, a highly accessible repository of validated outcome measures should be developed in collaboration with knowledge users who seek to promote the quality of research. This repository would serve as a tool researchers can consult when planning their trials again lending to the informed selection of measures. Knowledge users could also refer to this database to determine whether the measurement properties of the outcome measure used in a study they are evaluating have been assessed.

Conclusion

RCTs requires substantial resources, it has been estimated that the mean cost for a single RCT is $12 million US[10]. When we think about the number of pediatric RCTs conducted within the 10 year period examined, the number of children enrolled in these trials, and

the considerable expense, effort and resources dedicated to their conduct, determining that the trials are not valid highlights a suboptimal use and interpretation of RCTs. Furthermore, policy makers, peer reviewers, authors, journal editors and publishers among others will have not promoted the highest quality evidence expected and achievable of a gold standard. Great strides have been made in child health because of research, despite its imperfections. Improved primary outcome measurement and reporting is vital to pediatric clinical trials, and the decision-makers who rely on them: clinicians, policy-makers, and ultimately, patients themselves.

References

[1] Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement Issues in Trials of Pediatric Acute Diarrheal Diseases: A Systematic Review. *Pediatrics* 2010; 126;e222-e231.

[2] 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(4): e96.

[3] Reid GT, Walter FM, Brisbane JM, Emery JD. Family History Questionnaires Designed for Clinical Use: A Systematic Review. *Public Health Genomics* 2009; 12:73-83.

[4] Zhang B, Schmidt B. Do we measure the right end points? A systematic review of primary outcomes in recent neonatal randomized clinical trials. *J Pediatr*. 2001;138(1): 76-80

[5] Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Ann Int Med 2010:152. Epub 24 March.

[6] Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: Introduction. *J Rheum* 1993; 528–530.

[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;18:1115-1123.

[8] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2011; 21:651-657.

[9]Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, and de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34-42.

[10] Johnston SC, Rootenberg JD, Katrak S, Smith WS, Elkins JS. Effect of a US National Institutes of Health programmed of clinical trials on public health and costs. *Lancet* 2006; 367:1319-1327.