

Research Article

Optimizing Outcomes for Children With Phonological Impairment: A Systematic Search and Review of Outcome and Experience Measures Reported in Intervention Research

Elise Baker,^{a,b,c}  Sarah Masso,^{d,e}  Kylie Huynh,^d  and Ellie Sugden^a 

^aSchool of Health Sciences, Western Sydney University, New South Wales, Australia ^bSouth Western Sydney Local Health District, Liverpool, New South Wales, Australia ^cIngham Institute for Applied Medical Research, Liverpool, New South Wales, Australia ^dFaculty of Medicine and Health, The University of Sydney, New South Wales, Australia ^eCharles Sturt University, Bathurst, New South Wales, Australia

ARTICLE INFO

Article History:

Received August 25, 2021

Revision received November 29, 2021

Accepted January 11, 2022

Editor-in-Chief: Holly L. Storkel

Editor: Sherine R. Tambyraja

https://doi.org/10.1044/2022_LSHSS-21-00132**ABSTRACT**

Purpose: Reporting of outcome and experience measures is critical to our understanding of the effect of intervention for speech sound disorders (SSD) in children. There is currently no agreed-upon set of measures for reporting intervention outcomes and experiences. In this article, we introduce the Speech Outcome Reporting Taxonomy (SORT), a tool designed to assist with the classification of outcome and experience measures. In a systematic search and review using the SORT, we explore the type and frequency of these measures reported in intervention research addressing phonological impairment in children. Given the integral relationship between intervention fidelity and intervention outcomes, reporting of fidelity is also examined.

Method: Five literature databases were searched to identify articles written or translated into English published between 1975 and 2020. Using the SORT, outcome and experience measures were extracted and categorized. The number of intervention studies reporting fidelity was determined.

Results: A total of 220 articles met inclusion criteria. The most frequently reported outcome domain was broad generalization measures ($n = 142$, 64.5%), followed by specific measures of generalization of an intervention target ($n = 133$, 60.5%). Eleven (5.0%) articles reported measures of the impact of the phonological impairment on children's activity, participation, quality of life, or others. Twenty articles (9.1%) reported on parent, child, or clinician experience or child engagement. Fidelity data were reported for 13.4% of studies of interventions.

Conclusions: The measurement of intervention outcomes is challenging yet important. No single type of measure was reported across all articles. Through using tailored measures closely related to intervention targets in combination with a universal set of measures of intelligibility, the impact of phonological impairment on children's lives, and the experience of receiving and providing intervention, researchers and clinicians could work together to progress insights and innovations in science and practice for children with SSD.

Supplemental Material: <https://doi.org/10.23641/asha.19497803>

Correspondence to Elise Baker: elise.baker@westernsydney.edu.au.

Publisher Note: This article is part of the Forum: Innovations in Treatments for Children With Speech Sound Disorders. **Disclosure:** The authors acknowledge that Elise Baker and Sarah Masso are the authors of the Speech Outcome Reporting Taxonomy referred to in the article. Elise Baker is a co-author of the book "Children's Speech: An Evidence-based Approach to Assessment and Intervention" published by Pearson Education as referred to in the article, and receives royalty payments on the sale of the book. The other authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

Speech sound disorders (SSD) are common and can impact children's lives in different ways. SSD can impact children's speech and their ability to be understood. SSD can also impact children's social-emotional well-being and their ability to fully participate with family, friends, and the communities to which they belong (e.g.,

McCormack et al., 2010). When children with SSD receive intervention from speech-language pathologists (SLPs), there is an expectation that children's speech, social-emotional well-being, and participation will improve. Expectations of intervention, however, are not the same as outcomes resulting from intervention.

Outcome measures are fundamental to clinical practice, and there is an increasing expectation for their use in demonstrating the effectiveness of services (Cunningham & Oram Cardy, 2020). Outcome measures can provide data-driven answers to clinical questions about a child's response to intervention, and they are an important source of evidence for funding bodies, insurers, and other payers to justify services and track resource allocation (Mullen & Schooling, 2010; Sandbank et al., 2021). The very process of reporting outcomes into a registry can also improve outcomes (Nelson et al., 2016). Measures of the experience of intervention are a valuable complement to measures of outcome. Experience measures offer insight into the quality of care provided during intervention and can assist with the interpretation of outcomes following intervention (Kingsley & Patel, 2017; Roulstone, 2015). In this article, we explore the issue of the measurement of intervention outcomes and experiences for children with SSD. We describe a taxonomy for categorizing types of outcome and experience measures. Using a systematic search and review, we then explore the type and frequency of measures reported in intervention research for children considered to have phonologically based SSD (hereafter referred to as *phonological impairment*). Children with phonological impairment have difficulty learning the phonological system of their language, with their speech characterized by "pattern-based errors impacting phonemes (including one or more classes of phonemes), phonotactics (e.g., syllable and word shapes) and/or prosodic characteristics (e.g., lexical stress)" (Baker et al., 2020, p. 375). We focus on phonological impairment as it is the most common type of SSD and can form a high proportion of children on SLPs' caseloads (e.g., Farquharson et al., 2020; Mullen & Schooling, 2010). Our focus on phonological impairment also extends and facilitates comparison of the work of Kearney et al. (2015), who reported a narrative review on outcome measures for SSD with a motor basis. Throughout this article, we use the term *phonological impairment* as a collective term for phonological delay, phonological disorder, and inconsistent phonological disorder. The use of the term *phonological impairment* also avoids potential misunderstandings with the term *SSD*, given that SSD has been viewed as both an umbrella term for different types of SSD (e.g., Williams et al., 2021) and a synonym for phonological impairment (e.g., Brosseau-Lapr e et al., 2020).

Measuring the Outcome of Intervention for Children With Phonological Impairment

The World Health Organization (1998, p. 10) defines *health outcome* as "a change in the *health status* of an individual, group or population, which is attributable to a planned intervention or series of interventions." The measurement of outcomes has historically been grounded in scientific methods—sampling and measuring dependent variables (e.g., children's speech accuracy) to determine if variables have changed in response to independent variables (e.g., intervention). Although consideration of speech accuracy is central to determining the outcome of intervention for children with phonological impairment, operationalizing what, how, and when dependent variables are measured can be challenging. Likewise, the proximity of the outcome measure to the target of the intervention can influence our interpretation of the extent of change induced by a direct intervention (Sandbank et al., 2021). For measures that are proximal (i.e., closely related to what is targeted during intervention), positive change could indicate that a specific skill has improved (e.g., increased percent accuracy on a single-word generalization probe of targeted word-initial consonant clusters). For measures that are distal (i.e., extending beyond the target and clinical contexts to everyday life), positive change could indicate that cascading negative impacts of phonological impairment on activity, participation, and quality of life have indirectly been addressed (e.g., parent's rating of their child's participation in everyday conversations with peers has improved). Given the breadth of what could be measured, an understanding of the underlying nature of phonological impairment and the impact of the impairment could help guide what outcomes to measure, when, and how.

Children who have a phonological impairment need to learn the phonological system of their language; they need to learn the phonemes in the language and the rules about how those phonemes can combine to form abstract representations of words to communicate meaning. In terms of measuring outcomes, the most close or proximal outcomes relative to intervention are those that directly measure the targets of intervention within the context of intervention. Through intervention, however, children's productive phonological knowledge grows. The intervention approach, intensity, and targets selected for intervention can influence the extent and rate of change. Selected targets are considered particularly important as they are a means to an end for facilitating measurable stimulus and response generalization (e.g., Gierut & Morrisette, 2015; Rvachew & Nowak, 2001). Stimulus generalization refers to the occurrence of a targeted behavior with stimuli that are not present during intervention (e.g., other people such as siblings or parents, other places such as home or

preschool, or other materials such as a book or toy not used in therapy sessions). Response generalization (specifically phonological response generalization) refers to the occurrence of change in a child's productive phonological knowledge and can include both generalization of a targeted behavior to other response contexts (e.g., from treatment words to nontreatment words, other word positions, conversational speech, and other linguistic domains) and generalization related to the target in other response contexts (e.g., within-class and across-class phonological generalization; McLeod & Baker, 2017).

The challenge of what aspects of generalization to measure is further complicated by the different ways that measures can be reported, the type of tool used to collect the measure, and the meaningfulness of the measure to others. For instance, the accuracy of a particular phoneme or class of phonemes from an informal single-word probe could be reported as a raw score (e.g., 19/20 correct) or percent accuracy (e.g., 95%). Such measures could be readily interpreted by SLPs as they convey meaning within context: The child is progressing well toward achieving a specific goal. Alternatively, measures could be less transparent and reported as (a) composite scores across a range of tasks, (b) measures from instrumental analysis, or (c) standard scores or percentile ranks from standardized tests. Less transparent measures need interpretation. Although relatively proximal to the target of intervention, composite scores and outcome measures from instrumental analysis may not be as meaningful to children and their parents. Transformed measures (i.e., a percentile rank or standard score) from a standardized norm-referenced assessment should also be viewed with caution and not encouraged. Such measures have been criticized as inappropriate for measuring outcomes due to serious limitations resulting in over- or underestimation of change (Kearney et al., 2015; Payne, 2002). For instance, change (or lack of change) can be an artifact of a child changing (or not changing) the age-band from which standard scores are derived. A child's performance on a norm-referenced assessment also needs to improve faster than typical maturation to show change in a transformed score (Payne, 2002). In addition, diagnostic assessment tools usually measure a wider breadth of skills (e.g., full range of consonants) some of which a child may have already achieved, thereby limiting opportunities to capture change in specific skills targeted in intervention from transformed measures.

Measures of intelligibility and the impact of intervention on children's everyday activity and participation (e.g., Cunningham et al., 2017; McLeod et al., 2012) can be more meaningful to parents and align with their reasons for referral. Although such measures are more distal from specific abilities targeted during intervention, they are valuable for capturing important cascading changes in children's lives that may occur in response to intervention

(Sandbank et al., 2021). In a study where parents and clinicians observed children's lives across the International Classification of Functioning, Disability and Health: Children and Youth (ICF-CY) domains of body function, activity and participation, and personal factors, Thomas-Stonell et al. (2009) discovered that "parents noted twice as many changes in the participation and personal factors domains as clinicians" following intervention (p. 29), thus highlighting the value of measuring children's activity and participation, rather than the impairment alone.

Unlike some areas of speech-language pathology where comprehensive sets of universally agreed outcome measures and registries exist (e.g., a standard set of outcome measures for the comprehensive appraisal of cleft care [Allori et al., 2017] and the international population registry for aphasia after stroke [Ali et al., 2021]), there is currently no agreed-upon universal set of measures or tools for measuring the outcome of intervention for children with SSD. In 2011, Baker and McLeod reported a narrative review of 134 studies of intervention for children with phonological impairment and noted that "there was wide variation in the measures used to report the outcomes of intervention...[and that] the diversity of measures made it difficult to draw meaningful conclusions across the studies with respect to the outcome of intervention" (p. 115). Moreover, "the wider impact of intervention on children's activity and participation was rarely reported" (Baker & McLeod, 2011, p. 115). In Kearney et al.'s (2015) narrative review of outcome measures from 66 studies of intervention for children with SSD with a motor basis, they too "identified a wide variation of measures used to document change following treatment in children with SSD with a motor basis" (p. 260). Although Kearney et al. attempted to sort through the variation by categorizing outcome measures as perceptual measures (from standardized norm-referenced tests, criterion-referenced procedures, or assessments of intelligibility) and instrumental measures, the proportion of studies measuring the impact of intervention on children's activity and participation was not considered. They recommended that clinicians "choose an outcome measure" (p. 260) that aligned with the underlying nature of the child's motor speech difficulty. Although SLPs could read through individual research articles to identify outcomes measures, greater guidance is needed about the types of outcome measures reported across peer-reviewed published intervention research and measures that could be relevant for research and practice contexts. The need for such guidance is further illustrated in the findings from a recent review of case notes across three National Health Service SLP services in the United Kingdom for school-age children with SSD. Specifically, SLPs' documentation of outcomes varied, with data from post-intervention speech assessments only evident in 21.8%–50.0% of case notes (Morgan et al., 2021). Morgan et al. concluded that SLPs'

collection and documentation of intervention outcomes in routine practice was generally insufficient for determining intervention effect. Greater guidance is needed.

One solution to the challenge of what to measure is the use of rating scale–based outcome measure systems. For instance, the American Speech-Language-Hearing Association (ASHA) developed and encourages SLPs to use their National Outcomes Measurement System (NOMS) registry. This registry includes options for SLPs to report functional communication measures, patient-reported outcome measures, measures of patient satisfaction, and aspects of service delivery such as number of individual and group sessions, average session length, and whether services are delivered via telepractice (ASHA, 2021). The Royal College of Speech and Language Therapists' (2019) Online Outcome Tool (ROOT) is a similar distal rating scale outcome measurement system. Based on the Therapy Outcome Measures (TOMs) by Enderby and John (2015), ROOT captures SLPs' ratings of a client's degree of impairment, activity, participation, well-being/distress, and carer well-being/distress. The Australian TOMs (Perry et al., 2004) use similar distal outcome measures capturing impairment, activity, participation, and well-being based on the TOMs.

The majority of these rating systems are clinician-reported outcomes as they rely on observations and judgments by the clinician. Increasingly, parent- or patient-reported outcome measures (PROMs) are being called for as they can capture the more subjective, unobservable outcomes known to their clients rather than SLPs (Cohen & Hula, 2020). ASHA's revised NOMS includes a patient-reported outcome form for pediatric communication for children from 3 to 17 years (ASHA, 2020). The Focus on the Outcomes of Communication Under Six (FOCUS; Thomas-Stonell et al., 2010) is a validated tool designed specifically for parents to report on real-world changes in young children's communication outcomes, with a particular focus on children's abilities to participate. Broader generic PROMs such as the PROMIS Pediatric Scale Version 1.0 (PROMIS Health Organization, 2016) consider broader aspects of a child's quality of life. As noted, in the narrative review of 134 studies by Baker and McLeod (2011), measures of activity and participation were rarely reported. The extent to which such measures have been included over the past decade is unknown.

Measuring the Experience of Intervention for Children With Phonological Impairment

Measures that can assist with the interpretation of outcomes include measures of the intervention experience. Experience measures, typically referred to as patient-reported experience measures (PREMs), capture children and parents' views on their experience of the process of

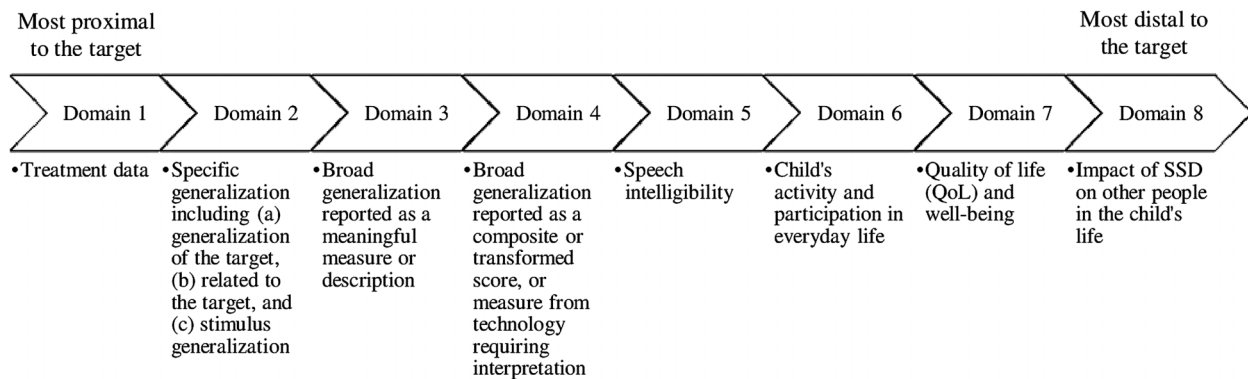
receiving care rather than the outcome of care (Kingsley & Patel, 2017). They are valuable to gather because parents and children's views about the intervention experience are not always positive and have the potential to impact on outcomes (Crowe et al., 2017; Roulstone, 2015). PREMs are usually gathered from questionnaires or interviews and can be classified as either relational (i.e., experience of the therapeutic relationship) or functional (i.e., practical experience of receiving intervention; Kingsley & Patel, 2017). Children's experience of intervention can also be captured indirectly via measures and observations of their engagement during intervention (Shriberg et al., 1990). Wray et al. (2018) codesigned a PREM questionnaire suitable for children receiving inpatient hospital services capturing relational insights (e.g., "How friendly were the people working at the hospital?") and functional insights (e.g., "How comfortable were the beds?"). It is currently unknown if or how often similar types of PREMs have been codesigned and used in research with children with phonological impairment.

Speech Outcome Reporting Taxonomy

To meet increasing expectations that clinical outcomes of intervention for children with phonological impairment are documented and that parents and children provide their perspectives of the outcome and experience of receiving intervention, greater guidance is needed. The Speech Outcome Reporting Taxonomy (SORT; Baker & Masso, 2021) is a system for organizing types of outcome and experience measures when managing SSD in children (see the Appendix). The SORT was developed by the first and second authors through a collaborative and iterative process by considering literature on outcome measures in research and practice (Cunningham et al., 2017; Enderby & John, 2015, 2019; Kearney et al., 2015), ways for monitoring and measuring generalization (Williams et al., 2021), and the concept that outcome measures can exist along a continuum from proximal to distal (Sandbank et al., 2021).

The SORT comprises eight outcome measurement domains. Figure 1 provides a summary overview of these domains. Outcome Domain 1 captures the most proximal type of data—treatment data. Treatment data could be further categorized as measures of speech production, time to achieve a predetermined goal or intervention criteria, and other abilities targeted during treatment (e.g., speech perception, phonological awareness, and expressive morphology). Domain 2 focuses on generalization. Given that generalization in phonological impairment in children is complex, Domain 2 is divided into three subdomains, each with the potential for the three categories of speech production, time, and other abilities. As depicted in Figure 1, Domain 2 includes (a) specific response generalization of the target (e.g., generalization of a target from treatment

Figure 1. Overview of the outcome measurement domains of the Speech Outcome Reporting Taxonomy. SSD = speech sound disorders.



words to nontreatment words or conversation), (b) specific response generalization related to the target (e.g., measurement of other implicationally related errors that are within or across classes, or specific errors not expected to change but closely monitored as a control behavior), and (c) stimulus generalization (e.g., measurement of the target behavior in other naturalistic contexts such as talking with family at home or friends at preschool).

Domain 3 refers to broader measures of generalization of children's overall phonological system or other abilities of interest (e.g., phonological awareness and expressive morphology) based on readily meaningful measures. Examples include percentage of consonants correct (PCC; Shriberg et al., 1997) and descriptions of children's overall phonological knowledge (e.g., total number of phonemes or sound classes within their system). By contrast, transformed and composite scores such as the target composite index (Tyler et al., 2006) and measures from technology (e.g., 3400 Hz as a formant frequency measure from acoustic analysis) are more distal and require interpretation to know what they mean; such measures are captured in Domain 4. Like Domains 1 and 2, measures within Domains 3 and 4 can be further categorized as direct measures of speech production, other abilities, or time to achieve a predetermined goal or intervention criteria. Domain 5 refers to measures of speech intelligibility, such as the number of intelligible words on a single-word test, intelligibility ratings completed by others (e.g., McLeod et al., 2012), and time to achieve a predetermined measure of intelligibility.

Outcome Domains 6 through 8 are measures or descriptions of the impact of a phonological impairment on children's lives and the lives of others. In keeping with the ICF-CY developed by the World Health Organization (2007), Domain 6 refers to measures or descriptions of children's activity and participation in everyday life. Domain 7 reflects broader measures or descriptions of children's quality of life and well-being, whereas Domain

8 captures measures or descriptions of the impact of children's impairment on other peoples' lives such as family members (e.g., parents and siblings) and others (e.g., educators and friends).

Contextually related to Outcome Domains 1 through 8 are measures of intervention experience and intervention fidelity. Experience measures include reports of child engagement during intervention, PREMs (i.e., parent/carer or child) of receiving intervention, and clinician-reported experience measures of providing intervention. In keeping with Wray et al. (2018), PREMs can be further categorized as relational and/or functional measures. Implementation of an intervention is captured via measures or descriptions of intervention fidelity. Fidelity refers to the degree to which an intervention is implemented in accordance with developers' intentions (Kaderavek & Justice, 2010). Measures of fidelity are included in the SORT as they are integral to interpreting intervention outcomes and avoiding Type III errors—reporting findings as significant or nonsignificant when the intervention was not implemented as planned (Bellg et al., 2004). Moreover, when researchers take steps to ensure high levels of fidelity, interventions can have larger effect size, and the rate of positive outcomes can increase (Durlak & DuPre, 2008). Together, the SORT provides a way of synthesizing and categorizing measures of intervention (see the Appendix).

Purpose

There is currently no universal set of measures or tools for evaluating the outcome and experiences of intervention for children with phonological impairment. Although a wide range of proximal outcome measures has been reported across empirical research, the diversity and frequency of use of these measures across published research is unknown. The extent to which measures of activity and participation have been reported is also unknown. Moreover, insight into

measures suitable for capturing intervention experience is lacking. Insight into the proportion of studies reporting measures of fidelity is also wanting. If clinicians and researchers are to move forward in documenting and using outcome and experience measures, there is a need to determine what has been reported across the empirical research. We need to learn from what has been done to build on and develop greater consensus and insight about what to measure, how, and when. This is because outcome and experience measures guide clinical decisions and monitoring of children's responses to intervention, drive benchmarking, promote research, and ensure accountability of funding from government and third-party payers (Kampstra et al., 2018).

Therefore, the purpose of this article was to use the SORT in the conduct of a systematic search and review of the type and frequency of outcome and experience measures documented in peer-reviewed, published intervention research for children with phonological impairment. According to Grant and Booth (2009), a systematic search and review is suitable over other forms of review when the breadth of published intervention research needing to be considered comprises a variety of research designs. Our specific research questions were as follows.

1. Across peer-reviewed published articles of intervention for children with phonological impairment,
 - (a) what proportion of outcome measures are reported for each SORT domain and category?
 - (b) what is the most frequently reported SORT outcome domain?
 - (c) what proportion of articles report direct outcome measures (Domains 1–5) across more than one domain of the SORT?
 - (d) what proportion of articles report measures of the impact of the child's impairment (Domains 6–8)?
 - (e) what proportion of articles report measures or descriptions of the intervention experience?
 - (f) what proportion of articles report measures of the fidelity of intervention implementation?
2. What types of outcome and experience measures and tools are reported across peer-reviewed published intervention research?

Method

Our methods for the systematic search and review followed the Preferred Reporting Items for Systematic Review and Meta-Analyses statement (Moher et al., 2009).

Search Strategy

A comprehensive search of five online databases was conducted (Linguistics and Language Behavior Abstracts, Scopus, Medline, speechBITE, and Education Resources Information Center) using the following search terms: phonological OR phonology OR articulation OR speech sound disorder, AND intervention OR therapy OR treatment, AND child OR children. The online databases were composed of literature for speech-language pathology, science, medicine, and education. The search strategy replicated the search terms used in the review by Sugden et al. (2018). Publication titles, abstracts, and key words were searched, with a limit placed on publications between the years 1975 and 2020, with articles published in 2020 having volume and page numbers (as opposed to being published early online). Additionally, a hand search was conducted by reviewing the reference lists of all articles included in the final review.

Inclusion Criteria

Articles were included if they were peer-reviewed intervention studies written or translated in English and published between 1975 and 2020. Studies included the following research designs: randomized controlled trials (RCTs), non-RCTs, quasi-experimental group designs, single-case experimental designs (SCEDs), case series studies, and case studies. Study participants included children (aged under 18 years) with phonological impairment, delay, or disorder. Additionally, intervention studies that used a different diagnostic label (e.g., speech sound disorder) but assessed or analyzed the children's speech phonologically (i.e., identified pattern-based errors) were included.¹

Exclusion Criteria

Articles were excluded if they were review studies, not peer-reviewed sources, or sources reporting on results that has already been published in previous articles and did not present new data. This review also excluded articles focused on children with SSD of known organic origin (i.e., cleft palate, sensorineural hearing loss, and Down

¹In this systematic search and review, the decision to include intervention articles addressing phonological impairment in children required consideration not only of the diagnostic labels but also of how researchers described the nature of the children's speech difficulty and the goals, purpose, and type of intervention provided. This was due to the diversity of diagnostic labels for phonological impairment over the past 4 decades. In addition, by focusing on interventions addressing phonological impairment in children, this meant that some non-phonological interventions such as morphosyntax intervention exploring cross-domain phonological generalization (e.g., Tyler et al., 2002) were also included.

syndrome), or children who were diagnosed with an articulation impairment who received an articulation intervention, or children who had a motor speech disorder such as childhood apraxia of speech, childhood dysarthria, or motor speech delay who received intervention specifically for that type of SSD.

Article Selection

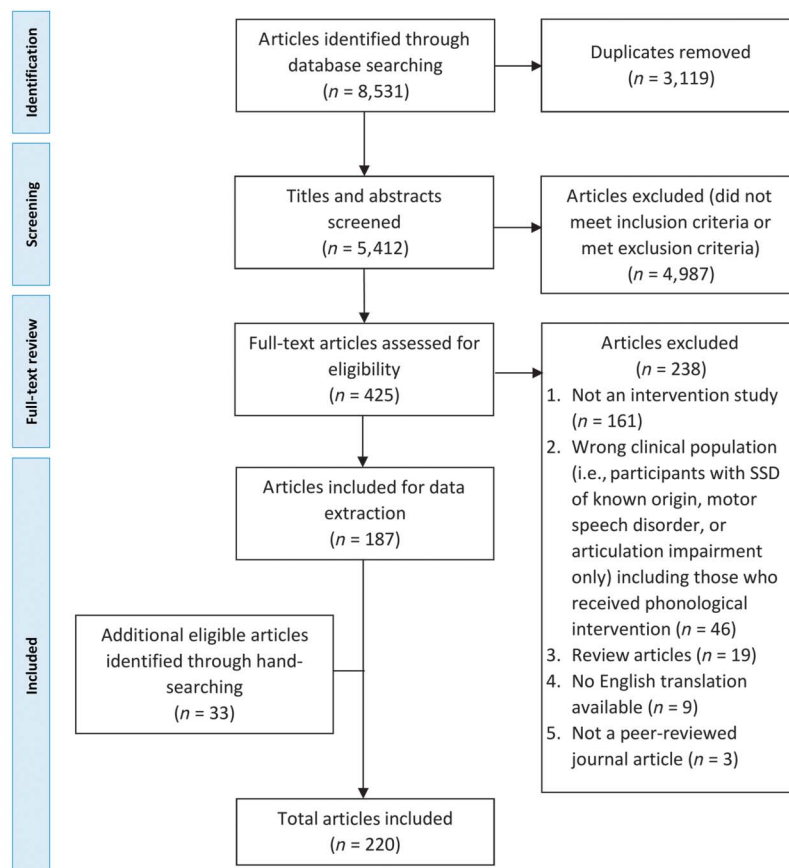
Following completion of the search strategy, citations from all databases were downloaded and managed using EndNote X8 (reference manager software; The EndNote Team, 2018). Citations were uploaded onto Covidence (systematic review software; Veritas Health Innovation, n.d.). Duplicate articles were automatically removed by EndNote and Covidence. The third author screened the titles and abstracts of all articles identified by the database search according to the inclusion/exclusion criteria. Reliability of this screening was conducted by the first and third authors (see below). The first and third authors then reviewed the full text of articles identified through screening. Articles were excluded if they did not

meet one or more of the specified inclusion criteria, or if they met one or more of the exclusion criteria. The first and fourth authors then rechecked the included articles according to the criteria to confirm eligibility for inclusion in the review. Disagreements were resolved by consensus (see Figure 2 for a flow diagram of the article inclusion process).

Data Extraction and Analysis

A range of data were extracted for each article and entered into a Microsoft Excel spreadsheet (Microsoft Corporation, 2021). Demographic data including reference, year of publication, intervention approach, study design, participant numbers, and age range per article were extracted by the second and third authors. Outcome and experience measurement data were extracted and categorized within SORT by the first and second authors. Following a period of training on 10% of articles, these two authors extracted data for 111 and 109 articles, respectively. If the authors had any doubts about how to classify any outcome measures within an article using the SORT,

Figure 2. Preferred Reporting Items for Systematic Review and Meta-Analyses flow diagram. SSD = speech sound disorders.



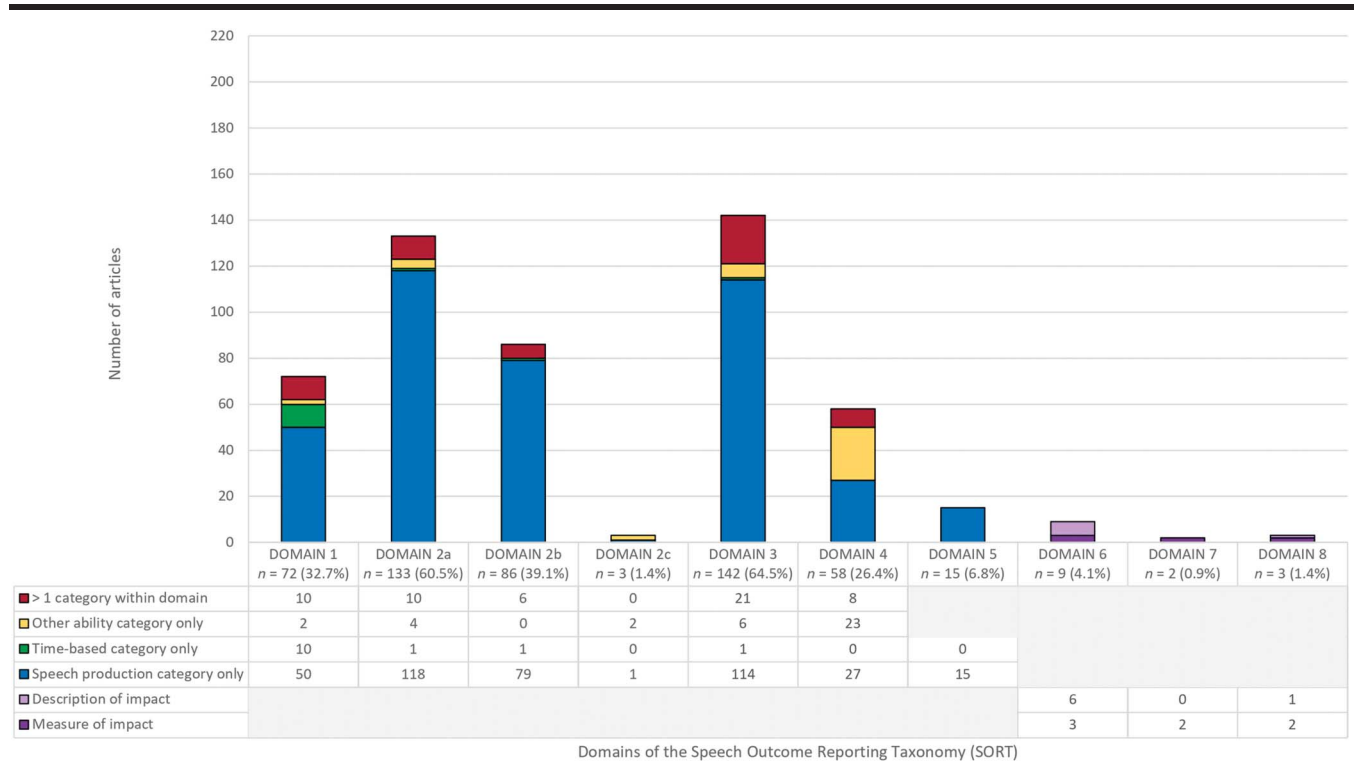
these articles were discussed, and data extraction was achieved by consensus. This occurred for 41 articles. The need for discussion typically was associated with decisions regarding whether outcome measures reflected Domain 2 or 3. In addition to this process, the first two authors also independently reextracted SORT data for 11% of articles. The reliability of data extraction is reported below. Finally, the first, second, and fourth authors used the Template for Intervention Description and Replication (TIDieR) checklist (Hoffmann et al., 2014) to rate the quality of the reporting of the interventions. They rated 60, 110, and 136 of the 306 interventions reported across the 220 articles, respectively. In keeping with the reviews by Sugden et al. (2018) and Baker and McLeod (2011), if one article included two or more interventions, then the TIDieR ratings (Hoffmann et al., 2014) for each intervention were entered separately into the spreadsheet. Across the 12 items on the checklist, reporting of information is tallied such as intervention title, materials, procedures, and fidelity. The outcome data were synthesized using frequency and summary counts in Excel and are presented in Figure 3. The coding of outcome measures for each

article (for each domain of the SORT) is tabulated in Supplemental Material S2, and that for the experience measures is tabulated in Supplemental Material S3.

Reliability

Three aspects of reliability were determined. First, interjudge reliability was undertaken for the screening of articles to be included in the review. The first and third authors screened titles and abstracts of 577 articles against inclusion and exclusion criteria. Interjudge agreement as measured by Cohen's kappa was 0.98. Interjudge reliability of the data extraction and classification according to the domains and categories within SORT was conducted by the first and second authors for 25 (11%) articles, equivalent to 698 decision points. Using Cohen's kappa, interjudge agreement was 0.81. Third, reliability of the coding of the articles using the TIDieR was completed by the first and fourth authors on 25 (11%) articles. Using Cohen's kappa, interjudge agreement was 0.92. All kappa values reflected strong agreement (McHugh, 2012). All disagreements were resolved by consensus.

Figure 3. Outcomes reported across the eight domains of the Speech Outcome Reporting Taxonomy.



KEY: Domain 1 = Treatment data; Domain 2a = Specific response generalization of the target to other response contexts; Domain 2b = Specific response generalization of related targets to other response contexts; Domain 2c = Stimulus Generalisation; Domain 3 = Broad generalization using a meaningful measure or description; Domain 4 = Broad generalization using a composite or transformed score; Domain 5 = Speech intelligibility; Domain 6 = Child's activity and participation in everyday life; Domain 7 = Child's quality of life; Domain 8 = Impact of child's SSD on other people in the child's life. (Note: If an article reported more than one outcome category within a domain [e.g., speech production measure such as percentage of consonants correct (PCC) and measure of another ability such as percent finite morphemes correct in conversation], this was captured separately as '> 1 category within domain'.)

Results

Article Selection

A systematic search of the five databases yielded 8,531 articles, and after screening, full-text review, and hand-searching, 220 articles met criteria. A total of 306 interventions were studied across the 220 articles. The mean TIDieR percent score for quality of reporting about the interventions was 61.0% (range: 9.1%–100%). Intervention studies with higher scores contained more information to support replication.

Article Characteristics

The 220 included articles were published across 41 journals from 1981 to 2020. The distribution of articles according to year of publication is depicted in Supplemental Material S1. On average, five articles were published a year, with a range of 0–13 articles per year. A variety of research designs were reported across the articles, including case studies and case series, SCEDs, quasi-experimental within-group design, non-RCTs (between-groups designs), and RCTs.

Participant Characteristics

A total of 4,151 children served as participants across the articles. Excluding outliers for one article (specifically Broomfield & Dodd, 2011; $n = 730$), the average number of participants per article was 16 (range: 1–159). This total and average number of participants includes children whose findings were reported across more than one article. The exact number of children involved in more than one article was unknown. The participant mean age range was 49.4–68.2 months.

Outcome Measures Reported According to SORT Domains and Categories

Regarding our first question, Figure 3 presents the number and proportion of outcome measures reported for each SORT domain and category. As depicted, across the 220 articles, Domain 3 (i.e., broad measures of generalization within meaningful contexts) was the most frequently reported domain ($n = 142$, 64.5%), with the majority of articles in this domain reporting outcomes for the category speech ($n = 135$, 61.2%). Outcome for Domain 2a focused on generalization of the intervention target was the second most frequent type of outcome reported ($n = 133$, 60.5%), with the majority reporting for the category speech ($n = 128$, 58.2%). Generalization data from less transparent measures of generalization (Domain 4) were reported for 58 articles (26.4%), with 11 of these 58 articles reporting transformed scores (e.g.,

standard scores or percentile ranks from norm-referenced assessment tools). Of these 11 articles, 10 included outcomes measures for other domains to support interpretation. Treatment data (Domain 1) were reported for 72 (32.7%) articles. Regarding the proportion of articles reporting direct measures (Domains 1–5), one article did not report any direct outcome measures, 73 (33.2%) articles reported outcome measures for one domain, 90 (40.9%) reported for two domains, 46 (20.9%) reported for three domains, and 10 (4.5%) reported for four domains.

Eleven articles (5.0%) reported outcomes on the impact of a child's impairment (i.e., Domains 6–8). These articles were published from 1990 to 2020, with two published in the past decade. Of the 11 articles, a total of 14 measures or descriptions were reported, as three articles reported measures for two domains. For everyday activity and participation (Domain 6), nine articles reported outcomes (three reported measures and six descriptions). Two articles reported outcome measures for quality of life (Domain 7), and three reported outcomes regarding the impact of a child's SSD on others (two reported measures and one description) (Domain 8). Supplement Material S2 provides the coding for all 220 articles for each domain and category.

Reporting of Measures of Experience and Intervention Fidelity

Across the 220 articles, 20 (9.1%) reported one or more measures or descriptions of the experience of intervention, with seven articles reporting child engagement, 14 articles reporting child and/or parent experience, and five articles reporting on the clinician experience of providing intervention (refer to Supplemental Material S3). Finally, measures of intervention fidelity were reported for 41 interventions (from a total of 306 interventions reported across the 220 included articles), equivalent to 13.4% of interventions.

Exploration of the Types of Measures and Tools Reported Across Articles

With regard to our second research question to explore the types of measures and tools, a plethora were reported across the 220 articles. Comprehensive synthesis of specific outcome measures and tools within and across articles was complicated as measures and tools were combined in different and unique ways to address specific research questions. For instance, Tyler et al. (1987) reported participants' percent accuracy on informal single-word probes that sampled each sound in a sound class affected by a phonological process for each word position in which the process occurred. The participants' performance on the probes was reported as percent accuracy for

treated and untreated processes over time, equivalent to Domains 2a and 2b (response generalization of treated and untreated processes). In contrast to this article reporting outcomes within a domain, other articles reported outcomes within and across multiple domains. For example, Lee (2018) reported (a) percent correct production of word-initial target phonemes in nontreatment words (Domain 2a) and related word-initial nontarget phonemes (Domain 2b) from an informal single-word probe, (b) overall phonemic inventory pre- and posttreatment from the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2; Goldman & Fristoe, 2000) (Domain 3), (c) the number/value of standard deviations from participant standard scores that were away from the mean on the GFTA-2 (Domain 4), and (d) ratings of overall intelligibility using the intelligibility rating scale of 1–6 by Fudala (2000; Domain 5).

One type of measure that was frequently mentioned across articles was PCC ($n = 67$, 30.5%). The PCC measure or derivatives of this measure were obtained from a variety of sampling tools and contexts, including unpublished and published word lists, conversational speech samples, narrative samples, and standardized articulation and phonological assessment tools. Rvachew et al. (2004) reported a derivative of PCC referred to as “PCC-difficult” (which focused on 13 consonants not mastered by most of the children in their study) in an effort to use a more sensitive metric of change rather than using a PCC score that includes consonants that are already accurate pretreatment and so not expected to change. Smit et al. (2018) reported similar derivatives of “PCC-type measures and inventory measures, including PCCList [list of single words], PCCconv [conversational speech], PCCList for late/difficult consonants, PCCconv for late/difficult consonants, and size of the late/difficult inventory derived from a word list and from a conversation” (p. 538).

Of the 11 articles (5.0%) reporting measures or descriptions of impact across Domains 6–8, five different tools were used to measure impact on children’s everyday life activity and participation (Domain 6), two tools were used to measure quality of life (Domain 7), and two tools were used to measure the impact on others (Domain 8). For an overview of the specific tools used by researchers who reported measures for Domains 6–8, readers are referred to Supplemental Material S2.

Of the 20 articles reporting experience measures, four measured engagement via structured observations during intervention sessions (e.g., Shriberg et al., 1990), and three provided descriptions of the child’s behavior (e.g., Culatta et al., 2005). Regarding measures or descriptions of the intervention experience from children, parents/carers, or clinicians, the majority (75%, $n = 12/16$ articles) used unpublished questionnaires designed for their study (e.g., Rvachew & Nowak, 2001) or collected anecdotal feedback exploring relational and/or functional

experiences, with four articles reporting PREMs from published tools. Refer to Supplemental Material S3 for a list of the informal tasks or procedures and published tools used across the 20 articles reporting experience or engagement.

Discussion

Evaluation of the outcome and experience of interventions for children with phonological impairment is fundamental to research and clinical practice. Outcome measures need to capture the extent of change in children’s speech production skills and/or other abilities targeted during intervention. They also need to capture improvements in the cascading impacts of phonological impairment on children’s everyday life activities and participation (Cunningham et al., 2017). Experience measures need to capture both relational and functional experiences of receiving intervention (Kingsley & Patel, 2017). Using the SORT, the type and frequency of outcome and experience measures documented across 220 peer-reviewed published articles of interventions for children with phonological impairment were identified. No single type of outcome or experience measure was documented across all articles. Rather, different types of outcomes directly measuring speech function were reported along a continuum from proximal measures of treatment data through to more distal measures of generalization. The most frequently reported domain or type of outcome measure across the articles was Domain 3 providing broad readily meaningful measures of generalization ($n = 142$, 64.5%) such as PCC from a single-word test or conversational speech sample. Outcome measures related to the impact of phonological impairment on children’s lives and experience measures were less frequently reported. In this discussion, we contextualize the results with previous reviews and reflect on the challenge of varied and disparate measures in research when professional organizations and clinicians are striving for a standard set of measures. We also consider the unique implications of the findings for researchers and clinicians.

Publication and Reporting of Outcome Measures: Similarities and Differences for Children With Different Types of SSD

There are similarities and differences with the findings from our review with previous related reviews (Cunningham et al., 2017; Kearney et al., 2015). Our review focused on outcome and experience measures for children with phonological impairment. To date, a review exclusively focused on these two broad types of measures has not been done. Our comparison with previous research therefore focuses on outcome measures.

First, most of the outcomes reported in the current review for Domains 1–5 were reliant on perceptual measures and clinical judgment. Although Kearney et al. (2015) reported a similar trend, they did note a wider range of measures reliant on various forms of technology not reported across articles in the current review (e.g., ultrasound tongue shape; tongue–palate contact patterns using electropalatography; speech kinematics measures such as jaw distance traveled and opening duration). In the current review, the use of specialized technology to gather measures was primarily limited to acoustic measures (e.g., MacLeod & Glaspey, 2014). These differences in the types of outcomes and technological tools used to measure outcomes perhaps reflect a difference between phonological impairment and motor speech disorders and the need for proximal measures to be tailored to the characteristics of each speech sound disorder.

Second, the proportion of articles reporting measures of intelligibility in this study (6.8%) was lower than the proportion reported by Kearney et al. (2015). They found that 19.7% of articles reported change in speech intelligibility as an intervention outcome. This may reflect a different focus on the goals and measures of interest by the researchers, with more articles in the current review focusing on aspects of children’s phonological systems (e.g., occurrence of phonological processes and PCC) rather than intelligibility. By contrast, Kearney et al. included articles targeting childhood motor speech disorders with an explicit focus on treating and measuring intelligibility. Another reason could be the way intelligibility has been conceptualized in the field and therefore categorized within SORT. Specifically, some authors have referred to intelligibility as a measure of speech function and therefore impairment (e.g., Cunningham et al., 2017)—measures captured in Domain 5. Others have included a rating of intelligibility as part of a combined measure of the impact of the impairment on a child’s every life activity and participation (e.g., ASHA, 2019; Enderby & John, 2019)—measures captured in Domain 6.

Third, a finding that was similar between the current review and Kearney et al.’s (2015) study was the low proportion and type of measures and tools used to capture the impact of phonological impairment on children’s lives and the lives of others (i.e., Domains 6–8). In the current review, only 5.0% of articles reported descriptions or measures for these domains. In the review by Kearney et al., the proportion was 4.5% (from 66 articles). A tool used to report those measures was common across both reviews—the FOCUS (Thomas-Stonell et al., 2010). In a review of the types of outcome measures and tools used to measure outcomes guided by the ICF-CY for preschoolers with communication impairments, Cunningham et al. (2017) noted “a dearth of measures in the pediatric speech-language literature that address Participation-based outcomes” (p. 447). The fact

that fewer measures in Domains 6–8 were reported could reflect not only the lack of available measures but also a reduced willingness to use these measures and/or understanding of the benefits of measuring change on the impact of phonological impairment on children’s lives. Such measures may also be seen by clinicians and researchers as inherently more difficult or less convenient to obtain as they rely on clients and/or parents. It could also be that a shift from an impairment-based model to a biopsychosocial model around the time of the publication of the ICF-CY (World Health Organization, 2007) occurred after the publication of 65.5% of the articles included in this review. Finally, a diversity of measures was identified in both the current review and Kearney et al.’s review. We explore this diversity in the next section.

The Challenge of Disparate Outcome Measures

Across the 220 articles, the type of outcome measures used to evaluate the effect and experience of intervention varied. One of the key reasons for this variation was that researchers needed to answer research questions unique to their study. Consider the outcome measures from the following two articles. Gierut and Morrisette (2015) were interested in discovering the effect of exposing children to words comprising dense neighbors versus non-dense neighbors in two different exposure contexts—before production activities to test a priming hypothesis and after production activities to test a phonological working memory hypothesis. Evaluation of the effect of these exposure contexts on phonological generalization was achieved by measuring accuracy of the treated sounds in treated words and nontreatment words on a single-word probe and by examining “percent gain in accuracy of production of sounds excluded from the phonemic inventory relative to baseline performance” in a single-word probe (Gierut & Morrisette, 2015, p. 1044). These measures, reflecting Domains 1, 2, and 3, revealed that exposure to neighbors prior to production training (i.e., priming) led to greater phonological generalization. In their study, single-word probes were individualized to measure potential change in each child’s phonological system. If a generic broad measure of PCC from a standardized single-word naming test or conversational speech sample was used, it may have been less sensitive to answer their research questions.

Rvachew and Brosseau-Lapr e (2015) provide another helpful illustration of the reason for the variation in outcome measures and tools across research. They were interested in the effect of different combinations of intervention approaches involving output-oriented individual intervention, articulation practice home program, dialogic reading home program, and input-oriented individual intervention on children’s speech production accuracy and phonological

awareness in French-speaking children with phonological impairment. Drawing on theoretical constructs in multi-linear phonology, Rvachew and Brosseau-Lapré used two types of measures to detect change in children's phonological systems with respect to segments, features, and prosodic structures and the interactions between these levels of the phonological hierarchy. Using a single-word picture-naming task, they measured PCC in addition to a measure of the child's targeted sounds matched ratio for features and word structures. The rationale for the latter measure was to address a limitation of broad measures of PCC—consonants children can already produce accurately before intervention and may be insensitive to change (Smit et al., 2018). Rvachew and Brosseau-Lapré also reported the number of correct responses on their measure of phonological awareness.

Could Rvachew and Brosseau-Lapré (2015) have used outcome measures and tools identical to Gierut and Morrisette (2015)? Their research questions, theoretical perspective, target abilities, and target language were all different. In each case, they needed to use measures suited to their study. If all researchers used the same measures (e.g., all used PCC from a 10-min conversational speech sample), research discoveries could be stifled. Does this mean that researchers should not strive for some degree of commonality or standard set when measuring outcomes for children with phonological impairment? No. Rather, we recommend that researchers and clinicians use a combination of unique measures suited to their specific research questions and/or client goals (Domains 1–4) coupled with common yet more personally meaningful distal measures for children and their families: measures evaluating intelligibility, the impact of SSD on children's everyday activity and participation, quality of life, and the impact on others (Domains 5–8). This idea aligns with efforts in the United States to encourage clinicians to participate in the NOMS registry (ASHA, 2020) and in the United Kingdom to encourage clinicians to use the ROOT system (Royal College of Speech and Language Therapists, 2019). Increasing the transparency of reporting outcome measures and the framework within which outcome measures are conceptualized (i.e., using the SORT) may also increase the transparency, replicability, and clinical applicability of different outcome measures specific to different interventions.

Outcome and Experience Measures: Implications for Researchers

The findings from this systematic search and review have four implications for researchers. First, given that only 11 (5.0%) articles included distal outcome measures or descriptions of the impact of intervention on children's lives and the lives of others, more researchers need to

routinely include measures using tools such as the FOCUS (Thomas-Stonell et al., 2010) or the TOMs (Enderby & John, 2019) in intervention research protocols. This recommendation aligns with Kearney et al. (2015), Smit et al. (2018), and Cunningham et al. (2017). When the year of publication of these measures was examined in the current review, there was no indication of a more recent increase in the use of such measures, as only two of the 11 articles reporting measures in Domains 6–8 were published in the past decade. There needs to be a greater appreciation about why these measures are valuable to collect in intervention research. They can help researchers determine if additional cascading benefits of their intervention have occurred (Cunningham et al., 2017).

A second implication for researchers is that information about measures and the tools used to collect the measures needs to be clearly reported. In the process of conducting the current review, data extraction and coding of outcome measures for 41 (18%) articles required discussion and consensus between the first two authors. These discussions predominantly focused on the distinction between Domains 2 and 3 where generalization probes were poorly described and/or difficult to interpret. Information in the method section of intervention studies needs to be clear about what the measures are, the purpose of the measure, the tools used to collect the measures, and if the person collecting the measures was blinded to the intervention. According to Smit et al. (2018), if an assessor is not blinded, pre-intervention measures can be prone to expectations that a child's speech is worse than it really is, and at post-intervention, it is better than it really is (i.e., the halo effect). Blinded assessors would therefore help minimize the potential for Type I errors. Reporting of the proximity of different measures to the intervention target would also increase transparency of reporting and interpretation of research findings.

Given that less than 10% of articles reported measures or descriptions of child, parent, or clinician experience of intervention, researchers are encouraged to include such measures as part of their research protocols. Across the articles that did report an experience measure, five different types of tools were used, which could serve as a guide for the researchers (see Supplemental Material S3 for further information). Of note, however, none of these tools were specifically designed to measure the experiences of children and parents in receiving phonological intervention or the experience of clinicians (or others) in providing intervention. This is an underexplored area of research and warrants engagement with stakeholders to codesign tools that could measure relevant aspects of the intervention experience (McNeill et al., 2021). As part of this research and in future intervention studies, it would be important to address the need for better quality and consistency of reporting of participant information (e.g., demographic,

developmental, and phonological profiles) given that lack of reporting about participants hampers the application of research to practice (Release Collaboration, 2020).

A final implication is the need for researchers to report intervention fidelity. Across the 306 interventions reported across the 220 articles, data on intervention fidelity were reported for 13.4% of interventions. This figure was similar to findings by Hinckley and Douglas (2013), where 14% of 149 aphasia intervention studies reported information about intervention fidelity. Although it could be argued that reporting of intervention fidelity was not expected in research published during the 1980s and 1990s and that reporting would be difficult for retrospective case studies, the expectation to plan, measure, and report intervention fidelity data has been growing since the early 1990s (Moncher & Prinz, 1991) and is identified as an essential component in speech-language pathology research and practice (Kaderavek & Justice, 2010). This is because the very act of planning, monitoring, and reporting intervention fidelity can improve intervention outcomes (Durlak & DuPre, 2008). We recommend that researchers be guided by the work of the Treatment Fidelity Workgroup of the National Institutes of Health Behavior Change Consortium (Bellg et al., 2004) as they aim to optimize and report intervention fidelity and, in doing so, optimize outcomes for children with phonological impairment.

Outcome and Experience Measures: Implications for Clinicians

The findings from this review raise challenges for clinicians wanting to identify outcome and experience measures suitable to collect during intervention. As no clear measure or set of measures dominated the empirical evidence base, answers to what, when, and how to collect outcome measures remain challenging for clinical practice. One option could be to use a common distal set of simple and efficient measures to capture intelligibility and the impact of impairment on everyday activity and participation (e.g., ASHA, 2020; Royal College of Speech and Language Therapists, 2019). However, sole reliance on such measures risks clinicians not knowing the immediate effect of intervention being implemented in clinical practice. By collecting proximal measures from treatment and phonological generalization data, clinicians have information to answer clinical questions about whether a child is responding to intervention cues and procedures, whether clinically significant and important change is occurring in a child's phonological system, and how long a specific intervention target should be treated (Baker & McLeod, 2004). As with our implications for researchers, we also recommend that clinicians combine tailored proximal measures of children's speech (and/or other abilities of interest) with common distal measures of the impact of

intervention on children's broader speech intelligibility, everyday activity, participation, and well-being using tools such as the FOCUS (Thomas-Stonell et al., 2010) and the TOMs (Enderby & John, 2019).

Finally, although the findings from this study emphasize a need for researchers to report measures of the intervention experience, there is value for clinicians in proactively gathering PREMs. Such measures would be particularly valuable when coaching parents to provide intervention, as unless they are explicitly asked, parents can be reluctant to question SLPs about the purpose of therapy or an activity despite their view or understanding being at odds with the SLP (Roulstone, 2015). Children with phonological impairment can also hold negative views on the intervention experience that can impact their motivation to participate (Crowe et al., 2017). By gathering and acting on PREMs data, not only could their experience of intervention be improved, but outcomes could also be optimized.

Limitations

The focus of this systematic search and review was on classifying and describing outcome and experience measures from peer-reviewed published articles of intervention for children with phonological impairment using the SORT (Baker & Masso, 2021). Although we documented the proportion reporting of intervention fidelity, we did not examine the types and measures of fidelity, strategies to optimize fidelity, or the rigor of the tools used to measure fidelity across the evidence base. This could be a valuable line of inquiry for future research. Second, the search strategy was limited to peer-reviewed published interventions in English or translated into English and so missed the body of intervention research and ways of measuring intervention outcomes and experiences published in other languages. Third, gray literature (i.e., documents such as reports, dissertations, and policy documents not controlled by a commercial publisher; Paez, 2017) was not included as we focused on peer-reviewed published intervention studies. Hence, a hand search of gray literature may be valuable to further our understanding of outcome and experience measures.

Conclusions

The measurement of intervention outcomes and experience is challenging yet important. Across 220 research articles on intervention for phonological impairment in children, one of the greatest challenges is the ability to compare outcomes across interventions. Although measures of phonological generalization are integral to evaluating the effect of intervention, there is a need for greater clarity about measures and the tools used to collect

those measures. Greater consideration is also needed to measure potential changes of the impact of phonological impairment on children's lives and the lives of others and the experience of receiving and providing intervention. By objectively measuring impact, we will be in a better position to detect and know the cascading outcomes of intervention. By objectively measuring experience, we will be in a better position to know and understand how children and their families experience intervention and how that experience might be optimized. Finally, by reporting fidelity of implementation, clinicians and researchers will have greater confidence when interpreting outcome and experience measures. Through increased transparency of reporting using a taxonomy such as the SORT, we will be better placed to understand how we might optimize outcomes for children with phonological impairment.

As this review has indicated, the adoption of one outcome measure that uses the same tool for all children with phonological impairment is unlikely to provide insight into every child's response to intervention or answer every research question. Rather, by using a suite of measures comprising proximal measures tailored to research and individual children's needs coupled with universal distal measures of intelligibility, the impact of phonological impairment on children's lives, and the experience of receiving and providing intervention, researchers and clinicians could work together progress insights and innovations in science and practice.

Author Contributions

Elise Baker: Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Investigation (Equal), Methodology (Equal), Project administration (Supporting), Validation (Equal), Visualization (Lead), Writing – original draft (Lead), Writing – review & editing (Lead). **Sarah Masso:** Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Investigation (Equal), Methodology (Equal), Project administration (Equal), Supervision (Lead), Validation (Equal), Visualization (Equal), Writing – review & editing (Equal). **Kylie Huynh:** Data curation (Equal), Investigation (Equal), Methodology (Equal), Visualization (Supporting), Writing – review & editing (Supporting). **Ellie Sugden:** Data curation (Equal), Formal analysis (Supporting), Investigation (Supporting), Methodology (Supporting), Project administration (Supporting), Validation (Equal), Writing – review & editing (Supporting).

Acknowledgments

Elise Baker would like to acknowledge funding from Western Sydney University's Research Start-up Award.

Sarah Masso would like to acknowledge funding from the Australian Research Council, Discovery Early Career Research Award (DE200101078).

References

- Ali, M., Ben Basat, A. L., Berthier, M., Blom Johansson, M., Breitenstein, C., Cadilhac, D. A., Constantinidou, F., Cruice, M., Davila, G., Gandolfi, M., Gil, M., Grima, R., Godecke, E., Jesus, L., Jiminez, L. M., Kambanaros, M., Kukkonen, T., Laska, A., Mavis, I., . . . Brady, M. C. (2021). Protocol for the development of the international population registry for aphasia after stroke (I-PRAISE). *Aphasiology*, 1–21. <https://doi.org/10.1080/02687038.2021.1914813>
- Allori, A. C., Kelley, T., Meara, J. G., Albert, A., Bonanahaya, K., Chapman, K., Cunningham, M., Daskalogiannakis, J., de Gier, H., Heggie, A. A., Hernandez, C., Jackson, O., Jones, Y., Kangesu, L., Koudstaal, M. J., Kuchhal, R., Lohmander, A., Long, R. E., Jr., Magee, L., . . . Wong, K. W. (2017). A standard set of outcome measures for the comprehensive appraisal of cleft care. *The Cleft Palate Craniofacial Journal*, 54(5), 540–554. <https://doi.org/10.1597/15-292>
- American Speech-Language-Hearing Association. (2019). *Pediatric communication* [Patient-reported outcome measure]. <https://www.asha.org/noms/patient-reported-outcome-forms/>
- American Speech-Language-Hearing Association. (2020). *ASHA NOMS SLP Healthcare Registry 2020 functional communication measures (FCMs)*.
- American Speech-Language-Hearing Association. (2021). *ASHA's National Outcomes Measurement System (NOMS) clinician user guide SLP healthcare*.
- Baker, E., & Masso, S. (2021). *Speech Outcome Reporting Taxonomy (SORT)*.
- Baker, E., McCauley, R. J., Williams, A. L., & McLeod, S. (2020). Elements in phonological intervention: A comparison of three approaches using the Phonological Intervention Taxonomy. In E. Babatsouli & M. J. Ball (Eds.), *On under-reported monolingual child phonology* (pp. 375–399). Multilingual Matters. <https://doi.org/10.21832/9781788928953-019>
- Baker, E., & McLeod, S. (2004). Evidence-based management of phonological impairment in children. *Child Language Teaching and Therapy*, 20(3), 261–285. <https://doi.org/10.1191/0265659004ct275oa>
- Baker, E., & McLeod, S. (2011). Evidence-based practice for children with speech sound disorders: Part 1. Narrative review. *Language, Speech, and Hearing Services in Schools*, 42(2), 102–139. [https://doi.org/10.1044/0161-1461\(2010/09-0075\)](https://doi.org/10.1044/0161-1461(2010/09-0075))
- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., & Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*, 23(5), 443–451. <https://doi.org/10.1037/0278-6133.23.5.443>
- Broomfield, J., & Dodd, B. (2011). Is speech and language therapy effective for children with primary speech and language impairment? Report of a randomized control trial. *International Journal of Language & Communication Disorders*, 46(6), 628–640. <https://doi.org/10.1111/j.1460-6984.2011.00039.x>
- Brousseau-Lapr e, F., Schumaker, J., & Kluender, K. R. (2020). Perception of medial consonants by children with and without speech and language disorders: A preliminary study. *American*

- Journal of Speech-Language Pathology*, 29(2), 883–889. https://doi.org/10.1044/2020_ajslp-19-00062
- Cohen, M. L., & Hula, W. D.** (2020). Patient-reported outcomes and evidence-based practice in speech-language pathology. *American Journal of Speech-Language Pathology*, 29(1), 357–370. https://doi.org/10.1044/2019_AJSLP-19-00076
- Crowe, K., Cumming, T., McCormack, J., Baker, E., McLeod, S., Wren, Y., Roulstone, S., & Masso, S.** (2017). Educators' perspectives on facilitating computer-assisted speech intervention in early childhood settings. *Child Language Teaching and Therapy*, 33(3), 267–285. <https://doi.org/10.1177/0265659017717437>
- Culatta, B., Setzer, L. A., & Horn, D.** (2005). Meaning-based intervention for a child with speech and language disorders. *Topics in Language Disorders*, 25(4), 388–401. <https://doi.org/10.1097/00011363-200510000-00013>
- Cunningham, B. J., & Oram Cardy, J.** (2020). Using implementation science to engage stakeholders and improve outcome measurement in a preschool speech-language service system. *Speech, Language and Hearing*, 23(1), 17–24. <https://doi.org/10.1080/2050571X.2019.1711307>
- Cunningham, B. J., Washington, K. N., Binns, A., Rolfe, K., Robertson, B., & 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(2), 447–464. https://doi.org/10.1044/2016_JSLHR-L-15-0329
- Durlak, J., & DuPre, E.** (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Enderby, P., & John, A.** (2015). *Therapy Outcome Measures for Rehabilitation Professionals—Third Edition*. J&R Press.
- Enderby, P., & John, A.** (2019). *Therapy Outcome Measures user guide*. J&R Press.
- Farquharson, K., Tambyraja, S. R., & Justice, L. M.** (2020). Contributions to gain in speech sound production accuracy for children with speech sound disorders: Exploring child and therapy factors. *Language, Speech, and Hearing Services in Schools*, 51(2), 457–468. https://doi.org/10.1044/2019_LSHSS-19-00079
- Fudala, J.** (2000). *Arizona Articulation Proficiency Scale—Third Edition*. Western Psychological Services.
- Gierut, J. A., & Morrisette, M. L.** (2015). Dense neighborhoods and mechanisms of learning: Evidence from children with phonological delay. *Journal of Child Language*, 42(5), 1036–1072. <https://doi.org/10.1017/S0305000914000701>
- Goldman, R., & Fristoe, M.** (2000). *Goldman-Fristoe Test of Articulation—Second Edition*. AGS. <https://doi.org/10.1037/t15098-000>
- Grant, M. J., & Booth, A.** (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Hinckley, J. J., & Douglas, N. F.** (2013). Treatment fidelity: Its importance and reported frequency in aphasia treatment studies. *American Journal of Speech-Language Pathology*, 22(2), S279–S284. [https://doi.org/10.1044/1058-0360\(2012/12-0092\)](https://doi.org/10.1044/1058-0360(2012/12-0092))
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A. W., & Michie, S.** (2014). Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. *BMJ*, 348, g1687. <https://doi.org/10.1136/bmj.g1687>
- Kaderavek, J. N., & Justice, L. M.** (2010). Fidelity: An essential component of evidence-based practice in speech-language pathology. *American Journal of Speech-Language Pathology*, 19(4), 369–379. [https://doi.org/10.1044/1058-0360\(2010/09-0097\)](https://doi.org/10.1044/1058-0360(2010/09-0097))
- Kampstra, N. A., Zipfel, N., van der Nat, P. B., Westert, G. P., van der Wees, P. J., & Groenewoud, A. S.** (2018). Health outcomes measurement and organizational readiness support quality improvement: A systematic review. *BMC Health Services Research*, 18(1), 1005. <https://doi.org/10.1186/s12913-018-3828-9>
- Kearney, E., Granata, F., Yunusova, Y., van Lieshout, P., Hayden, D., & Namasivayam, A.** (2015). Outcome measures in developmental speech sound disorders with a motor basis. *Current Developmental Disorders Reports*, 2(3), 253–272. <https://doi.org/10.1007/s40474-015-0058-2>
- Kingsley, C., & Patel, S.** (2017). Patient-reported outcome measures and patient-reported experience measures. *BJA Education*, 17(4), 137–144. <https://doi.org/10.1093/bjaed/mkw060>
- Lee, S. A. S.** (2018). The treatment efficacy of multiple opposition phonological approach via telepractice for two children with severe phonological disorders in rural areas of West Texas in the USA. *Child Language Teaching and Therapy*, 34(1), 63–78. <https://doi.org/10.1177/0265659018755527>
- MacLeod, A. A., & Glaspey, A. M.** (2014). A multidimensional view of gradient change in velar acquisition in three-year-olds receiving phonological treatment. *Clinical Linguistics & Phonetics*, 28(9), 664–681. <https://doi.org/10.3109/02699206.2013.878855>
- McCormack, J., McLeod, S., Harrison, L. J., & McAllister, L.** (2010). The impact of speech impairment in early childhood: Investigating parents' and speech-language pathologists' perspectives using the ICF-CY. *Journal of Communication Disorders*, 43(5), 378–396. <https://doi.org/10.1016/j.jcomdis.2010.04.009>
- McHugh, M. L.** (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22, 276–282. <https://doi.org/10.11613/bm.2012.031>
- McLeod, S., & Baker, E.** (2017). *Children's speech: An evidence-based approach to assessment and intervention*. Pearson Education.
- McLeod, S., Harrison, L. J., & McCormack, J.** (2012). *Intelligibility in Context Scale*. Charles Sturt University. <http://www.csu.edu.au/research/multilingual-speech/ics>
- McNeill, M., Noyek, S., Engeda, E., & Fayed, N.** (2021). Assessing the engagement of children and families in selecting patient-reported outcomes (PROs) and developing their measures: A systematic review. *Quality of Life Research*, 30(4), 983–995. <https://doi.org/10.1007/s11136-020-02690-4>
- Microsoft Corporation.** (2021). *Microsoft Excel Microsoft 365*. <https://office.microsoft.com/excel>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group.** (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), Article e1000097. <https://doi.org/10.1371/journal.pmed1000097>
- Moncher, F. J., & Prinz, R. J.** (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, 11(3), 247–266. [https://doi.org/10.1016/0272-7358\(91\)90103-2](https://doi.org/10.1016/0272-7358(91)90103-2)
- Morgan, L., Overton, S., Bates, S., Titterton, J., & Wren, Y.** (2021). Making the case for the collection of a minimal dataset for children with speech sound disorder. *International Journal of Language & Communication Disorders*, 56(5), 1097–1107. <https://doi.org/https://doi.org/10.1111/1460-6984.12649>
- Mullen, R., & Schooling, T.** (2010). The National Outcomes Measurement System for pediatric speech-language pathology. *Language, Speech, and Hearing Services in Schools*, 41(1), 44–60. [https://doi.org/10.1044/0161-1461\(2009/08-0051\)](https://doi.org/10.1044/0161-1461(2009/08-0051))

- Nelson, E. C., Dixon-Woods, M., Batalden, P. B., Homa, K., Van Citters, A. D., Morgan, T. S., Eftimovska, E., Fisher, E. S., Ovreteit, J., Harrison, W., Lind, C., & Lindblad, S. (2016). Patient focused registries can improve health, care, and science. *BMJ*, *354*, i3319. <https://doi.org/10.1136/bmj.i3319>
- Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, *10*(3), 233–240. <https://doi.org/10.1111/jebm.12266>
- Payne, S. (2002). Standardised tests: An appropriate way to measure the outcome of paediatric occupational therapy? *British Journal of Occupational Therapy*, *65*(3), 117–122. <https://doi.org/10.1177/030802260206500303>
- Perry, A., Morris, M., Unsworth, C., Duckett, S., Skeat, J., Dodd, K., Taylor, N., & Reilly, K. (2004). Therapy outcome measures for allied health practitioners in Australia: The AusTOMs. *International Journal for Quality in Health Care*, *16*(4), 285–291. <https://doi.org/10.1093/intqhc/mzh059>
- PROMIS Health Organization. (2016). *PROMIS Pediatric Scale v1.0—Global Health 7*. <https://www.healthmeasures.net/explore-measurement-systems/promis>
- RELEASE Collaboration. (2020). Communicating simply, but not too simply: Reporting of participants and speech and language interventions for aphasia after stroke. *International Journal of Speech-Language Pathology*, *22*(3), 302–312. <https://doi.org/10.1080/17549507.2020.1762000>
- Roulstone, S. (2015). Exploring the relationship between client perspectives, clinical expertise and research evidence. *International Journal of Speech-Language Pathology*, *17*(3), 211–221. <https://doi.org/10.3109/17549507.2015.1016112>
- Royal College of Speech and Language Therapists. (2019) *RCSLT Online Outcome Tool (ROOT)*. Different Class Solutions. <http://www.rcslt-root.org>
- Rvachew, S., & Brosseau-Lapr e, F. (2015). A randomized trial of 12-week interventions for the treatment of developmental phonological disorder in francophone children. *American Journal of Speech-Language Pathology*, *24*(4), 637–658. https://doi.org/10.1044/2015_ajslp-14-0056
- Rvachew, S., & Nowak, M. (2001). The effect of target-selection strategy on phonological learning. *Journal of Speech, Language, and Hearing Research*, *44*(3), 610–623. [https://doi.org/10.1044/1092-4388\(2001\)050](https://doi.org/10.1044/1092-4388(2001)050)
- Rvachew, S., Nowak, M., & Cloutier, G. A. (2004). Effect of phonemic perception training on the speech production and phonological awareness skills of children with expressive phonological delay. *American Journal of Speech-Language Pathology*, *13*(3), 250–263. [https://doi.org/10.1044/1058-0360\(2004\)026](https://doi.org/10.1044/1058-0360(2004)026)
- Sandbank, M., Chow, J., Bottema-Beutel, K., & Woynaroski, T. (2021). Evaluating evidence-based practice in light of the boundedness and proximity of outcomes: Capturing the scope of change. *Autism Research*, *14*(8), 1536–1542. <https://doi.org/10.1002/aur.2527>
- Shriberg, L. D., Austin, D., Lewis, B. A., McSweeney, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, *40*(4), 708–722. <https://doi.org/10.1044/jslhr.4004.708>
- Shriberg, L. D., Kwiatkowski, J., & Snyder, T. (1990). Tabletop versus microcomputer-assisted speech management. *Journal of Speech and Hearing Disorders*, *55*(4), 635–655. <https://doi.org/10.1044/jshd.5504.635>
- Smit, A. B., Brumbaugh, K. M., Weltsch, B., & Hilgers, M. (2018). Treatment of phonological disorder: A feasibility study with focus on outcome measures. *American Journal of Speech-Language Pathology*, *27*(2), 536–552. https://doi.org/10.1044/2017_AJSLP-16-0225
- Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *International Journal of Language & Communication Disorders*, *53*(4), 718–734. <https://doi.org/10.1111/1460-6984.12399>
- The EndNote Team. (2018). *EndNote Version EndNote X8*. Clarivate.
- Thomas-Stonell, N. L., Oddson, B., Robertson, B., & Rosenbaum, P. (2009). Predicted and observed outcomes in preschool children following speech and language treatment: Parent and clinician perspectives. *Journal of Communication Disorders*, *42*(1), 29–42. <https://doi.org/10.1016/j.jcomdis.2008.08.002>
- Thomas-Stonell, N. L., Oddson, B., Robertson, B., & 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*(1), 47–53. <https://doi.org/10.1111/j.1469-8749.2009.03410.x>
- Tyler, A. A., Edwards, M. L., & Saxman, J. H. (1987). Clinical application of two phonologically based treatment procedures. *Journal of Speech and Hearing Disorders*, *52*(4), 393–409. <https://doi.org/10.1044/jshd.5204.393>
- Tyler, A. A., Lewis, K. E., Haskill, A., & Tolbert, L. C. (2002). Efficacy and cross-domain effects of a morphosyntax and a phonology intervention. *Language, Speech, and Hearing Services in Schools*, *33*(1), 52–66. [https://doi.org/10.1044/0161-1461\(2002\)005](https://doi.org/10.1044/0161-1461(2002)005)
- Tyler, A. A., Williams, M. J., & Lewis, K. E. (2006). Error consistency and the evaluation of treatment outcomes. *Clinical Linguistics and Phonetics*, *20*(6), 411–422. <https://doi.org/10.1080/02699200500097769>
- Veritas Health Innovation. (n.d.). *Covidence systematic review software*. <https://www.covidence.org/>
- Williams, A. L., McLeod, S., & McCauley, R. J. (2021). *Interventions for speech sound disorders in children* (2nd ed.). Brookes.
- World Health Organization. (1998). *Health promotion glossary*.
- World Health Organization. (2007). *International Classification of Functioning, Disability and Health: Children and Youth Version (ICF-CY)*.
- Wray, J., Hobden, S., Knibbs, S., & Oldham, G. (2018). Hearing the voices of children and young people to develop and test a patient-reported experience measure in a specialist paediatric setting. *Archives of Disease in Childhood*, *103*(3), 272–279. <https://doi.org/10.1136/archdischild-2017-313032>

Appendix

Speech Outcome Reporting Taxonomy

