Designs and decisions: The creation of informal measures for assessing speech production in children

Nicole Limbrick, Jane McCormack & Sharynne McLeod

Designs and decisions: The creation of informal measures for assessing speech production in children

NICOLE LIMBRICK[1], JANE McCORMACK[1] & SHARYNNE McLEOD[2]

[1]*Charles Sturt University, Albury, Australia, and* [2]*Charles Sturt University, Bathurst, Australia*

Abstract
Speech-language pathologists (SLPs) frequently assess children's speech to diagnose and identify areas of difficulty, then determine appropriate intervention goals. Formal measures are available for assessment; however, many SLPs use informal measures within clinical practice. The purpose of this two-part mixed methods study was to describe informal measures created to assess children's speech. Study 1 involved a systematic review of 39 informal measures identified via journal database and internet searches, scanning of reference lists, and submission by SLPs and researchers. The measures were reviewed in terms of their conceptualization (content and format) and operationalization (evaluation and validation). Common conceptual features included assessment of consonant singletons, single words, computer format, and picture-naming. Few measures provided information addressing operational criteria; in particular, they lacked evaluation of their effectiveness. Study 2 involved an inductive thematic analysis of journal entries from eight creators of informal measures that explored key considerations in the development process. Informal measures were created due to the absence of measures which were sufficiently comprehensive and culturally appropriate, plus a desire to incorporate technology. Considerations in the creation of informal measures included sourcing research and existing measures to inform the measures' development, maximizing children's engagement, and utility. SLPs must be cautious when using informal measures due to their lack of operationalization. However, these measures often address SLPs' needs and so operationalization of informal measures would be beneficial for the profession.

Keywords: *Assessment, speech sound disorder, articulation, phonology, speech-language pathologist.*

Introduction

Speech impairment (also known as speech sound disorder) is common in childhood (Mullen & Schooling, 2010), affecting between 2.3–24.6% of children (Law, Boyle, Harris, Harkness, & Nye, 2000). It refers to problems with speech sound production, perception, and/or phonological representation which may make speech difficult to understand. Ongoing speech impairment may have long-term effects on children's literacy, emotional health, and ability to build relationships, cope with stress, and manage their behaviour (Felsenfeld, Broen, & McGue, 1994; Leitão & Fletcher, 2004; McCormack, McLeod, Harrison, & McAllister, 2010; Teverovsky, Bickel, & Feldman, 2009). Intervention is effective in treating speech impairment (Almost & Rosenbaum, 1998; Law, Garrett, & Nye, 2003) but for intervention to be most appropriate, speech-language pathologists (SLPs) need to conduct thorough and accurate assessment (Baker & Bernhardt, 2004; Stow & Dodd, 2003).

Purposes of assessment

An understanding of the reasons why SLPs assess children's speech is an essential precursor to considering which measures are used and how they are selected (Leitão, 2011). SLPs may assess children's speech production for a number of purposes, including screening, diagnosing, selecting intervention targets, monitoring progress, and determining when to discharge. The two main reasons are screening and diagnosing (Bankson, Bernthal, & Flipsen, 2009).

Screening aims to distinguish between children who have typically-developing speech, and children who may have speech impairment and require more comprehensive assessment to make a diagnosis (Bankson et al., 2009). Currently, there are no universally accepted protocols to guide SLPs when screening (Nelson, Nygren, Walker, & Panoscha, 2006). Often, commercially-available screening measures lack validity and rely on SLPs making subjective judgements about children's abilities (Sturner et al., 1994). Diagnostic assessment involves more

comprehensive sampling of children's speech. The results of diagnostic assessments indicate the presence and severity of speech impairment and are used to determine the need for SLP intervention. They also provide information about skills to be targeted and possible strategies to be implemented in therapy (Bankson et al., 2009).

Types of measures: Formal vs informal

There are many different measures available for assessing speech, and it is important that SLPs choose measures that are accurate, fair, and trustworthy because they "serve as gateways to services" (Crais, 2011, p. 342). There are consequences associated with using measures which have poor *sensitivity* (accurately identifying the presence of speech impairment in children who have impairment) and *specificity* (accurately identifying the absence of speech impairment in children who do not have impairment), including the over- or under-diagnosis of speech impairment among children. This may result in "inappropriate provision or denial of clinical services" (Friberg, 2010, p. 86), meaning some children may miss out on necessary intervention at a young age, when they have the best chance of positive outcomes (Nelson et al., 2006).

Of the measures available for assessing children's speech, some are published or formal measures, such as the Goldman-Fristoe Test of Articulation-2 (Goldman & Fristoe, 2000) and the Bankson-Bernthal Test of Phonology (Bankson & Bernthal, 1990), which are commercially-available and have undergone standardization. These formal measures are usually norm-referenced and enable children's scores to be compared with data from typically-developing, same-aged children to aid SLPs' decision-making surrounding the presence of speech impairment. Alternatively, some measures are classed as informal measures, meaning that they are not commercially published and have not undergone a formal standardization process. Informal measures are usually developed specifically for the intended population and do not have their own normative data, meaning the results are compared with functional standards (AERA, APA, & NCME, 1985; Bankson et al., 2009).

The use of informal measures in SLP practice

While SLPs use many of the available formal measures for assessing children's speech (Joffe & Pring, 2008; McLeod & Baker, 2004, 2012; Skahan, Watson, & Lof, 2007), they also frequently use informal measures. Indeed, the majority of a sample of 231 Australian SLPs (59.1%) report using informal measures when evaluating the speech of English-speaking children (McLeod & Baker, 2012), while 36% of a sample of 85 Dutch SLPs use a "self-made test" to assess Dutch children's speech production (Priester, Post, & Goorhuis-Brouwer, 2009, p. 1101). When assessing the speech of multi-lingual children, most (76.7%) of a sample of 128 Australian SLPs report always using informal measures (Williams & McLeod, 2012), which is mirrored by 67% of a sample of 309 SLPs in the US (Skahan et al., 2007).

Factors influencing the selection of measures

It is unclear why SLPs create and use informal measures instead of, or in conjunction with, formal measures. The reasons for creating informal measures have not previously been studied. However, SLPs' assessment practices and factors impacting on their selection of measures have been investigated in prior research (Joffe & Pring, 2008; McLeod & Baker, 2012; Priester et al., 2009; Skahan et al., 2007; Tyler & Tolbert, 2002). A number of factors have been reported to contribute to decision-making surrounding assessment, therefore it is possible to speculate about why SLPs may choose to create or use an informal measure.

One factor is comprehensiveness. Some formal measures do not assess specific aspects of speech production in depth. For instance, a review of procedures used to assess toddlers' polysyllable productions revealed that "mainstream picture-naming tests" do not include many polysyllabic words and SLPs need to look beyond those measures for a comprehensive sample (Baker & Munro, 2011, p. 61). SLPs may therefore decide to create informal measures to supplement data obtained from formal measures (Newman & Creaghead, 1985).

A second factor is children's enjoyment or engagement in the assessment process. Newman and Creaghead (1985) described the "interest value" and "creativity" (p. 70) of stimuli (e.g., incorporating current events and popular characters) as an advantage of self-made measures. Therefore, SLPs may create informal measures so that they are more appealing for children than the formal measures available.

Other factors include population, purpose, and financial constraints. Informal measures may be used when formal measures are inappropriate (e.g., for cultural reasons) or not available (Jordaan, 2008; Stow & Dodd, 2003; Vetter, 1988). Informal measures may be more appropriate to use when planning therapy as they enable SLPs to probe children's strengths and weaknesses (Vetter, 1988). Financial restrictions may also impact upon the accessibility of formal measures (Joffe & Pring, 2008).

Conceptualization and operationalization of measures

There are two stages involved in the development of assessment measures: conceptualization and operationalization (Frytak, 2000; McLeod, 2012).

Conceptualization of SLP measures. Conceptualization refers to the purpose and scope of measures, including the way they are designed and the features they include (McLeod, 2012). The conceptualization of measures may impact estimates of the severity of expression of a child's speech sound disorder. For instance, the apparent severity of speech impairment may vary depending on the type of words included in the word list (James, 1997; James, van Doorn, & McLeod, 2008) and the method of eliciting target sounds or skills (Morrison & Shriberg, 1992; Wolk & Meisler, 1998).

Operationalization of SLP measures. Operationalization refers to the evaluation, validation, or psychometric qualities of measures and can also impact on diagnostic accuracy (McLeod, 2012). SLPs can more confidently rely on the accuracy of measures if they demonstrate high levels of *validity* and *reliability*; that is, the measures assess what they are claimed to assess and provide consistent and stable results over time (Crais, 2011; McCauley & Swisher, 1984).

Evaluation of measures: *Conceptualization and operationalization*. Previous research into the development and review of assessment measures has identified criteria to be used in the evaluation of conceptualization and operationalization, known as conceptual and operational criteria (Friberg, 2010; McCauley, 1996; McCauley & Swisher, 1984; McLeod, 2012; Vetter, 1988). Examples of conceptual criteria include mode of elicitation, word shapes included, and inclusion of pictures. Examples of operational criteria include an explanation of test administration procedures, a description of tester qualifications, and evidence of content validity.

Reviews of the conceptualization and operationalization of formal speech and language measures have been undertaken using these criteria. McCauley and Swisher (1984) reviewed the operationalization of five formal measures for assessing articulation and 25 formal measures for assessing language and vocabulary. Four of the five measures of articulation adequately described the administration procedures, but criteria relating to validity and reliability of the measures were poorly addressed. When all 30 measures were considered, they found that five of the operational criteria were met by five or fewer measures. Predictive validity and inter-rater reliability criteria were not addressed by any of the 30 measures. Eisenberg and Hitchcock (2010) reviewed the conceptualization of 11 formal measures for assessing articulation and phonology. They reported that none of the measures sampled consonants and vowels in sufficient depth to determine a phonetic inventory. Sturner et al. (1994) reviewed 51 formal speech-language screening measures in terms of conceptualization and operationalization. Only nine of the measures met the conceptual criteria for "brevity and comprehensiveness", while just six of the measures provided operational data which could be used to determine their validity (Sturner et al., 1994, p. 25). Nelson et al. (2006) reviewed 24 speech and language screening measures in terms of their conceptualization and operationalization. They found that the measures varied widely in terms of their scope, as well as their ranges of sensitivity (17–100%) and specificity (45–100%).

Similar reviews of formal measures have been undertaken in other domains, including non-verbal oral and speech motor performance (McCauley & Strand, 2008), speech intelligibility (Kent, Miolo, & Bloedel, 1994), and language (Friberg, 2010; Spaulding, Swartwout Szulga, & Figueroa, 2012). These reviewers also found that formal measures of other speech and language skills varied in terms of the degree to which they met conceptual and operational criteria.

Each of these comprehensive reviews has primarily evaluated formal measures. While "less rigor can realistically be expected for measures that can be characterized as informal", reliability and validity are still important (McCauley, 1996, p. 128). Guidelines for designing informal measures have been outlined which address essential conceptual and operational qualities (McCauley, 1996; Vetter, 1988). However, the conceptualization and operationalization of informal speech measures have not previously been investigated.

Research aims

Although there is an abundance of formal measures available, many SLPs use informal measures to assess children's speech. Reasons behind the creation and use of informal measures are unclear. Furthermore, the conceptualization and operationalization of informal measures have never been reviewed. Thus, the purpose, scope and validation of informal measures used by SLPs to assess children's speech are currently unknown.

Two studies are reported in the current paper, both of which explore the creation of informal measures used to assess speech production in children. Study 1 adopted a quantitative methodology in the form of a systematic review of informal measures and statistical analysis to investigate the following research question: What are the conceptual and operational characteristics of informal measures used to assess children's speech production? Study 2 employed a qualitative methodology involving inductive thematic analysis of written journal entries/narratives to explore the following research questions: Why are informal measures created? and What do SLPs and researchers think are important considerations when creating informal measures?

Method

Study 1: Systematic review of informal speech measures

A systematic review of informal measures for assessing speech was conducted. To be included in the systematic review, informal measures needed to be designed to assess speech production by sampling a wide range of target sounds or skills. Measures located on the internet, within SLP workplaces, or provided as appendices to journal articles were eligible for inclusion. Exclusionary criteria for the systematic review can be found in the Appendix.

Identification of informal measures. Informal measures were identified for the systematic review in three main ways: (1) journal database searches, (2) internet searches, and (3) submission by SLPs and researchers.

STEP 1: JOURNAL DATABASE SEARCHES

Online database searches were conducted using EBSCO Host (Health) and Medline. These databases were selected as those most likely accessed by SLPs and researchers. The full list of databases and search terms can be found in Table I. Truncation of search terms was used to eliminate the possibility of articles being overlooked due to morphological differences in terms. Subject fields were searched, and the searches were limited to articles published between January 1980 and February 2012.

The initial search was conducted in EBSCO Host (Health) and resulted in 4536 papers. The titles and abstracts of the first 200 articles were scanned for relevance and full articles were read when relevant. The majority of the articles were deemed irrelevant based on an inappropriate purpose (e.g., intervention, analysis; $n = 55$), excluded condition (e.g., language disorder, brain injury; $n = 49$), or population (e.g., Spanish, Mandarin, multilingual; $n = 45$). Subject words and keywords to be excluded from the search were identified and the search terms were modified accordingly. The search was limited to full text with reference lists.

A refined search limited to full text articles with reference lists returned 996 papers. Papers were excluded if their aims were inconsistent with the purpose of the current study (e.g., intervention or analysis focus rather than assessment), the primary focus was not on speech production (e.g., speech perception), incomplete word lists were provided for measures, the words were from a naturalistic sample or not predetermined by researchers, the words were from a published test, or the word lists used to create the measure were unavailable. Nine measures were identified as appropriate for the systematic review.

An identical search was conducted in the Medline database. The search was limited to English language, full text, and articles published between January 1980 and February 2012. The search result returned 311 results. Titles, abstracts, and articles were scanned for relevance, but no new informal measures were identified.

Following the database searches, reference lists of relevant journal articles were scanned for further articles that may contain informal measures. Potentially relevant articles were sourced and their abstracts were scanned for relevance. Seven journal articles containing informal measures were added.

Table I. Search terms for database and internet searches.

Databases searched	Terms included	Terms excluded
EBSCO Host (Health)	articulation	multilingual*, bilingual*, therapy, intervention,
• Academic Search Complete	phonolog*	treatment*, (English "as" a second language),
• Dentistry & Oral Sciences Source	speech	(Language* other than English), cultural
• ERIC	consonant*	pluralism, narrative*, cogniti*, grammar,
• Health Source: Nursing/Academic Edition	syllab*	dyslexi*, reading disability, reading disorder,
• Health Source - Consumer Edition	vowel*	Down* syndrome, Cerebral palsy, autis*,
• SocINDEX with Full Text	assess*	cochlear implant, stutter*, fluency, voice,
• Health Business Elite	test*	hearing impair*, hearing disorder, hearing
• Psychology and Behavioral Sciences Collection	evaluat*	loss, hearing aid*, Deaf*, traumatic brain
• CINAHL Plus with Full Text	measur*	injur*, acquired brain injur*, neurological
Medline[a]	screen*	injur*, brain injur*, aphasia, stroke,
	child*	Parkinson* disease, epilepsy, dysphagia,
	paediatric*	Multiple Sclerosis, literacy, writing, spelling,
	pediatric*	expository text, problem-solving, executive
		function, working memory, activities of daily
		living, intelligence, attachment, self-care
Google web search engine	Articulation tests for children,	No terms excluded
Yahoo! search engine	Articulation tests,	
	Informal articulation tests,	
	Articulation screener,	
	Phonology test	

[a]No terms excluded in Medline search.

STEP 2: INTERNET SEARCHES

Secondly, searches were conducted using Google and Yahoo! search engines to reflect those commonly used in SLP workplaces, potentially with a purpose of locating existing measures. Five strings of search terms were entered into both Google and Yahoo! search engines to locate informal measures (Table I). Each of the search terms returned between 1–6 million hits. The first 10 pages of each search were scanned for relevant measures. The titles and descriptions of the search results were read and the links were accessed when relevant. The linked site or document was then scanned to determine its suitability for inclusion. Eleven informal measures were identified using this method.

STEP 3: SUBMISSION BY SLPs AND RESEARCHERS

Informal measures were also identified via submission by SLPs and researchers. Firstly, a message was posted on the phonological therapy list (Bowen, 2001) inviting members to submit any informal measures they have used or created to the researchers. The list has 8058 international members and is a forum for SLPs to discuss the assessment and management of childhood speech impairment. The message contained links to an information sheet and a consent form. SLPs emailed or posted the completed consent forms and their informal measures to the research team.

Secondly, an invitation was emailed to members of the authors' professional networks, requesting SLPs and researchers submit informal measures they had used or created. The information sheet and consent form were provided as attachments. The procedure was the same as for the phonological therapy list members. A snowball method of recruitment was utilized, whereby SLPs and researchers were asked to forward the invitation to their own professional networks.

These approaches to identification of informal measures aimed to capture SLPs and researchers with a known interest in speech impairment, in order to increase the response rate. They enabled an international sample of informal measures to be obtained, and provided an insight into informal measures that had not been disseminated on the internet. Twelve informal measures were identified via submission by SLPs and researchers. A total of 39 informal speech measures were included in the systematic review using these methods of identification. A reference list for the 39 informal measures is located in Supplementary Appendix A to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552.

Procedure. The 39 informal measures were systematically reviewed in terms of their conceptualization and operationalization (Frytak, 2000; McCauley & Swisher, 1984; McLeod, 2012; Vetter, 1988). A full list of criteria used in the systematic review can be found in Supplementary Appendix B online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552. There were 10 categories relating to *demographic criteria*, including geographic location and intended population. There were eight categories relating to *conceptualization* of the measures: format, purpose, target skill, scope, presentation, elicitation, scoring, and analysis (Supplementary Appendix B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552). There were six categories relating to *operationalization* of the measures: formulation of a clinical question, selection of stimulus items, identification of desirable responses, formulation of instructions, development of decision-making guidelines, and evaluation of the measure (Supplementary Appendix B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552). Operational and conceptual criteria were informed by previous research in test development and evaluation (Eisenberg & Hitchcock, 2010; Friberg, 2010; McCauley, 1996; McCauley & Strand, 2008; McCauley & Swisher, 1984; McLeod, 2012; Vetter, 1988). Information for each measure was obtained from a number of sources including manuals, instruction sheets, test stimuli, scoresheets, and attached journal articles, to address the criteria.

Data relating to demographics and conceptualization of the measures were coded numerically, typically using binary coding (e.g., Does the word list include bisyllabic words? $0 = $ no, $1 = $ yes). Data relating to operationalization were also coded numerically according to a three-way code: $0 = $ no relevant information provided about the characteristic being examined; $1 = $ provides some information but does not meet the criterion; and $2 = $ sufficient detail provided to meet the criterion (McCauley & Strand, 2008).

Analysis. Data were entered into the Statistical Package for Social Sciences (PASW Statistics, 2009) and analysed using descriptive statistics. Frequency of occurrence was calculated to determine the number of measures meeting each criterion. This enabled identification of the most common characteristics of the informal measures.

Reliability and validity. Development of criteria for the systematic review and protocol for coding of informal measures was accomplished by consensus between the first and third author to enhance reliability of the study. Next, coding of the systematic review was undertaken by the first author. Following the systematic review, four (10%) measures were randomly selected and independently coded by the third author. Point-to-point agreement between the two authors was 88.1% for 664 data points. This was considered to be an acceptable level of

inter-rater reliability. The first author re-coded four (10%) of the measures within 1 month of completing the initial evaluations to determine intra-rater reliability. Point-to-point agreement between the coding was 94.4% for 664 data points, which was considered an acceptable level of intra-rater reliability.

Study 2: Journal entries

The process of creating informal measures was explored through analysis of the clinical reasoning and decision-making processes of SLPs and researchers, as reported in journal entries.

Participants. To be eligible for study 2, participants needed to be SLPs or researchers who had created their own informal measure for assessing children's speech production. Eight participants (referred to hereinafter as *creators*) volunteered for study 2; four were practising SLPs and four were researchers. Five of the creators resided in Australia and three resided in the US.

Procedure. The creators were recruited during study 1. The information sheet disseminated in study 1 also contained details for study 2 and the consent form enabled individuals to identify their willingness to participate in study 2.

Data were collected in a written form via journal entries (narratives/recounts), to allow creators time to reflect on and describe the process of creating their informal measure. Creators were provided with directions to guide their journal entries. They were asked to describe the clinical problem or incident which prompted them to create their own informal measure for assessing children's speech. They were encouraged to describe their experience of creating their informal measure, as well as their clinical reasoning throughout the development process. Creators emailed or posted their completed journal entries and consent forms to the research team.

Analysis. Journal entries were read and analysed via inductive thematic analysis. Segments of text were selected and coded, then compared with a diagram of the steps in the development of an informal measure by Vetter (1988). This diagram "articulates steps in decision-making that can benefit the development of any informal measure" and describes a procedure for creating an informal measure in the absence of an appropriate formal measure (McCauley, 1996, p. 129). Statements from each journal entry were compiled beneath the relevant steps of Vetter's (1988) diagram. Key content was identified from the excerpts and compared between all journal entries to search for common themes. New codes were added whenever new themes arose from the data and the established codes did not represent the meaning of a text segment (Creswell, 2009).

Scientific rigour. The trustworthiness of the analysis was facilitated through the use of data immersion and constant comparative analysis. Data immersion involved repeated readings of the journal entries to gain a deeper understanding of the meanings being expressed (Davidson & McAllister, 2002; Yin, 2011). Constant comparative analysis involved checking each journal entry with others as the analysis progressed to ensure coding was consistent. Following coding, all journal entries were re-read to ensure no sections of text had lost their intended meaning during analysis. Journal entries were read and analysed by both the first and second authors to enhance reliability of coding.

Results

Study 1: Systematic review of informal speech measures

The 39 informal measures were reviewed according to criteria applying to informal measures. Informal measures are not expected to demonstrate the same rigour as formal measures (McCauley, 1996). Therefore, only results from criteria relating to informal measures have been reported in this paper. Supplementary Appendix B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552 lists all criteria used in this study.

Demographic information. The 39 informal measures were created between the years 1985–2012 (see Table II). The majority of the measures were created

Table II. Demographic criteria and results.

Demographic criteria[b]	n	%[a]
Location		
US	18	46.2
Australia	12	30.8
UK	5	12.8
Dialect of English		
American English	20	51.3
Australian English	9	23.1
British English	5	12.8
New Zealand English	3	7.7
Other dialects	1	2.6
Context		
University	22	56.4
School	4	10.3
Clinic	2	5.1
Hospital	1	2.6
Creator		
Researcher	28	71.8
Speech-Language Pathologist	4	10.3
Administrator		
Speech-Language Pathologist	9	23.1
Teacher	4	10.3
Parent	2	5.1
Psychologist	1	2.6

[a]% correct to one decimal place.
[b]demographic data was unavailable for some measures

in the US ($n = 18$, 46.2%) and Australia ($n = 12$, 30.8%). All of the measures assessed English speakers since measures intended for multilingual children were excluded, and the most frequently assessed English dialect was American English ($n = 20$, 51.3%). Twenty-two measures (56.4%) specified an age range of the children to be tested, with ages ranging between 1 year (1;0) and 14 years 9 months (14;9). Most of the informal measures were developed by researchers ($n = 28$, 71.8%) in universities ($n = 22$, 56.4%). Eight of the measures (20.5%) had been adapted from an existing measure and they were most commonly designed to be administered by an SLP ($n = 9$, 23.1%).

Conceptualization of the informal measures. Table III provides an overview of the conceptual criteria that were used and the number of measures that met each criterion.

FORMAT

Almost all of the informal measures utilized technology ($n = 38$, 97.4%), such as computer formatted pages ($n = 12$, 30.8%) and Microsoft PowerPoint™ slideshows ($n = 7$, 17.9%), which contained animations, sounds, and transition effects ($n = 2$, 5.1%). No mobile device applications or eBooks met the exclusionary criteria. More than half of the measures provided instructions ($n = 31$, 79.5%), scoresheets ($n = 22$, 56.4%), and word lists ($n = 20$, 51.3%), and many provided picture stimuli ($n = 16$, 41.0%).

PURPOSE AND TARGET SKILL

The informal measures were created for the purposes of screening ($n = 26$, 66.7%), probing ($n = 12$, 30.8%), intervention planning ($n = 7$, 17.9%), measuring progress ($n = 7$, 17.9%), and diagnosis ($n = 4$, 10.3%). The exclusionary criteria meant all of the measures primarily assessed speech sound production; however, other additional speech-related skills, including intelligibility ($n = 5$, 12.8%), were also examined.

SCOPE OF THE MEASURES

The measures most often assessed consonant singletons ($n = 22$, 56.4%), while consonant clusters ($n = 17$, 43.6%), vowels ($n = 8$, 20.5%), phonological processes ($n = 7$, 17.9%), and polysyllabic words ($n = 5$, 12.8%) were also examined. Some measures focused on just one of these categories of sounds or skills ($n = 19$, 48.7%), while others assessed sounds from more than one of these categories ($n = 18$, 46.2%). The number of items in the target word list varied widely, ranging between 16–310 words. Some of the measures assessed more than one production of each sound or skill across all word positions ($n = 11$, 28.2%). The word lists of

Table III. Conceptual criteria and results.

Conceptual criteria	n	%[a]
Format		
Computer formatted pages (e.g., Microsoft Word)	12	30.8
Microsoft PowerPoint slideshow	7	17.9
Provided as an appendix to a journal article	20	51.3
Instructions provided	31	79.5
Score sheet provided	22	56.4
Word list provided	20	51.3
Picture stimuli provided	16	41.0
Analysis form provided	7	17.9
Purpose		
Screening	26	66.7
Probing	12	30.8
Intervention planning	7	17.9
Measuring intervention progress	7	17.9
Diagnosing	4	10.3
Target Skills		
Speech production	39	100
Intelligibility	5	12.8
Stimulability	4	10.3
Inconsistency	1	2.6
Scope		
Consonant singletons	22	56.4
Consonant clusters	17	43.6
Vowels	8	20.5
Monosyllabic words included	37	94.9
Bisyllabic words included	35	89.7
Polysyllabic words included	31	79.5
Nouns	39	100
Other word types (e.g., adjectives)	31	79.5
Verbs	30	76.9
Initial position	16	41.0
Within word position	14	35.9
Final position	16	41.0
Presentation		
Pictures incorporated	29	74.4
Illustrations (e.g., Microsoft ClipArt™)	14	35.9
Photos	7	17.9
Uniform style	6	15.4
Picture scenes (multiple words elicited from one picture)	4	10.3
Elicitation		
Single words	30	76.9
Connected speech	11	28.2
Real words	39	100
Picture-naming	26	66.7
Imitation	9	23.1
Cueing hierarchy	23	59.0
Scoring and Analysis		
Transcription of children's responses	24	61.5
Independent analysis	12	30.8
Relational analysis	26	66.7

[a]% correct to one decimal place.

the informal measures most commonly consisted of monosyllabic word shapes ($n = 37$, 94.9%) and nouns ($n = 39$, 100%). The majority of measures assessed target sounds or skills in word-initial position ($n = 16$, 41.0%) and word-final position ($n = 16$, 41.0%). Target sounds needed to be specified on the score sheet for word positions to be coded.

PRESENTATION

Pictures were incorporated in the majority ($n = 29$, 74.4%) of the measures. Most pictures were

illustrations ($n = 14$, 35.9%) and were presented in colour ($n = 12$, 30.8%). Picture scenes were used to elicit multiple words in four (10.3%) of the measures.

ELICITATION

Speech production was most commonly assessed at the single word level ($n = 30$, 76.9%), while some informal measures assessed connected speech ($n = 11$, 28.2%). All measures assessed target sounds or skills in real words ($n = 39$, 100%), and some also incorporated nonsense words ($n = 4$, 10.3%). Picture-naming was the most common method of eliciting target words ($n = 26$, 66.7%). A cueing hierarchy for eliciting target words was outlined for over half of the measures ($n = 23$, 59.0%).

SCORING AND ANALYSIS

Scorers were most frequently required to transcribe children's productions ($n = 24$, 61.5%), primarily by transcribing whole words ($n = 20$, 51.3%). Some of the measures provided information regarding analysis of children's speech productions ($n = 14$, 35.9%). Both independent ($n = 12$, 30.8%) and relational ($n = 26$, 66.7%) analyses were described. The most common type of independent analysis described was a stimulability inventory ($n = 4$, 10.3%), while the most common type of relational analysis was a phonological process analysis ($n = 13$, 33.3%). Two of the measures (5.1%) provided normative data from another source to aid analysis.

Operationalization of informal measures. The informal measures addressed the operational criteria (McCauley, 1996; Vetter, 1988) outlined for informal measures to varying degrees. None of the measures met all of the operational criteria (see Table IV).

FORMULATION OF A CLINICAL QUESTION

Most of the measures ($n = 28$, 71.8%) provided some description of their purpose or specified a clinical question to be answered concerning the client (e.g., probing the production of a specific consonant).

SELECTION OF STIMULUS ITEMS

Overall, some evidence was provided for stimulus content or breadth, relevance, and difficulty for the measures; however, none of the measures fully addressed all three criteria for stimulus selection. The majority of the measures ($n = 31$, 79.5%) comprised stimuli which examined target skills with some breadth. Eight of the measures (20.5%) met the criterion regarding stimulus breadth and content. Most of the measures ($n = 35$, 89.7%) contained stimulus items which had some relevance to the purpose of the measure or the clinical question. Four measures (10.3%) met the criterion based on stimulus relevance. Almost all of the measures ($n = 36$, 92.3%) contained stimuli which were judged to be of appropriate difficulty for a child. Only three measures (7.7%) had references suggesting that the stimuli were at a suitable level of difficulty for a child, thus meeting this criterion.

IDENTIFICATION OF DESIRABLE RESPONSES

While all of the measures provided some information regarding the identification and reliable scoring of desirable responses, none met both criteria. Almost all of the measures ($n = 38$, 97.4%) provided some definition or identification of target words or responses which could reasonably be executed by a child. One of the measures (2.6%) met this identification criterion. Most of the measures ($n = 33$, 84.6%) had also provided some form of scoring information and specified target words on the

Table IV. Operational criteria and results.

Operational criteria	Met criterion		Provided some information	
	n	%[a]	n	%[a]
Formulation of a clinical question				
Identification of a specific clinical question/purpose	0	0	28	71.8
Selection of stimuli				
Stimulus breadth and content	8	20.5	31	79.5
Stimulus relevance	4	10.3	35	89.7
Stimulus difficulty	3	7.7	36	92.3
Identification of desirable responses				
Desirable responses identified that can be reasonably executed by children	1	2.6	38	97.4
Desirable responses identified that can be reliably scored	2	5.1	33	84.6
Formulation of instructions				
Instructions outlined that are likely to be understood by the client	5	12.8	28	71.8
Decision-making guidelines devised				
Decision-making guidelines developed to help determine correct vs incorrect responses on test stimuli	2	5.1	20	51.3
Evaluation of the effectiveness of measures				
The measure has been evaluated and steps in the design process have been revised to enhance effectiveness	0	0	3	7.7

[a]% correct to one decimal place.

scoresheet, suggesting that desirable responses may be scored somewhat reliably. Only two of the measures (5.1%) provided enough information to address this scoring criterion.

FORMULATION OF INSTRUCTIONS

The majority of the measures ($n = 28$, 71.8%) provided some instructions which were likely to be understood by the test user to guide their administration and scoring, and prompts for the child were targeted at an appropriate level. Five measures (12.8%) provided a sufficient amount of detail in their instructions to meet this criterion.

DECISION-MAKING GUIDELINES DEVISED

Administration, scoring, and analysis of the measures, provision of normative data, and information regarding what constituted a correct vs incorrect response were considered for this criterion. Around half of the measures ($n = 20$, 51.3%) provided some of this information to aid decision-making surrounding whether a child should be identified with having a speech impairment or as typically-developing. Two of the measures (5.1%) included decision-making (performance) guidelines which were sufficient to meet the criterion. One of these measures (2.5%) included normative data generated from their own test.

EVALUATION OF THE EFFECTIVENESS OF THE MEASURE

There was little evidence of evaluation of the measures' effectiveness or subsequent revision processes. Three (7.7%) of the measures provided some evidence or information regarding evaluation or validation. One example was pilot testing of a word list, conducted by a researcher. None of the measures fully addressed the criterion relating to evaluation/validation.

Study 2: Journal entries

Eight journal entries were submitted for study 2 by creators of informal measures. The journal entries ranged between one-to-four pages in length.

Reasons for creation. Creators described a number of different catalysts for developing their informal measures. Themes related to: a *need* to develop a measure for a specific purpose or population vs a *want* for particular design features; and designing the measure for a *research purpose* vs a *clinical purpose*.

NEED VS WANT

Some creators *needed* to develop a measure due to the absence of any appropriate measures to suit their purpose or population. A lack of formal measures which were culturally appropriate prompted some of the creators to make their own to suit their intended population:

> ... The Goldman-Fristoe assessment (Goldman & Fristoe, 2000) ... was a culturally inappropriate assessment to be using as it was made in America and had American items in it, such as a gun, an American flag ... some of which many Australian children did not know (Creator 3).

A lack of formal measures which sufficiently sampled specific target sounds or skills also led to the decision to create a new measure. The limited sampling of consonant clusters and polysyllabic words in formal measures was an issue:

> Commercially-available tests ... provide only a snapshot of what children can produce ... these abbreviated tasks do not adequately sample clusters (Creator 5).

The inability to sample the target sounds of interest across all word positions using formal measures was also a reason for creators to design their own measure:

> ... some SLPs had been previously using screeners which only examined sounds in the initial position (Creator 6).

Despite the availability of other measures, some creators *wanted* to develop a measure with certain desirable qualities. For instance, some of the creators described existing measures as lengthy, time-consuming, and complex to administer and score. Thus, creating a measure that was "simple" (Creator 7) and "wouldn't take too long to administer" (Creator 3) was important. Creators also wanted to develop measures which were "fun" or "interesting" for children (Creator 4). Desiring a particular format for the measure was also a catalyst for designing a self-made measure. Creators "wanted something that was not paper-based" (Creator 4) and could be "easily carried" (Creator 6) to places as required.

RESEARCH PURPOSE VS CLINICAL PURPOSE

Creators included the purpose of their self-made measure in the reasons for their development. Some creators described a need to develop a measure for *research purposes*:

> I needed to devise a speech and language test battery to use in my doctoral research (Creator 2).

Others had a *clinical purpose* in mind when deciding to create their own measure. Financial restrictions imposed by their workplace were a contributing factor:

> Our budget was extremely tight and so purchasing a standardized measure was essentially out of the question—we had to conserve our funds to purchase full assessments, office supplies, and therapy materials (Creator 6).

The desire to use "the same measure" consistently by all SLPs to minimize "confusion" and regulate access to SLP services was also a catalyst for developing an informal measure (Creator 6).

Creating the measure: Important considerations. In describing factors they consider when developing informal measures, the creators' responses were found to relate to three key stages in the process: preparation, development, and use. *Preparation* encompassed the process undertaken by creators once they had decided to make an informal measure, which involved devising a rationale and ideas for the design. *Development* referred to making the measure and considering what stimulus items it would consist of. *Use* related to decisions surrounding administration and how the targets would be elicited from the children.

PREPARATION

Creators described sourcing journal articles when designing their measures, and a review of the literature was apparent within some of the journal entries. Research conducted by the creator or others was used to inform and guide the development of the measures:

> ... a journal article (James, 2002) ... indicated that articulation assessments needed to include not only nouns and single syllable words, but adjectives, verbs, and multisyllabic words and, where possible, a phoneme needs to be targeted more than once in the same word position (Creator 3).

Research was also used to support the creators' own ideas and decisions to make a measure:

> Studies have suggested that three-element clusters (/skr-/, /spr-/, /str-/, /spl-/, /skw-/) should only be taught if the child has the second and third consonants in his/her phonemic inventory (Gierut & Champion, 1999) ... Therefore, SLPs needed a probe that facilitates efficient determination of the phonemic status of these six sounds (Creator 5).

Creators also made reference to other existing measures when planning the design of their own measure. Existing measures were used to inform the development of a new measure or were used as a starting point from which modification occurred (the creators did not mention whether or not permission had been gained from the authors and publishers to adapt the measures):

> I took the Metaphon Screening Test (Dean et al., 1990) stimuli and ... modified it over a period of years (Creator 2).

DEVELOPMENT

Creators described the selection of stimulus items as a key consideration during the actual development of their measure, particularly the words in the word list. The difficulty of the words was an important factor, with creators considering appropriateness, familiarity, and presence within the vocabulary of children:

> The words selected were mainly common household and community items from topics familiar to children such as clothing, food, animals, etc. (Creator 7).

Ensuring the words sampled a range of different sounds, word shapes, word types, and word positions was also considered:

> I assembled an independent probe that samples singletons at least five times across word positions and initial clusters at least twice. This allowed me to more fully describe a child's sound repertoire (Creator 5).

To ensure the target words sampled the sounds of interest to the creator, some decided which sounds to include first, then identified suitable words containing those sounds:

> First I selected which consonant clusters to include. I decided to elicit word-initial and word-final consonant clusters that were produced in Australian English ... Next I worked out which words I should use to elicit each consonant cluster (Creator 1).

The selection of the words was closely related to the selection of pictures. "Imageability" (Creator 6) was a priority for some, and the words selected were sometimes the result of what pictures were available:

> Sometimes, the photographs we used were mediated by what was available. There were a few consonant clusters that were not included since there were no picturable words that would readily be known by young children (Creator 1).

Other creators sought appropriate pictures to suit their specific needs. One creator consulted other professionals, and other creators sourced them from image databases:

> I employed a photographer to take photos ... we talked about good pictures to take for different words. I also consulted a primary school teacher (Creator 1).
> Once the words were chosen, I found photographs on Art Explosion© Photo Objects 150,000 (Nova Development, 2006; Creator 5).

The choice of format of the measures was also a key consideration during the development process. Some creators reported the use of a computer was necessary, with child enjoyment, interaction, and engagement at the forefront of their decision-making regarding creation of their measure:

> Computer-based delivery ... meant that I could include some animations and sound-files to enhance the test experience for the child (Creator 4).

Use

When deciding how the measure would be used, the elicitation of target skills from the children was considered, with creators weighing up single word vs connected speech samples and spontaneous vs imitated responses:

> It was necessary for the child to be able to look at the picture and name it without a model, if at all possible. I wanted this format so that we could examine the child's true production, not an imitation (Creator 6).

Cueing hierarchies were also devised by some creators:

> I aimed to elicit spontaneous single word responses. However, I also generated a hierarchy to assist children if they did not know the words (Creator 1).

One creator developed two versions of her informal measure to provide options for administration, in an effort to suit the needs of SLPs and individual children:

> The long and short versions of the (informal measure) give SLPs options for administration ... depending upon the child's profile (Creator 5).

Discussion

This two-part research project explored informal measures for assessing speech production in children. Study 1 involved a systematic review of 39 informal measures in terms of demographic, conceptual, and operational criteria. Study 2 involved inductive thematic analysis of written journal entries describing the creation of informal measures.

Demographic, conceptual, and operational characteristics of informal measures

The informal measures were most commonly created in universities, which is not surprising given that the majority of the informal measures were created by researchers. Few of the measures were created by SLPs working clinically, which contrasts with the high proportion of practising SLPs using informal measures indicated by previous research (McLeod & Baker, 2012; Priester et al., 2009; Skahan et al., 2007; Williams & McLeod, 2012). This may reflect time limitations and the caseload size of practising SLPs. However, it may also indicate that some practising SLPs use informal measures which have been created by SLPs other than themselves, rather than a self-made measure.

The main purpose of the informal measures was for screening. This may indicate that SLPs are aware of the need to use measures demonstrating the operational qualities of reliability and validity when making a diagnosis, which may be perceived as less important during screening. However, some of the informal measures had a diagnostic purpose. None of these diagnostic measures provided normative data to aid decision-making when making a diagnosis, which is potentially problematic.

The use of technology was an innovative conceptual feature of many informal measures. Some of the measures utilized a Microsoft PowerPoint™ slideshow format with animations, sounds, and transition effects. It is possible that the computer format of these informal measures was selected in order to be more appealing for children. However, given the broad range of applications available for Apple™ and Android™ devices, it was surprising that none of the informal measures reviewed were mobile device applications. In future, the profession may see some of these informal measures converted into forms which are more easily downloaded onto mobile devices.

Some of the informal measures were designed to probe target skills in speech production that have not yet been the focus of formal measures. This may be an attempt to compensate for the inadequacy of formal measures for in-depth sampling of specific sounds and skills, such as productions of polysyllabic words (Baker & Munro, 2011; James, 2006), vowels (Eisenberg & Hitchcock, 2010; Pollock, 1991), and consonant clusters (Powell, 1995).

Pictures were incorporated in the majority of the informal measures, which is parallelled by the stimuli of formal measures. Scoring of the measures often involved transcription of whole words, which is similar to the scoring of formal measures. The analyses described for the informal measures included both independent and relational methods which are also applied with formal measures.

Few of the informal measures fully addressed the operational criteria relating to the selection of stimuli, formulation of instructions, identification of desired responses, and formulation of decision-making guidelines. In terms of stimulus selection, measures mostly lacked research evidence to support the decisions made. It is possible that the creators of the measures had consulted the literature, but not reported it within the documentation. Clear instructions and knowledge of desirable responses and decision-making guidelines are pertinent to the replicability, reliability, and objectivity of assessment findings (Vetter, 1988). It may be that the measures were designed for their creators' use only, and this could have impacted on the amount of information provided explicitly. Additionally, the informal measures lacked evidence that their effectiveness had been evaluated.

The lack of operationalization demonstrated by the majority of the measures is consistent with the expectation that informal measures do not demonstrate the same rigour as formal measures (McCauley, 1996). Previous research indicates that

even formal measures for assessing speech and language domains have a number of limitations in terms of their operationalization (Eisenberg & Hitchcock, 2010; Friberg, 2010; McCauley & Strand, 2008; McCauley & Swisher, 1984; Nelson et al., 2006; Spaulding et al., 2012). In view of the resources, time, expenses, and expertise in psychometrics (Streiner & Norman, 2003) involved in operationalization, this stage is usually only accessible to publishing companies. However, while many informal measures lack scientific rigour, positive features include being user-friendly for SLPs and appropriate for the intended population and context.

Overall, while preliminary comparisons indicate some differences in the conceptualization and operationalization of informal and formal measures, there are also features which are common to both. Since they share a number of similarities with formal measures, this raises the question of why SLPs are creating and using informal measures.

Reasons for creating informal measures

An understanding of the reasons why SLPs decide to create their own informal measures and their considerations during the development process was provided in Study 2. Creators identified that a catalyst for developing their own measure was that formal measures did not sample specific target skills of interest (e.g., consonant clusters, vowels, polysyllabic words) in sufficient depth. The inadequacy of formal measures in sampling these skills has been reported in prior research (Baker & Munro, 2011; Eisenberg & Hitchcock, 2010; James, 2006; Pollock, 1991; Powell, 1995). Costs associated with purchasing formal measures and time efficiency contributed to creators' decisions to develop their own measure. Financial restrictions (Joffe & Pring, 2008) and time (Tyler & Tolbert, 2002) have been reported in previous research as factors influencing SLPs' selection of measures.

Some creators desired particular conceptual qualities within a measure, such as a computer format or portability. Enhancing children's enjoyment and engagement in the assessment process was desired. The level of appeal for children, in terms of the appearance and creativity of pictures, has been described as a factor influencing SLPs' selection of measures (Khan, 2002) and an advantage of creating an informal measure (Newman & Creaghead, 1985).

Considerations when creating informal measures

Creators identified a number of considerations relating to preparation, development, and use of an informal measure. In terms of preparation, they used existing measures and their own or others' research to inform and support the design. The results of study 1 support this notion, in that the methods of elicitation, pictures, scoring, and analyses all resemble existing formal measures. The inclusion of a range of word types and shapes in many of the measures suggests some awareness of the research identifying the importance of these considerations (James, 1997; McLeod, Hand, Rosenthal, & Hayes, 1994). Some of the measures also included references, and measures provided as appendices to journal articles evidenced some form of literature review.

In terms of the development phase, creators felt it was important to choose items that were familiar and of an appropriate level of difficulty for children, which has been identified as an important consideration in the development of an informal measure (McLeod, 2012; Newman & Creaghead, 1985). While study 1 revealed that the measures largely consisted of stimulus items which were of appropriate difficulty, few provided supporting references and met the criteria for stimulus difficulty. Another consideration was inclusion of stimuli sampling a variety of sounds, word shapes, and word positions, which has also been outlined in the literature (Newman & Creaghead, 1985; Winitz, 1969). The scope of the measures in study 1 largely reflected this consideration; however, less than a third of the measures sampled productions of target sounds across all word positions. The scope of the measures in study 1 may have been impacted by the creators' purpose for developing them, and perhaps sampling across one or two word positions met their needs. Creators also reported that a computer format which would be engaging for children was an important aspect in developing their informal measure. Research supports the use of a computer format for intervention (Shriberg, Kwiatkowski, & Snyder, 1989, 1990; Wren & Roulstone, 2008); however, the use of a computer format has not yet been explored in relation to assessment.

Regarding the use of their informal measures, creators described decision-making surrounding single word vs connected speech sampling. The majority of the measures in study 1 incorporated single word sampling, despite the benefits of obtaining a connected speech sample reported in the literature (Shriberg & Kwiatkowski, 1985; Wolk & Meisler, 1998). However, creators may have weighed up the advantages and disadvantages of both sampling methods (Khan, 2002; McLeod et al., 1994; Morrison & Shriberg, 1992) and made their decision based on their needs and clinical judgement. The need for cueing hierarchies was described by some creators when developing their measure, and these were also found to be present in over half of the measures in study 1.

Clinical implications

Study 2 revealed that SLPs and researchers created their own measures because they were unable to locate formal measures to suit their needs. The

informal measures in study 1 showed the creativity of SLPs and researchers, with the measures showcasing some innovative design features and probing specific target sounds and skills (e.g., vowels, polysyllabic words). The creators should be commended for their innovation and contribution to the range of measures available. Informal measures may be the only measures which suit certain children and meet the specific needs of SLPs, so SLPs should not be discouraged from creating and using them. The advantages of informal measures, including their contextual appropriateness and utility, may make them more suitable for assessment of children's speech than formal measures. However, SLPs need to take appropriate caution when interpreting the results of informal measures.

Using measures for purposes for which they have not been designed may be problematic (Spaulding et al., 2012). Typically, the use of informal measures is more appropriate for the purposes of screening, intervention planning, and monitoring progress than for diagnosing speech impairment, because of their lack of operationalization. If measures which have not undergone operationalization are used to diagnose speech impairment, SLPs cannot be certain that their diagnoses are accurate. Inaccurate diagnoses can negatively impact children and their families. Labelling children with speech impairment when it is not appropriate may cause emotional harm and place unnecessary financial burden on families. Ethically, SLPs must ensure prioritization of services is fair; otherwise children may miss out on necessary intervention to the detriment of their literacy development, emotional health, and relationships (Nathan, Stackhouse, Goulandris & Snowling, 2004).

It may be beneficial to raise the awareness of SLPs regarding the importance of using measures which are reliable, valid, and appropriate for their intended purpose when assessing children's speech. SLPs need to consider the various conceptual and operational qualities of measures to decide which are most important for them to focus on when selecting a measure (Friberg, 2010). SLPs could be provided with a framework of conceptual and operational criteria to guide their thinking when developing their own measure. The criteria provided in Supplementary Appendix B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552 could be used as a guide. SLPs could also use the criteria to evaluate the informal measures that they use in practice. This might reassure SLPs that a diagnosis of speech impairment is accurate, rather than being influenced by other factors (e.g., type of words used).

Limitations

It is possible that more informal measures exist, but were not able to be accessed for this systematic review. Efforts were made to obtain measures from a variety of sources through extensive searching of databases and the internet, and submission by individuals. However, a selection bias in the identification of informal measures may have impacted the measures included in the systematic review. For instance, SLPs were invited to submit their informal measures via email, which may have excluded measures that were paper-based if SLPs were unable to scan and email them to the authors. This may have partially contributed to the large proportion of informal measures with a computer format. Additionally, the monetary limit of $20 and exclusion of word lists from books may have limited the scope of the informal measures included. It is also possible that researchers create their informal measures with others in mind and are more inclined to distribute their informal measures on the internet, thus contributing to the large number of measures created by researchers in the review. In contrast, practising SLPs may create informal measures solely for their own individual or workplace use and retain the measure within that workplace. Thus, the results may not fully represent the range of informal measures created and used.

Inviting members of the authors' professional networks to participate in the research may have introduced selection bias into the sample of creators. This, coupled with the small sample size in study 2, means that data is exploratory and the results may not represent all the different reasons for creating a measure and considerations involved in their development. Participant checking of the authors' interpretation of the journal entries was not undertaken, which may have impacted the accuracy of the results. However, both studies provide interesting preliminary findings which could form the basis of future research.

Future directions

This research provides an initial understanding of the purpose, scope and evaluation of informal measures, as well as the reasons for their creation and considerations in their development. Future research could investigate informal measures which were beyond the scope of this systematic review, such as those assessing speech perception or intelligibility. This would provide an understanding of the conceptualization and operationalization of a broader range of informal measures. Further research could also explore decision-making surrounding the creation and use of informal measures with a larger participant sample to determine whether the findings from this research represent the broader population.

Another avenue of research is the potential operationalization of informal measures by undertaking validity and reliability testing. Funding should be provided to further the development of

informal measures (McCauley & Strand, 2008). Publishing companies and developers of formal measures would benefit from acknowledging the ideas of SLPs and researchers and using them to inform the creation of formal measures in the future. Perhaps publishing companies could consult and work collaboratively with practising SLPs and researchers when developing and revising formal measures. This may lead to the development of formal measures which are more appealing for SLPs and the children they assess, and more appropriate for the demands of clinical practice.

Conclusion

Despite the availability of formal measures, many SLPs report using informal measures when assessing children's speech (McLeod & Baker, 2012; Priester et al., 2009; Skahan et al., 2007; Williams & McLeod, 2012). The present research investigated informal measures of speech production in terms of their conceptualization and operationalization. Some of the features of informal measures were innovative (e.g., use of technology) and others were similar to formal measures (e.g., inclusion of pictures). Reasons for creating informal measures were identified and considerations in their development were explored. As outlined by McCauley and Strand (2008, p. 82), "… taking stock of their [informal measures'] current status can serve as an important step toward promoting their judicious use and improving the quality of ongoing test development …". The findings of this study have implications for practising SLPs utilizing and creating informal measures; researchers and publishing companies involved in developing formal measures; and funding bodies, which have the potential to invest in the operationalization of informal measures. Further research exploring informal measures and their creation may help to guide future test development to ensure accurate and timely identification of speech impairment in a format that is useful and appealing for SLPs, as well as for the children they assess.

Acknowledgement

References

Almost, D., & Rosenbaum, P. (1998). Effectiveness of speech intervention for phonological disorders: A randomised controlled trial. *Developmental Medicine and Child Neurology*, 40, 319–325.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Baker, E., & Bernhardt, B. (2004). From hindsight to foresight: Working around barriers to success in phonological intervention. *Child Language Teaching and Therapy*, 20, 287–318.

Baker, E., & Munro, N. (2011). An overview of resources for assessing toddlers' productions of polysyllables. *ACQuiring Knowledge in Speech, Language and Hearing*, 13, 58–62.

Bankson, N. W., & Bernthal, J. E. (1990). *Bankson-Bernthal Test of Phonology*. San Antonio, TX: Special Press.

Bankson, N. W., Bernthal, J. E., & Flipsen, P. (2009). Phonological assessment procedures. In J. E. Bernthal, N. W. Bankson, & P. Flipsen (Eds.), *Articulation and phonological disorders: Speech sound disorders in children*. (pp. 187–250). Boston, MA: Pearson.

Bowen, C. (2001). *Phonological therapy: Children's speech sound disorders (PHONO-TX)*. http://health.groups.yahoo.com/group/phonologicaltherapy/.

Crais, E. R. (2011). Testing and beyond: Strategies and tools for evaluating and assessing infants and toddlers. *Language, Speech, and Hearing Services in Schools*, 42, 341–364.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). London: SAGE.

Davidson, B., & McAllister, L. (2002). An introduction to qualitative research approaches. *ACQuiring Knowledge in Speech, Language and Hearing*, 4, 28–31.

Dean, E., Howell, J., Hill, A., & Waters, D. (1990). *Metaphon Resource Pack*. Windsor, Berks: NFER-Nelson.

Eisenberg, S. L., & Hitchcock, E. R. (2010). Using standardized tests to inventory consonant and vowel production: A comparison of 11 tests of articulation and phonology. *Language, Speech, and Hearing Services in Schools*, 41, 488–503.

Felsenfeld, S., Broen, P. A., & McGue, M. (1994). A 28-year follow up of adults with a history of moderate phonological disorder: Educational and occupational results. *Journal of Speech and Hearing Research*, 37, 1341–1353.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy*, 26, 77–92.

Frytak, J. (2000). Measurement. *Journal of Rehabilitation Outcomes*, 4, 15–31.

Gierut, J., & Champion, A. H. (1999). Interacting error patterns and their resistance to treatment. *Clinical Linguistics and Phonetics*, 13, 421–431.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation-2*. Circle Pines, MN: American Guidance Service.

James, D. (1997). The need to use polysyllabic words in the assessment and analysis of speech. *Australian Communication Quarterly*, Autumn, 6–8.

James, D. (2002). Part II: On assessing normal speech development. *ACQuiring Knowledge in Speech, Language and Hearing*, 4, 148–150.

James, D. (2006). *Hippopotamus is so hard to say: Children's acquisition of polysyllabic words*. Unpublished PhD, University of Sydney, Sydney.

James, D., van Doorn, J., & McLeod, S. (2008). The contribution of polysyllabic words in clinical decision making about children's speech. *Clinical Linguistics and Phonetics*, 22, 345–353.

Joffe, V., & Pring, T. (2008). Children with phonological problems: A survey of clinical practice. *International Journal of Language and Communication Disorders*, 43, 154–164.

Jordaan, H. (2008). Clinical intervention for bilingual children: An international survey. *Folia Phoniatrica et Logopaedica*, *60*, 97–105.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology*, *3*, 81–95.

Khan, L. M. (2002). The sixth view: Assessing preschoolers' articulation and phonology from the trenches. *American Journal of Speech-Language Pathology*, *11*, 250–254.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (2000). Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature. *International Journal of Language and Communication Disorders*, *35*, 165–188.

Law, J., Garrett, Z., & Nye, C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *Cochrane Database of Systematic Reviews*, *3*, CD004110.

Leitão, S. (2011). My top 10 assessment resources (with a paediatric slant). *ACQuiring Knowledge in Speech, Language and Hearing*, *13*, 94–95.

Leitão, S., & Fletcher, J. (2004). Literacy outcomes for students with speech impairment: Long-term follow-up. *International Journal of Language and Communication Disorders*, *39*, 245–256.

McCauley, R. J. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Speech, and Hearing Services in Schools*, *27*, 122–131.

McCauley, R. J., & Strand, E. I. (2008). A review of standardised tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology*, *17*, 81–91.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, *49*, 34–42.

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*, 378–396.

McLeod, S. (2012). Translation to practice: Creating sampling tools to assess multilingual children's speech. In S. McLeod, & B. A. Goldstein (Eds.), *Multilingual aspects of speech sound disorders in children*. Clevedon, UK: Multilingual Matters.

McLeod, S., & Baker, E. (2004). Current clinical practice for children with speech impairment. In B. E. Murdoch, J. Goozee, B. M. Whelan, & K. Docking (Eds.), *Proceedings of the 26th World Congress of the International Association of Logopedics and Phoniatrics*. Brisbane: The University of Queensland.

McLeod, S., & Baker, E. (2012). Speech-language pathologists' practices regarding assessment, analysis, target selection and intervention for children with speech sound disorder. [Manuscript in preparation].

McLeod, S., Hand, L., Rosenthal, J. B., & Hayes, B. (1994). The effect of sampling condition on children's productions of consonant clusters. *Journal of Speech and Hearing Research*, *37*, 868–882.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research*, *35*, 259–273.

Mullen, R., & Schooling, T. (2010). The National Outcomes Measurement System for paediatric speech-language pathology. *Language, Speech, and Hearing Services in Schools*, *41*, 44–60.

Nathan, L., Stackhouse, J., Goulandris, N., & Snowling, M. J. (2004a). The development of early literacy skills among children with speech difficulties: A test of the "critical age hypothesis". *Journal of Speech, Language, and Hearing Research*, *47*, 377–391.

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the U.S. Preventive Services Task Force. *Pediatrics*, *117*, 298–319.

Newman, P. W., & Creaghead, N. A. (1985). Assessment of articulatory and phonological disorders. In N. A. Creaghead, P. W. Newman, & W. A. Secord (Eds.), *Assessment and remediation of articulation and phonological disorders*. (pp. 70–80). Columbus, OH: Merrill.

Nova Development. (2006). *Art Explosion Photo Objects 150,000*. Calabasas, CA: Nova Development Corporation.

PASW Statistics. (2009). *Statistical Program for the Social Sciences* (Version 17.0.2). Chicago, IL: SPSS Inc.

Pollock, K. E. (1991). The identification of vowel errors using traditional articulation or phonological process test stimuli. *Language, Speech, and Hearing Services in Schools*, *22*, 39–50.

Powell, T. (1995). A clinical screening procedure for assessing consonant cluster production. *American Journal of Speech-Language Pathology*, *4*, 59–65.

Priester, G. H., Post, W. J., & Goorhuis-Brouwer, S. M. (2009). Problems in speech sound production in young children. An inventory study of the opinions of speech therapists. *International Journal of Pediatric Otorhinolaryngology*, *73*, 1100–1104.

Shriberg, L. D., & Kwiatkowski, J. (1985). Continuous speech sampling for phonologic analyses of speech-language delayed children. *Journal of Speech and Hearing Disorders*, *50*, 323–334.

Shriberg, L. D., Kwiatkowski, J., & Snyder, T. (1989). Tabletop versus microcomputer-assisted speech management: Stabilization phase. *Journal of Speech and Hearing Disorders*, *54*, 233–248.

Shriberg, L. D., Kwiatkowski, J., & Snyder, T. (1990). Tabletop versus microcomputer-assisted speech management: Response evocation phase. *Journal of Speech and Hearing Disorders*, *55*, 635–655.

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology*, *16*, 246–259.

Spaulding, T. J., Swartwout Szulga, M., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools*, *43*, 176–190.

Stow, C., & Dodd, B. (2003). Providing an equitable service to bilingual children in the UK: A review. *International Journal of Language and Communication Disorders*, *38*, 351–377.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). New York: Oxford University Press.

Sturner, R. A., Layton, T. L., Evans, A. W., Heller, J. H., Funk, S. G., & Machon, M. W. (1994). Preschool speech and language screening: A review of currently available tests. *American Journal of Speech-Language Pathology*, *January*, 25–36.

Teverovsky, E. G., Bickel, J. O., & Feldman, H. M. (2009). Functional characteristics of children diagnosed with childhood apraxia of speech. *Disability and Rehabilitation*, *31*, 94–102.

Tyler, A. A., & Tolbert, L. C. (2002). Speech-language assessment in the clinical setting. *American Journal of Speech-Language Pathology*, *11*, 215–220.

Vetter, D. K. (1988). Designing informal assessment procedures. In D. E. Yoder, & R. D. Kent (Eds.), *Decision making in speech-language pathology*. (pp. 192–193). Philadelphia, PA: B. C. Decker.

Williams, C., & McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology*, *14*, 292–305.

Winitz, H. (1969). *Articulatory acquisition and behavior*. New York: Appleton-Century-Crofts.

Wolk, L., & Meisler, A. W. (1998). Phonological assessment: A systematic comparison of conversation and picture naming. *Journal of Communication Disorders*, *31*, 291–313.

Wren, Y., & Roulstone, S. (2008). A comparison between computer and tabletop delivery of phonology therapy. *International Journal of Speech-Language Pathology*, *10*, 346–363.

Yin, R. K. (2011). *Qualitative research from start to finish*. New York: Guilford Press.

Appendix: Exclusionary criteria for study 1.

- Measures published by commercial publishing houses (e.g., Pro-Ed) at the time of the systematic review.
- Measures with costs greater than $20 AUD.
- Measures created before 1980.
- Measures found within published books.
- Measures with a primary focus other than assessing speech impairment (e.g., specific language impairment)
- Measures with the primary purpose of describing characteristics of speech impairment (e.g., intelligibility, prosody, rate, stimulability) rather than the presence or severity of speech impairment.
- Measures evaluating speech impairment with a known cause (e.g., acquired dysarthria, cleft palate/velopharyngeal impairment).
- Measures designed for assessing children speaking languages other than English or for whom English is a second language.
- Measures involving collection of spontaneous conversation samples without pre-determined targets.
- Measures primarily consisting of non-words (e.g., non-word repetition tests).
- Word lists provided as appendices to journal articles where assessment was not the primary focus of the journal article (e.g., focus on intervention or analysis).
- Measures probing a small sub-set of specific phonemes or speech patterns (e.g., vowel tests probing unrounded vowels only; consonant cluster tests probing /s/ clusters only; polysyllabic word tests probing a small range of word shapes).
- Measures with incomplete word lists.
- Measures described in journal articles that were not accessible (e.g., could not be obtained by contacting the author).

Supplementary material available online

Supplementary Appendix A and B to be found online at http://informahealthcare.com/doi/abs/10.3109/17549507.2013.770552.