2-12-2021

Strengths and challenges of the COSMIN tools in the appraisal of outcome measures: A case example for speech-language therapy

Elaine Kwok
McMaster University, kwoke3@mcmaster.ca

Peter Rosenbaum
McMaster University

Nancy Thomas-Stonell
McMaster University

Barbara Jane Cunningham
Western University, bjcunningham@uwo.ca

Citation of this paper:
Kwok, Elaine; Rosenbaum, Peter; Thomas-Stonell, Nancy; and Cunningham, Barbara Jane, "Strengths and challenges of the COSMIN tools in the appraisal of outcome measures: A case example for speech-language therapy" (2021). *International Journal of Language & Communication Disorders* 8.
https://ir.lib.uwo.ca/precisepreschoolpubs/8

Title: Strengths and challenges of the COSMIN tools in the appraisal of outcome measures: A case example for speech-language therapy

Abstract:

Background: The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) is an international initiative that offers standardized and validated tools to guide the appraisal of outcome measures in healthcare.

Aims: This study aimed to explore the use of a new set of tools from COSMIN to appraise studies on outcome measures available to speech and language therapists (SLTs).

Methods: We used the COSMIN tools to appraise seven studies and a user manual that reported the measurement properties of the Focus on the Outcomes of Children Under Six (FOCUS), a validated measure of preschoolers' communicative participation that is used in various contexts around the world.

Results: Using COSMIN guidelines, the FOCUS was categorized as a "Category A" tool because there was a sufficient level of evidence to support its content validity and internal consistency. According to the COSMIN guidelines, this means that the FOCUS can be recommended for clinical use. The quality of evidence supporting measurement properties of the FOCUS received a rating of 'moderate', meaning users can have moderate confidence in its measurement properties. Since these ratings from the COSMIN tools may be unclear to users of the FOCUS, we have provided more specific recommendations.

Conclusions & Implications: The COSMIN tools offer detailed standards to support the appraisal of outcome measures available to SLTs. However, several limitations were observed, and recommendations to support the application of the COSMIN tools are provided.

Declaration of Interest: The authors declare no conflicts of interest.

What is already known on this subject?

Collecting outcome data is essential to ensuring speech and language therapy is effective. Until the development of Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) there was a lack of standards in the way the measurement properties of outcome measures were appraised.

What this study adds?

This paper used the Focus on the Outcomes of Children Under Six (FOCUS), a measure of preschoolers' communicative participation outcomes in speech and language therapy, as a case example to illustrate the applications of the COSMIN tools. In doing so the strengths and limitations of the current COSMIN tools in appraising the quality of outcome measure instruments are emphasized.

Clinical implications of this study?

The COSMIN tools offer a step-by-step, standardized approach to appraising various measurement properties in outcome instruments. Due to existing limitations of the COSMIN tools, appraisal should provide clear and specific recommendations so users of outcome measures (e.g., SLTs, researchers) can identify the appropriate uses of each instrument.

Introduction

Outcome measures are important tools for assessing the impact of a healthcare system (Agency for Health Research and Quality 2011, Donabedian 1988). Across the globe, speech language therapists (SLTs) are encouraged by their professional organizations to use outcome measures (Mullen and Schooling 2010, Royal College of Speech & Language Therapists 2020, Speech-Language & Audiology Canada 2010). Amongst many benefits, data collected using outcome measures allow for the evaluation of clinical effectiveness, inform quality improvement efforts, and support best practices (Royal College of Speech & Language Therapists 2020). Moreover, when used in large health systems, data collected using valid and reliable outcome measures can generate evidence to inform decisions about services (e.g., service type, length, and intensity). For SLTs, outcome measures can be used to gather objective data on clients' skills and progress – which can be used to guide clinical decisions (Garland *et al.* 2003). Patient-reported outcome measures provide clients and families with opportunities to express their perspectives, values, and preferences about their own care, improving SLTs' accountability to their clients (Ronen *et al.* 2000).

To realize the many benefits associated with outcome measures, it is imperative to select tools that have appropriate measurement properties (Enderby and John 2015, 2020, Speech-Language & Audiology Canada 2012, Threats 2013, World Health Organization 2001). Despite some graduate training to support understanding of psychometrics, in practice SLTs report a lack of confidence (Kerr *et al.* 2003), time (Kerr *et al.* 2003), and resources (e.g., access to literature Vallino-Napoli and Reilly, 2004) to evaluate the properties of the outcome measures they use. These barriers may explain why measurement properties of instruments were not a major factor influencing SLTs' choice of instrument (Betz *et al.* 2013) and that “misuses” of measurement instruments were frequently reported (Kerr *et al.* 2003). One way to support SLTs in choosing appropriate instruments is to appraise existing outcome measures systematically and critically.

The Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) offers an internationally agreed-upon taxonomy for evaluating the psychometric properties of health-related patient reported outcome measures (Mokkink, Terwee, Patrick, *et al.* 2010). According to COSMIN, an outcome measure can be evaluated based on (1) the methods used in *tool development* and (2) psychometric properties (*validity*, *reliability*, and *responsiveness*) (Barten *et al.* 2012, Lambert and Hawkins 2004, Mokkink, Terwee, Patrick, *et al.* 2010). Additionally, COSMIN acknowledges the importance of two additional properties: *interpretability* and *feasibility*. Interpretability refers to the ease of deriving meaning from an instrument's scores, and feasibility refers to the ease with which an instrument is adopted into its intended context (Mokkink *et al.* 2018, Mokkink, Terwee, Patrick, *et al.* 2010). At the moment, *interpretability* and *feasibility* are not formally assessed with the COSMIN tools, but rather are viewed as important considerations for the practical use of an outcome measure (Mokkink, Terwee, Patrick, *et al.* 2010).

To support the evaluation of patient-reported outcome measures, COSMIN offers a validated Risk of Bias checklist (Mokkink *et al.* 2018) and a user manual with step-by-step guide to support instrument appraisals (COnsensus-based Standards for the selection of health Measurement INstruments 2020, Prinsen *et al.* 2018, Terwee *et al.* 2018). The Risk of Bias checklist was developed based on a literature review of the measurement properties of health-related measurement instruments and the consensus of a panel of 57 experts involved in a Delphi study (Mokkink *et al.* 2018, Mokkink, Terwee, Knol, *et al.* 2010). Although the COSMIN tools were originally developed and validated to appraise patient-reported outcome measures, it has been argued that their criteria are also applicable to the evaluation of non-patient reported outcome measures (Tate 2019). Since becoming available, these COSMIN tools have been used to evaluate a range of patient/parent/clinician-reported outcome measures within healthcare (e.g., Bull et al., 2019; Howell et al., 2020; Williams and Beovich, 2020), and were found to be some of the more carefully developed and comprehensive tools to appraise outcome measures (Tate 2019). Additionally, some professional organizations have begun to use COSMIN tools to recommend instruments that met standards for clinical use (England *et al.* 2019, Pick *et al.* 2020).

The purpose of this paper is to illustrate use of the COSMIN tools (i.e., the Risk of Bias checklist) (Mokkink *et al.* 2018) and the criteria for good measurement properties (Prinsen *et al.* 2018, Terwee *et al.* 2018) to appraise an outcome measure in speech-language therapy. Importantly, this paper illustrates *how* appraisal results can be considered to draw clinically meaningful recommendations regarding the use of existing outcome measures. We argue that existing outcome measures should be considered on an instrument-by-instrument basis for three reasons. First, most existing outcome measures are found not to meet all standards of the COSMIN tools, which is not surprising given that most tools were developed prior to the COSMIN standards (Bull *et al.* 2019, Howell *et al.* 2020, Williams and Beovich 2020). Second, not all measurement properties are equally important for all clinical decisions, so the properties important to individual clinical decisions or purposes should be considered in context, on a measure-by-measure basis rather than categorizing tools as ‘good’ versus ‘bad’ (Bull *et al.* 2019, Messick 1993, 1995). Third, in practice, clinicians are limited by the resources available to them (i.e., instruments available in their clinic/district). Therefore, considering what each outcome measure can and cannot do is perhaps more practical and appropriate than identifying one “best” tool.

To contextualize the considerations when applying the COSMIN tools, we have included an evaluation of an outcome measure that is currently implemented in at least one large clinical health system. The Focus on the Outcomes of Children Under Six (FOCUS) (Thomas-Stonell *et al.* 2015) is one of a handful of validated tools explicitly designed to measure *outcomes* for preschoolers receiving SLT interventions (Thomas-Stonell *et al.* 2010). Furthermore, the FOCUS is the only validated tool available to assess how preschoolers use their communication to participate in real-world situations (Cunningham *et al.* 2017) – an outcome that has been identified as important and meaningful to families (Lindsay and Dockrell 2004, Roulstone *et al.* 2013).

With the FOCUS as a case example, the goal of this paper is to illustrate how the COSMIN tools can be used as a guide to identify the strengths and limitations that are associated with any outcome measure. We intend for this paper to serve as a support for researchers and SLTs in selecting tools that are both psychometrically strong and meaningful for practice. The paper will also have implications for test developers who will want to understand a new standard for evaluating outcome measures. This work involves the secondary analysis of data collected during a recent scoping review of the literature related to the FOCUS (Cunningham *et al.* 2020). More specifically, we used findings from this review to identify studies that reported psychometric properties of the FOCUS, which we appraised using the COSMIN tools.

Methods

Search strategy and inclusion criteria

Cunningham et al. identified 25 publications that reported on either the development or application of the FOCUS (Cunningham *et al.* 2020). In the current study, we reviewed these 25 publications to identify those that reported psychometric properties of the FOCUS using the following inclusion criteria: (i) the article was peer-reviewed and about the English-language version of the FOCUS or its derivative (i.e., the FOCUS-34); (ii) the study evaluated the psychometric properties of the FOCUS; and (iii) the article was published in English. Although not peer reviewed, the published FOCUS user manual was also included in order to complete a comprehensive appraisal (Thomas-Stonell *et al.* 2020). This user manual presented information for both the original and shortened FOCUS tools (i.e., FOCUS-34), and was created by drawing upon peer-reviewed research.

Appraisal of psychometric properties.

An extraction spreadsheet (see Appendix 1) was developed to record the following data from each identified publication: (1) FOCUS version, (2) measurement properties investigated, and (3) study methodology and results. Based on the psychometric properties investigated in each publication, the relevant portions of the COSMIN Risk of Bias checklist (Mokkink *et al.* 2018) and the criteria for good measurement properties were completed (i.e., the reliability portion for studies that investigated reliability) (Prinsen *et al.* 2018, Terwee *et al.* 2018). Descriptions of each measurement property as well as explanations of how each applies to the FOCUS are presented in table 1.

< Table 1 Here >

Appraisal of a validation study's methodology. The COSMIN Risk of Bias checklist has specific sections dedicated to evaluating the *methodology* of studies conducted to demonstrate measurement properties of an outcome measure. Under each section (i.e., for each specific measurement property), the checklist provides a list of items to evaluate a study's quality. For example, the section on content validity contains 31 items concerning the appropriateness of data collection methods, participant sample, sample size, and data analysis approach. For each item, a study's methodology is rated as *very good*, *adequate*, *doubtful* or *inadequate* based on well-defined standards (these ratings corresponding to *excellent*, *good*, *fair*, *poor* on the original rating scale (Mokkink *et al.* 2018)). For example, the standards for sample sizes are: ≥50 (for *very good* rating), ≥30 (*adequate*), <30 (*doubtful*), unclear sample size (*inadequate*). Across all items, the lowest rating is selected as the ratings for a study's methodologies (i.e., the worst score counts) (Mokkink *et al.* 2018). In cases where multiple studies evaluated the same measurement property and received different quality ratings, the higher quality rating was taken as the overall rating. We reasoned that measurement properties are subject to continuous evaluation; thus, if multiple studies were conducted over several years, the study with the most rigorous design should be used. Data extraction and quality rating were first completed by E.K. and reviewed by NT-S with 93% agreement. Due to the "worst score counts" rule, no disagreements resulted in changes in the rating of any psychometric property. All disagreements were resolved through discussion with BJC and PR.

Appraisal for validation study's results. Apart from evaluating the methodology of validation studies, the results of each primary study were considered using the COSMIN criteria for good measurement properties (Prinsen *et al.* 2018, Terwee *et al.* 2018). This provides a quality rating based on the results reported in validation studies regarding a psychometric property of an outcome measure, and can range from "+" sufficient; "-" insufficient; "?" indeterminate; "±" inconsistent. For example, for reliability testing, a study would receive a "+" if it found an ICC or weighted Kappa value ≥ 0.70. In contrast, a "–" rating would be assigned if the reported ICC or weighted Kappa value was <0.70. A "?" would be assigned if no ICC or weighted Kappa was reported. If more than one study was conducted for reliability but had mixed findings, an "±" inconsistent rating was assigned. For the evaluation of a study's results, E.K. and NT-S had 100% agreement in their ratings.

Together, these COSMIN tools provide standards to appraise study methodologies and reported results (Mokkink *et al.* 2018). There are, however, some limitations to the COSMIN rating scales. When validating the COSMIN Risk of Bias rating, the COSMIN developers noted that "*a study often received a 'fair' quality rating (i.e., doubtful on the new rating scale) only* *because it was not reported how missing items were handled. It was argued that this would not necessarily lead to biased results of the study*" (Mokkink et al., 2018, p.2). This is why, when a study receives *doubtful*/*inadequate* rating on Risk of Bias or an *indeterminate* rating on the criteria for good measurement properties (Prinsen *et al.* 2018, Terwee *et al.* 2018), the reasons behind the rating should be scrutinized before assigning an overall quality rating for the measurement tool. To this end, COSMIN also provides a rating scale (*High*, *Moderate*, *Low*, *Very Low*) and key factors to consider when indicating the overall quality of a measurement tool. The four key factors to consider include: high risk of bias in study methodology and reporting, inconsistent findings across studies, imprecision (referring to a small sample size) and indirectness (referring to validation studies completed in a population dissimilar from the intended users of the instrument).

Lastly, to improve transparency, COSMIN recommends categorizing outcome measures into the following categories: (A) instrument with evidence for sufficient content validity AND at least low-quality evidence for sufficient internal consistency; (B) instrument categorized not in A or C; and (C) instrument with high quality evidence for an insufficient measurement property. According to COSMIN, outcome measures in Category A can be recommended for use and their results can be trusted; those in category B can be recommended provisionally, subject to further evidence being provided; and category C tools should not be recommended.

Results

Full-text screening identified 7 articles that met the inclusion criteria. These articles and the FOCUS user manual were included in the appraisal (reasons for studies being excluded are shown in figure 1). The studies included described the methodology used to develop FOCUS items (Thomas-Stonell *et al.* 2009), content validity (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2010), construct validity (Thomas-Stonell *et al.* 2010, Washington, Thomas-Stonell, *et al.* 2013), internal-consistency reliability (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2010), inter-rater reliability (Oddson *et al.* 2013, Thomas-Stonell *et al.* 2010, 2013, Washington, Oddson, *et al.* 2013), test-retest reliability (Thomas-Stonell *et al.* 2010, Washington, Oddson, *et al.* 2013), responsiveness (Thomas-Stonell *et al.* 2013), and interpretability (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2013) of the FOCUS. It should be noted that the FOCUS-34 is a streamlined version of the original 50-item FOCUS, as 16 items were removed based on empirical findings from item response analysis (Oddson *et al.* 2019). For readability, measurement properties that apply to both the original FOCUS and the FOCUS-34 are described as *FOCUS tools*.

< Insert Figure 1 Here >

The COSMIN quality ratings for the FOCUS validation studies are presented in table 2. Considering all available evidence related to the measurement properties of the FOCUS, we rated the quality of evidence as *Moderate* and categorized the FOCUS tools as Category A due to sufficient content validity and internal consistency. According to COSMIN, this means that the FOCUS tools can be recommended for use, and that we are *moderately confident* that the FOCUS provides an estimate close to what has been stated in the literature (i.e., the reported measurement properties). The major considerations that led to this overall rating are described in the sections that follow. A detailed rationale behind each quality rating presented in table 2 can be found in appendix 1. Item-by-item scoring of the COSMIN tools is available from the authors upon request. We acknowledge that the overall *Moderate* rating and Category A nomenclature are not very informative, therefore, we provided the following sections to describe the clinical implications from the appraisal findings.

Tool Development and Content Validity (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2009, 2010): Current findings suggest that the FOCUS measures communicative participation outcomes that are meaningful and important to both parents and SLTs. Each of the FOCUS items was found to be clear and relevant to users (the development and testing of the FOCUS items involved 349 parents and SLTs). Through a combination of quantitative and qualitative analysis, the items on the FOCUS-34 were also found to provide a comprehensive measure of communicative participation outcomes. These studies received positive ratings because a clear description was provided of the aim of the FOCUS, the high ecological validity in the item generation, selection and reduction process that involved parents, SLTs and statisticians. For SLTs, since the FOCUS was validated in various real-world clinical settings that serve a range of clinical populations, it provides a consistent tool to measure gains in preschoolers' communicative participation skills during speech-language interventions.

Validity (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2010, 2013, Washington, Thomas-Stonell, *et al.* 2013): FOCUS scores were shown to correlate moderately with existing instruments that measure related, but dissimilar constructs (e.g., the Pediatric Quality of Life Inventory (Varni 1998), the communication domains on the Ages and Stages Questionnaire – Social/ Emotional (ASQ-SE) (Squires *et al.* 2003), Communication and Socialization domains of the Vineland Adaptive Behavior Scales (VABS) (Sparrow *et al.* 2005)). Meanwhile, FOCUS scores did not correlate with domains that are not related to communication. These studies on construct validity received an overall *very good* rating on methodology and *sufficient* rating on results. For SLTs, these findings clarify what is being measured by the FOCUS, namely an aspect of communication that relates to how children use communication skills to participate in everyday situations. Additionally, there is a very high correlation between scores on the FOCUS-34 and the original FOCUS in the criterion validity study (Oddson *et al.* 2019), which suggests that the FOCUS-34 sufficiently reflects the original tool and can provide a more efficient option for data collection for those who need or want it.

Reliability (Oddson *et al.* 2013, 2019, Thomas-Stonell *et al.* 2010, Washington, Oddson, *et al.* 2013): Two studies reported on internal consistency (Oddson *et al.* 2019, Thomas-Stonell *et al.* 2010) and received an overall rating of *very good* for methodology and *sufficient* for results. Studies that explored inter-rater reliability received an overall *adequate* rating for methodology and *sufficient* rating for results. There was a moderate level of inter-rater reliability between SLTs and parents or amongst SLTs, which suggests that even when completed by different individuals, scores on the FOCUS reliably reflect preschoolers' communicative participation skills. Thus, it is not necessary for both parents and SLTs to complete the FOCUS in order to capture change. With regards to test-retest reliability, the current *doubtful* and *indeterminate* ratings were due to the fact that Pearson's correlation (*r* =0.96) (Washington, Oddson, *et al.* 2013) were reported instead of ICC values in the FOCUS validation studies. While the quality ratings were limited by the reported statistics, we do not believe it t should not limit use of the FOCUS in clinical practice.

Responsiveness (Thomas-Stonell *et al.* 2013): There is no single agreed-upon approach for measuring responsiveness (i.e., an outcome measure's ability to detect change) (Thomas-Stonell *et al.* 2007) and COSMIN offers a range of checklists to assess responsiveness of an outcome measure. The responsiveness of the FOCUS was evaluated in two ways. First, change scores on the FOCUS were compared to the change scores measured by three established measures of speech, intelligibility and language (i.e., Children's Speech Intelligibility Measure (Wilcox and Morris 1999), Percent Consonant Correct-Revised (Schriberg *et al.* 1997) and Developmental Sentence Scoring (Lee and Canter 1971)) (Thomas-Stonell *et al.* 2013). There was a fair level of agreement between measures when a minimally clinically important difference was observed. A fair but not excellent level of agreement is to be expected, since the FOCUS and these comparator measures do not measure the same construct. This study received a *very good* rating for methodology and *sufficient* rating for results. Second, responsiveness of the FOCUS was demonstrated when preschoolers receiving interventions showed more change than a group of children on a waitlist (*M* =18.2 and *M* = 5.87 points respectively, and that the average change scores in the waitlist group was lower than the 16-point cut-off scores to be considered minimally clinically significant change) (Thomas-Stonell *et al.* 2013). However, because the study was observational rather than experimental, and was conducted within a practice context, this finding was limited by the unequal intervention (90 days) and waitlist (60 days) intervals. This resulted in a *doubtful* rating for the study's methodology. For the FOCUS-34 tool, the change scores (i.e., pre-to-post intervention) highly correlate with those from the original FOCUS tool.

For SLTs, these findings suggest that the FOCUS demonstrates comparable responsiveness with speech, intelligibility and language outcome measures that are commonly used in practice. One area of constraint related to responsiveness is that with the published evidence, the possibility that changes observed on the FOCUS were due to natural development or some other factors cannot be fully ruled out. However, SLTs using the recommended criterion to interpret when a minimally clinically significant change has occurred will minimize contributions from natural development and random error. In the absence of a control group, we caution SLTs against attributing change on the FOCUS solely to specific treatment effects, as this is difficult to determine given the many factors that can affect children's development at any given time (e.g., growth spurt, change in language learning environment).

Feasibility & Interpretability (Thomas-Stonell *et al.* 2020): These two properties are not formally evaluated by the COSMIN tools, so here we summarize the major findings related to the FOCUS for these two properties. The streamlined FOCUS-34 provides a reliable and efficient option for data collection, which can be completed by parents or SLTs within 10 minutes. The FOCUS tools offer criterion scores to support SLTs in interpreting change in children’s communicative participation during intervention. The criterion score had 95% agreement between parents’ and SLTs’ judgements of whether a clinically important change had occurred. This criterion score allows SLTs and researchers to determine whether meaningful change occurred during an intervention period, and theoretically this can be done without a control group, making it particularly useful for both research and practice.

<Insert Table 2 Here>

Discussion:

The purpose of this paper was to illustrate the use of relatively new tools from the COSMIN (Mokkink *et al.* 2018) to guide the appraisal of outcome measurement tools. As such, the discussion is focused on the benefits and limitations of the COSMIN.

The COSMIN tools (2018) are comprehensive and offer standards to appraise patient-reported outcome measures s. For trainees and SLTs, the standards provide an objective way of appraising measurement properties of outcome measures. We reiterate that the appraisal should be conducted on an instrument-by-instrument basis and guided by a clearly articulated clinical or research question(s). For tool developers, COSMIN provides a standard to improve the quality of reporting for the development and validation of outcome measurement tools. In fact, two authors of this paper (NS-T and PR) were involved in the development and validation of the FOCUS tools, and this checklist has helped identify additional details that could be included in future editions of the FOCUS user’s manual to continue to support clinical practice and research.

We also observed important limitations in our efforts to apply the COSMIN tools. The first is the time needed to complete the appraisal. It took over 25 hours for our team of four authors with graduate-level training in tool development to complete the evaluation of the FOCUS tools (not including the time to become acquainted with the COSMIN tools). Clinicians focused on providing quality care to clients are unlikely to have the time or academic background to complete this type of detailed evaluation (e.g., the knowledge to evaluate statistical analysis and results such as item-response analysis). Thus, the onus of evaluating measurement properties may necessarily fall to interested researchers, professional colleges and tool developers.

Second, the COSMIN tool has not yet developed a rating scale to evaluate *interpretability* or *feasibility*, but we believe these are among the most important clinical considerations for SLTs; they are interested in understanding whether observed changes are clinically meaningful, and whether a tool can be easily adopted into practice. When using COSMIN to appraise outcome measurement tools, these two properties should not be overlooked simply due to a lack of clear appraisal standards. Until a rating scale is available on COSMIN, we recommend referring to the detailed data extraction matrix that is available in the COSMIN user manual (Prinsen *et al.* 2018, Terwee *et al.* 2018) to identify information related to interpretability and feasibility. We also recommend using other tools to supplement appraisals in these areas, for example, the criteria from the *Acceptability and Utility* checklist from the Allied Health Professions (AHP) Outcome Measures UK Working Group (Allied Health Professions (AHP) Outcome Measures UK Working Group, 2019) and the *Interpretability* and *Burden* tool from the Scientific Advisory Committee of the Medical Outcomes Trust (Lohr 2002).

Thirdly, we emphasize the need for any appraisal completed using the COSMIN tools to consider the practical implications of appraisal findings (i.e., making clear recommendations regarding tool use). One important reason for making a clear recommendation statement is concern about the categories on the COSMIN risk of bias checklist (i.e., the *very good*, *adequate*, *doubtful* and *inadequate* scale) and the criteria for determining good measurement properties (i.e., the + sufficient, ? indeterminate , - insufficient). Using COSMIN, each measurement property receives ratings that reflect the design and reporting of validation studies, and not the measurement property of the tool itself. Clinicians, policymakers and other stakeholder groups who are unfamiliar with the COSMIN tools may take these ratings to mean that a tool is "very good", "adequate", "doubtful" or "inadequate". Current knowledge on measurement properties suggests that users consider the *purpose for measurement* (i.e., What is the clinical/research question?*)*. A measurement tool is 'good' or 'bad' for specific usesl, but tool should not be viewed as categorically good versus bad (Bull *et al.* 2019, Messick 1993, 1995). Another reason for making clear recommendations is the fact that systematic reviews have reported many existing outcome measurement tools would not be considered to be of adequate/sufficient quality based on appraisals done using the COSMIN tools (Bull *et al.* 2019, Howell *et al.* 2020, Williams and Beovich 2020). This is a commonly reported limitation in studies that retrospectively applied the COSMIN tools to evaluate a measurement tool that was developed prior to COSMIN being published; it reflects a lack of standards in the reporting of measurement properties as well as an evolving understanding of best-practice in tool development and validation (Bull *et al.* 2019, Van Tiggelen *et al.* 2020, Williams and Beovich 2020). Providing clear recommendations will help users interpret appraisal findings accurately, and understand the appropriate use of existing outcome measurement tools. The results reported in this paper serve as a case example for how clear recommendations can (and should) be made on an instrument-by-instrument basis, depending on the purpose of the measure and the question(s) to be answered.

Another limitation of COSMIN relates to the scope of application. The COSMIN tools were originally developed and validated to appraise *patient-reported* outcome measures. While it has been argued that the criteria in the COSMIN tools are applicable to evaluate non-patient reported outcome measures (Tate 2019), it is possible that more criteria should be considered when appraising non-patient reported outcome measures. Recent work is expanding the scope of the COSMIN tools for the appraisal of clinician-reported, performance-based and laboratory-based outcome measure instruments (Mokkink *et al.* 2020). As such, when more comprehensive, validated appraisal tools become available, the work described here will be updated.

A future direction of our work is to appraise multiple functional outcome measures used by SLTs, particularly those used with young children. These appraisals will allow us to identify the strengths and limitations, and the specific purposes, of existing outcome measures, and the appropriate uses of each of the available measures for SLTs. Recommendations will be developed based on these appraisals, and will be available as an online resource for SLTs.

Conclusion:

The study illustrates the use, and limitations, of the COSMIN tools (Mokkink *et al.* 2018, Prinsen *et al.* 2018, Terwee *et al.* 2018), which were designed to appraise outcome measures systematically. The COSMIN tools provide an up-to-date, comprehensive list of factors to consider in psychometric appraisals, but due to an evolving understanding of psychometric properties and reporting standards, many existing clinical tools (i.e., those developed prior to the COSMIN tools) may receive doubtful/indeterminate ratings on COSMIN. Appraisal of all existing outcome measurement tools should consider carefully the reasons behind quality ratings and how these may impact clinical decisions. This paper demonstrates how measurement properties should be considered in conjunction with clinical decisions to be made based on using the outcome measurement instrument(s). Lastly, for researchers and tool developers, this paper introduces a newly available tool that can be used to guide the development and reporting of outcome measurement instrument(s). We believe this study will be a useful reference for SLTs, researchers, and developers in appraising, choosing and creating appropriate outcome measurement tools.

References

Agency for Health Research and Quality, 2011, Types of health care quality measures [online]. Available: https://www.ahrq.gov/talkingquality/measures/types.html [Accessed 8 January 2020].

Allied Health Professions (AHP) Outcome Measures UK Working Group, 2019, Key questions to ask when selecting outcome measures: a checklist for allied health professionals.

Barten, J.A., Pisters, M.F., Huisman, P.A., Takken, T., and Veenhof, C., 2012, Measurement properties of patient-specific instruments measuring physical function. *Journal of Clinical Epidemiology*, **65**, 590–601. https://doi.org/10.1016/j.jclinepi.2011.12.005.

Betz, S.K., Eickhoff, J.R., and Sullivan, S.F., 2013, Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, **44**, 133–146. https://doi.org/10.1044/0161-1461(2012/12-0093).

Bull, C., Byrnes, J., Hettiarachchi, R., and Downes, M., 2019, A systematic review of the validity and reliability of patient-reported experience measures. *Health Services Research*, **54**, 1023–1035. https://doi.org/10.1111/1475-6773.13187.

COnsensus-based Standards for the selection of health Measurement INstruments, 2020, COSMIN Tools [online]. Available: https://www.cosmin.nl/cosmin-tools/ [Accessed 2 November 2020].

Cunningham, B.J., Thomas-Stonell, N.L., and Rosenbaum, P., 2020, Assessing communicative participation in preschool children with the Focus on the Outcomes of Communication Under Six: a scoping review. *Developmental Medicine & Child Neurology*.

https://doi.org/10.1111/dmcn.14665.

CUNNINGHAM, B.J., WASHINGTON, K.N., BINNS, A., ROLFE, K., ROBERTSON, B., and ROSENBAUM, P., 2017, Current methods of evaluating speech-language outcomes for preschoolers with communication disorders: A scoping review using the ICF-CY. *Journal of Speech, Language, and Hearing Research,* **60**, 447–464. https://doi.org/10.1044/2016.

DONABEDIAN, A., 1988, The quality of care: How can it be assessed? *JAMA*, **260**, 1743–1748. https://doi.org/10.1001/jama.1988.03410120089033.

ENDERBY, P. and JOHN, A., 2015, *Therapy outcome measures for rehabilitation professionals*. Third Edit. (Guildford: J&R Press Ltd).

ENDERBY, P. and JOHN, A., 2020, *Therapy Outcome Measures theoretical underpinning and case studies* (Havant: J&R Press Ltd).

ENGLAND, B.R., TIONG, B.K., BERGMAN, M.J., CURTIS, J.R., KAZI, S., MIKULS, T.R., O'DELL, J.R., RANGANATH, V.K., LIMANNI, A., SUTER, L.G., and MICHAUD, K., 2019, 2019 Update of the American College of Rheumatology recommended rheumatoid arthritis disease activity measures. *Arthritis Care and Research,* **71**, 1540–1555. https://doi.org/10.1002/acr.24042.

GARLAND, A.E., KRUSE, M., and AARONS, G.A., 2003, Clinicians and outcome measurement: What's the use? *Journal of Behavioral Health Services and Research*, **30**, 393–405. https://doi.org/10.1007/BF02287427.

HOWELL, M., BRADSHAW, J., and LANGDON, P.E., 2020, A Systematic Review of Behaviour-Related Outcome Assessments for Children on the Autism Spectrum with Intellectual Disabilities in Education Settings. *Review Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s40489-020-00205-y.

KERR, M.A., GUILDFORD, S., and BIRD, E.K.., 2003, Standardized language test use: A Canadian survey utilisation. *Journal of Speech-Language Pathology and Audiology*, **27**, 10–28.

LAMBERT, M.J. and HAWKINS, E.J., 2004, Measuring outcome in professional practice: Considerations in selecting and using brief outcome instruments. *Professional Psychology: Research and Practice*, **35**, 492–499. https://doi.org/10.1037/0735-7028.35.5.492.

LEE, L. and CANTER, S., 1971, Developmental sentence scoring: a clinical procedure for estimating syntactic development in children's spontaneous speech. *J Speech Lang Hear Disord*, **36**, 315–40.

LINDSAY, G. and DOCKRELL, J.E., 2004, Whose job is it? Parents' concerns about the needs of their children with language problems. *Journal of Special Education*, **37**, 225–235. https://doi.org/10.1177/00224669040370040201.

LOHR, K.N., 2002, Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*, **11**, 193–205. https://doi.org/10.1023/A:1015291021312.

MESSICK, S., 1993, Foundations of validity: Meaning and consequences of psychological assessment. *Educational Testing Service Research Report Series*, **2**.

MESSICK, S., 1995, Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, **14**, 5–8.

MOKKINK, L., BOERS, M., VLEUTEN, C. VAN DER, BOUTER, L., ALONSO, J., PATRICK, D., VET, H. DE, and TERWEE, C., 2020, COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Medical Research Methodology*, **1**, 1–13. https://doi.org/10.21203/rs.3.rs-40864/v1.

MOKKINK, L.B., TERWEE, C.B., KNOL, D.L., STRATFORD, P.W., ALONSO, J., PATRICK, D.L., BOUTER, L.M., and DE VET, H.C.W.W., 2010, The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology*, **10**. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., and de Vet, H.C.W., 2010, The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, **63**, 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Mokkink, L.B., de Vet, H.C.W., Prinsen, C.A.C., Patrick, D.L., Alonso, J., Bouter, L.M., and Terwee, C.B., 2018, COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, **27**, 1171–1179. https://doi.org/10.1007/s11136-017-1765-4.

Mullen, R. and Schooling, T., 2010, The national outcomes measurement system for pediatric speech-language pathology. *Language Speech and Hearing Services in Schools*, **41**, 44. https://doi.org/10.1044/0161-1461(2009/08-0051).

Oddson, B., Thomas-Stonell, N.L., Robertson, B., and Rosenbaum, P., 2019, Validity of a streamlined version of the Focus on the Outcomes of Communication Under Six: Process and outcome. *Child: Care, Health and Development*, 600–605. https://doi.org/10.1111/cch.12669.

Oddson, B., Washington, K.N., Robertson, B., Thomas-Stonell, N., and Rosenbaum, P., 2013, Inter-rater reliability of clinicians' ratings of preschool children using the FOCUS©: Focus on the outcomes of communication under six. *Canadian Journal of Speech-Language Pathology and Audiology*, **37**, 170–174.

Pick, S., Anderson, D.G., Asadi-Pooya, A.A., Asadi-Pooya, A.A., Aybek, S., Baslet, G., Bloem, B.R., Nicholson, T.R., Brown, R.J., Carson, A.J., Chalder, T., Damianova, M., David, A.S., Edwards, M.J., Epstein, S.A., Espay, A.J., Garcin, B., Goldstein, L.H., Hallett, M., Jankovic, J., Joyce, E.M., Kanaan, R.A., Keynejad, R.C., Kozlowska, K., Lafaver, K., Curt Lafrance, W., Lang, A.E., Lehn, A., Lidstone, S., Maurer, C.W., Mildon, B., Morgante, F., Myers, L., Nicholson, C., Nielsen, G., Perez, D.L., Popkirov, S., Reuber, M., Rommelfanger, K.S., Schwingenshuh, P., Serranova, T., Shotbolt, P., Stebbins, G.T., Stone, J., Tijssen, M.A.J., and Tinazzi, M., 2020, Outcome measurement in functional neurological disorder: A systematic review and recommendations. *Journal of Neurology, Neurosurgery and Psychiatry*, **91**, 638–649. https://doi.org/10.1136/jnnp-2019-322180.

Prinsen, C.A.C., Mokkink, L.B., Bouter, L.M., Alonso, J., Patrick, D.L., de Vet, H.C.W., and Terwee, C.B., 2018, COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, **27**, 1147–1157. https://doi.org/10.1007/s11136-018-1798-3.

Ronen, G.M., Rosenbaum, P., and Streiner, D.L., 2000, Outcome measures in pediatric neurology: Why do we need them? *Journal of Child Neurology*, **15**, 775–780. https://doi.org/10.1177/088307380001501201.

Roulstone, S., Coad, J., Ayre, A., Hambly, H., and Lindsay, G., 2013, *The preferred outcomes of children with speech, language and communication needs and their parents.* (UK Department for Education Research report).

Royal College of Speech & Language Therapists, 2020, Outcome measurement [online]. Available: https://www.rcslt.org/speech-and-language-therapy/guidance-for-delivering-slt-services/outcome-measurement [Accessed 11 January 2020].

Schriberg, L.D., Austin, D., Lewis, B.A., McSweeney, J.L., and Wilson, D.L., 1997, The speech disorders classification system (SDCS): extensions and lifespan reference data. *JSLHR*, **40**, 723–40.

Sparrow, S., Cicchetti, D., and Balla, D., 2005, *Vineland Adaptive Behavior Scales: Survey*

Interview, 2nd edn. (Minneapolis: Pearson).

SPEECH-LANGUAGE & AUDIOLOGY CANADA, 2010, *Position statement on outcomes measures* (Canada: Ottawa).

SPEECH-LANGUAGE & AUDIOLOGY CANADA, 2012, SAC Position Paper: Early Identification of Speech & Language Disorders [online]. Available: http://www.sac-oac.ca/professional-resources/resource-library/early-identification-speech-language-disorders?_ga=1.225121811.849982984.1468963630.

SQUIRES, J., BRICKER, D., and TWOMBLY, E., 2003, *The ASQ-SE User's Guide: Ages and Stages Questionnaires Social-Emotional. A Parent-Completed, Child-Monitoring Systemfor Social-Emotional Behaviours.* (Baltimore: Paul H. Brookes Publishing Co.).

TATE, R.L., 2019, Measuring outcomes and monitoring progress in the era of evidence-based clinical practice. *Brain Impairment*, 276–288. https://doi.org/10.1017/BrImp.2019.28.

TERWEE, C.B., PRINSEN, C.A.C., CHIAROTTO, A., WESTERMAN, M.J., PATRICK, D.L., ALONSO, J., BOUTER, L.M., DE VET, H.C.W., and MOKKINK, L.B., 2018, COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, **27**, 1159–1170. https://doi.org/10.1007/s11136-018-1829-0.

THOMAS-STONELL, N.L., CUNNINGHAM, B.J., ROBERTSON, B., and ROSENBAUM, P., 2020, *The Focus on the Outcomes of Communication Under Six (FOCUS©) Manual, Second Edition* (Hamilton, ON: CanChild).

THOMAS-STONELL, N.L., MCCONNEY-ELLIS, S., ODDSON, B., ROBERTSON, B., and ROSENBAUM, P., 2007, An evaluation of the responsiveness of the Pre-Kindergarden ASHA NOMS. *Canadian Journal of Speech-Language Pathology and Audiology*, **31**, 74–82.

THOMAS-STONELL, N.L., ODDSON, B., ROBERTSON, B., and ROSENBAUM, P., 2009, Predicted and observed outcomes in preschool children following speech and language treatment: Parent and clinician perspectives. *Journal of Communication Disorders*, **42**, 29–42. https://doi.org/10.1016/j.jcomdis.2008.08.002.

THOMAS-STONELL, N.L., ODDSON, B., ROBERTSON, B., and ROSENBAUM, P.L., 2010, Development of the FOCUS (Focus on the Outcomes of Communication Under Six), a communication outcome measure for preschool children. *Developmental Medicine and Child Neurology*, **52**, 47–53. https://doi.org/10.1111/j.1469-8749.2009.03410.x.

THOMAS-STONELL, N.L., ODDSON, B., ROBERTSON, B., and ROSENBAUM, P.L., 2013, Validation of the Focus on the Outcomes of Communication under Six outcome measure. *Developmental Medicine and Child Neurology*, **55**, 546–552. https://doi.org/10.1111/dmcn.12123.

THOMAS-STONELL, N.L., ROBERTSON, B., WALKER, J., ODDSON, B., WASHINGTON, K.N., and ROSENBAUM, P., 2015, *FOCUS©: Focus on the Outcomes of Communication Under Six Manual* (Toronto: ON, Canada: Holland Bloorview Kids Rehabilitation Hospital).

THREATS, T.T., 2013, WHO's International Classification of Functioning, Disability, and Health: A framework for clinical and research outcomes. In *Outcomes in speech-language pathology* ((New York, NY)), pp. 58–72.

VAN TIGGELEN, H., KOTTNER, J., CAMPBELL, K., LEBLANC, K., WOO, K., VERHAEGHE, S., VAN HECKE, A., and BEECKMAN, D., 2020, Measurement properties of classifications for skin tears: A systematic review. *International Journal of Nursing Studies*, **110**, 103694. https://doi.org/10.1016/j.ijnurstu.2020.103694.

VALLINO-NAPOLI, L.D. and REILLY, S., 2004, Evidence-based health care: A survey of speech pathology practice. *International Journal of Speech-Language Pathology*, **6**, 107–112. https://doi.org/10.1080/14417040410001708530.

VARNI, J.W., 1998, The PedsQL Measurement Model for the Pediatric Quality of Life Inventory [online]. Available: http://www.pedsql.org/%0Aabout_pedsql.html [Accessed 3 March 2020].

WASHINGTON, K.N., ODDSON, B., ROBERTSON, B., ROSENBAUM, P., and THOMAS-STONELL, N., 2013, Reliability of the Focus on the Outcomes of Communication Under Six (FOCUS©). *Journal of Clinical Practice in Speech-Language Pathology*, **15**, 25–31. https://doi.org/10.1111/cch.12049.

WASHINGTON, K.N., THOMAS-STONELL, N., ODDSON, B., MCLEOD, S., WARR-LEEPER, G., ROBERTSON, B., and ROSENBAUM, P., 2013, Construct validity of the FOCUS© (Focus on the Outcomes of Communication Under Six): A communicative participation outcome measure for preschool children. *Child: Care, Health and Development*, **39**, 481–489. https://doi.org/10.1111/cch.12043.

WILCOX, K. and MORRIS, S., 1999, *Children's Speech Intelligibility Measure. 1999* (San Antonio: The Psychological Corporation Harcourt Brace & Company,).

WILLIAMS, B. and BEOVICH, B., 2020, A systematic review of psychometric assessment of the Jefferson Scale of Empathy using the COSMIN Risk of Bias checklist. *Journal of Evaluation in Clinical Practice*, **26**, 1302–1315. https://doi.org/10.1111/jep.13293.

WORLD HEALTH ORGANIZATION, 2001, *International classification of functioning, disability and health: ICF*. (Geneva).