The (un)Met Need for Creation of Speech Sampling Tools

Although speech sampling tools are available in a range of languages (see McLeod, Chapter 13), speech-language pathologists (SLPs) around the world often rely on English or informal assessment tools when working with multilingual children (McLeod, 2007; Skahan et al., 2007). For example, when considering assessment of Samoan children, Ballard and Faro (2008: 379) wrote: ‘As information about different cultures and languages is limited, few practitioners have the multicultural assessment skills or resources necessary to make such a judgment or a culturally appropriate assessment’. When speech assessments are not available, sampling tools are often created by SLPs themselves. There are at least four situations that may necessitate the creation of speech sampling tools: when no assessment tools have been developed for a particular language (e.g. when working with children from a developing nation); when working with children who speak non-dominant languages (e.g. Stow & Pert, 1998); when assessment tools are available in a particular language, but are difficult to access (e.g. PAL (1995) can be purchased after attending a training workshop in Greece); or when available sampling tools are not appropriate for use in the SLPs’ context (e.g. the words are irrelevant for a particular dialect or the accompanying images are culturally inappropriate/insensitive).

Determine whether Suitable Measurement Tools already Exist
Before creating a new tool to assess children’s speech, it is important to conduct a thorough search to determine whether suitable tools already exist. There are a number of places to look:

1. Internet and library databases including: Google scholar, Medline, ERIC and CINAHL. Journal papers can include word lists (sampling tools) as appendices (e.g. a single word test of Samoan is included in the appendix of Ballard and Farao (2008); the Mansoura Arabic Articulation Test (MAAT) is included as the appendix in Abou-Elsaad et al. (2009); and a single word test for Cantonese tone production is included in Lee et al. (2010)).

2. Professional associations in the countries that speak the language(s) of interest. The International Directory of Communication Disorders (IDCD) (Bleile, 2006) provides a list of professional associations around the world. For example, the Panhellenic Association of Logopedists has published a phonology test in Greek (PAL, 1995) and the Norwegian Association of Speech Therapists has published a speech sound test in Norwegian (Vidsjå et al., 1983).

3. Hospitals, community health centers, schools and universities in countries that speak the language(s) of interest. Within universities, contact linguistic and phonetic departments, teacher education departments (e.g. in Korea, teachers learn the International Phonetic Alphabet), as well as speech-language pathology departments (and try an alternative nomenclature such as speech and language therapy, logopedics and phoniatriecs).

4. Craniofacial/cleft palate medical teams, such as Operation Smile (http://www.operationsmile.org/), Interplast (http://www.intreplast.org) and The Smile Train (http://www.thesmiletrain.org) often include SLPs who assess children’s speech.

5. Non-government organizations (NGOs), charities and mission organizations in countries that speak the language(s) of interest, as well as the office in the home country, to identify SLPs who no longer work in the country. The Summer Institute of Linguistics (www.sil.org) also develops translation tools, including phonetic fonts.

6. Books and chapters, including the International Guide to Speech Acquisition (McLeod, 2007), which lists available assessments in 12 English dialects and 23 languages other than English, and Chapter 13 of the current book, which contains a list of speech assessments in 18 different languages.

[H1]Creating a Sampling Tool

If a suitable tool is unavailable, then according to Frytak (2000) there are two steps to creating one: conceptualization and operationalization. Most SLPs who create sampling tools
undertake many of the components of the conceptualization step, but few undertake the second step to operationalize the tool.

[H2] Conceptualization: Decide on the purpose and scope of the tool

(1) **Purpose.** Will the tool assess one language and/or dialect, or be designed to compare skills in different languages and dialects? Alternatively, will the tool assess a child’s ability to learn a language (e.g. Jacobs & Coufal, 2001)? Will the tool be used for screening, diagnostic assessment, describing speech, goal setting for intervention or determination of outcomes? It is unlikely that one tool will be able to achieve each of these purposes.

(2) **Intended population.** Who is the tool designed for? Consider the country, language, dialect and age group(s). Is the test for monolingual or bilingual speakers? What qualifications do the examiners need?

(3) **Target skill.** Will the tool be used to assess production (articulation of speech sounds, phonology, nonword repetition), perception, phonological awareness (emergent literacy skills), stimulability or intelligibility/acceptability? The underlying theoretical construct of the sampling tool should guide this step.

(4) **Scope.** Will the tool assess all or some of the following: consonants, vowels, consonant clusters, onsets, codas, multisyllabic words, prosody, tones (for tonal languages such as Cantonese and Norwegian), phonological processes/patterns and/or phonological awareness?

[H3] Guidelines for developing a single word test of consonant production

One of the most common speech sampling tools used throughout the world is the single word test of consonants. This may be (and is typically) the first test that is developed for a language or dialect. The principles provided for developing a single word test of consonants can be applied to the development of other tools.

(1) Prepare an inventory of the consonants that occur within the language and dialect. There are a number of sources to assist with this step, including:

   (a) **Handbook of the International Phonetic Alphabet** (IPA, 1999), and Illustrations of the IPA in the **Journal of the International Phonetic Alphabet** (http://journals.cambridge.org/action/displayJournal?jid=IPA) (e.g. Bengali is illustrated in Khan (2010) and includes supplementary sound files).

   (b) Books such as **The Sounds of the World’s Languages** (Ladefoged & Maddieson, 1996) and **The International Guide to Speech Acquisition** (McLeod, 2007).
(c) Journal papers and book chapters (e.g. Salameh et al. (2003) include inventories of Arabic and Swedish consonants).
(d) Websites, including the Speech Accent Archive (http://accent.gmu.edu/browse_native.php), American Speech-Language-Hearing Association (ASHA) Multicultural Affairs and Resources page (http://www.asha.org/practice/multicultural/Phono.htm) and Wikipedia IPA charts (http://en.wikipedia.org/wiki/International_Phonetic_Alphabet); however, remember that not all websites have undergone rigorous peer review.

2) Prepare a phonotactic inventory (i.e. sound structure) of the language and dialect. Which sounds occur in all word positions, and which occur only in syllable-initial word-initial, syllable-final within-word, syllable-initial within-word and syllable-final word-final?

3) Decide on the number of productions of each speech sound to elicit (typically between one and five). Children’s speech varies and different phonetic contexts create different co-articulatory situations (e.g. in English, if the only word that is elicited is yellow /jɛlʊ/, then it is likely that /j/ will be misarticulated due to co-articulatory effects with /l/).

4) Decide on the mode, and the phonotactic and morphological context of the words.
   (a) Mode. Consonants in single words are typically more accurate than consonants in connected speech; however, different phonological processes occur in each mode (Morrison & Shriberg, 1992).
   (b) Number of syllables. Consonants in monosyllabic and bisyllabic words are typically more accurate than in polysyllabic words (James et al., 2008) (e.g. cat vs. caterpillar).
   (c) Consonant clusters. Consonants within consonant clusters are typically less accurate than consonants in singleton contexts (Smit et al., 1990) (e.g. bread vs. bed vs. red).
   (d) Morphophonemes. Consonants within morphophonemic contexts may be less accurate than consonants in words that do not contain a grammatical morpheme (Song et al., 2009) (e.g. keys /kɪz/ (plural) vs. buzz /bʌz/). Consequently, a morphophonological context should be avoided so that phonology and morphology are not conflated.

5) Locate frequently used words for the language and dialect. If possible, this list should comprise words frequently used by children. Some word frequency lists on the Internet are taken from compiling words printed in newspapers and will not provide as relevant information for compiling a pediatric assessment tool. The CHILDES database may act as a good reference. An alternative source is children’s dictionaries, first word books and early school books.
(6) Select words. The selected words should be:

(a) Familiar to children, and occur frequently, so that children can produce the word spontaneously, and not rely on imitated productions.

(b) Culturally sensitive in both word choice and pictures (e.g. the number 4 in Cantonese is considered to be unlucky because it sounds like the word for death).

(c) Picturable. Often this means that the test primarily contains nouns, unless computerized animation is available to demonstrate action verbs. Be aware that languages differ regarding the emphasis on nouns vs. verbs (Peña et al., 2003).

(7) Consider presentation. Words can be presented as objects, pictures, photographs, drawings or computerized images (possibly with animation and an audio example). Clarity and ease of identification is paramount in this step. If pictures are sourced from the Internet, they should be within the public domain (not copyrighted). Try not to mix different illustration styles (e.g. cartoon style drawings vs. line drawings vs. photographs). Color vs. black and white illustrations has been found to effect children’s picture-naming ability. For example, Barrow et al. (2000) found that when children were less familiar with words, color illustrations were more easily identifiable than black and white illustrations during a picture-naming task.

(8) Consider test administration. Will the tool be administered in a standardized (i.e. static) manner or as a dynamic assessment? Standardized test administration may involve presenting instructions via pre-recorded speech. Dynamic speech assessments typically involve systematic assessment of stimulability of sounds (Glaspey & Stoel-Gammon, 2007). Will the children be allowed to code-switch from one language (and/or dialect?) to another during the assessment (e.g. Grech & Dodd, 2008)? Will the tool be administered by a native language speaker? Will testing occur in a formal clinical setting or a more familiar setting where children are more likely to use their home dialect or an informal register (e.g. Ballard & Farao, 2008)?

(9) Consider transcription, scoring, recording and analysis. Document acceptable pronunciations by transcribing ‘accurate'/adult-like productions of the words in the chosen dialect(s). Decide whether scoring of the tool will include transcription of consonants only, and/or whole words, and whether broad vs. narrow phonetic transcription will be used (using the International Phonetic Alphabet). Will transcription occur using pen and paper or via a computer program? Will the speech sample be audio- and/or video-recorded for reliability checking and later comparison of outcomes after assessment?
Creating a measurement tool is a fine balance between the need to accurately represent the complexity of speech and a child’s capacity to speak, with the realities of children’s attention span, SLPs’ time and the need to identify and quantify a speech sound disorder.

[H2] Operationalization: Testing a test

Once a sampling tool is created, it can then be operationalized and validated. It is useful to work in conjunction with a (bio)statistician or psychometrician to establish the psychometric properties of the tool. Key concepts when operationalizing a sampling tool include determination of:

(1) Reliability: Does the test provide a consistent measure?
   (a) *Internal consistency*. To estimate the extent to which a group of items measure the same overall construct (relevant statistical analyses include: Cronbach’s alpha, confirmatory factor analysis). A test may generate one overall score, or a set of subscale scores, depending on the number of domains being assessed.
   (b) *Test-retest reliability*. To estimate the extent to which similar scores would be achieved if the test was readministered under the same circumstances (relevant statistical analyses include: correlation, intra-class correlation).
   (c) *Rater reliability including inter- and intra-rater reliability*. To estimate the extent to which the test results are the same when the test is administered by different people or by the same people at different times. Rater reliability can be affected by training and by provision of external standard comparison measures (relevant statistical analyses include: inter-class correlation, Cohen’s kappa).

(2) Validity: Does the test measure what it claims to measure?
   (a) *Content validity*. To consider ‘the degree to which the items in the measure cover the domain of interest’ (Frytak, 2000: 22). Two common approaches to establishing content validity of a test are: by systematic examination of literature and consultation with experts (Brown, 1985; Chan & Lee, 1999; Millman & Greene, 1993; Thorn & Deitz, 1989). A systematic literature review is important for specifying the initial test content, while expert professional judgments play an integral part in further defining the test areas and evaluating the degree to which the items are relevant and representative to the target construct (AERA, APA, NCME, 1999; Bubela et al., 1990; Haynes et al., 1995). Relevance of test content refers to the appropriateness of the test items for measuring the target construct (Messick, 1993). For example, testing the English phoneme /t/ in the medial position using the word *butter* would be an item with low relevance as the medial
/t/ is often realized as /d/ or a flap in some dialects of English. Representativeness refers to the extent that test items are proportional to the elements contained in the targeted construct (Nunnally & Bernstein, 1994). A speech sampling tool may have content validity if there is comprehensive coverage of the phonemes within a language with test items that are appropriately devised (relevant statistical analyses include: correlation).

(b) **Criterion validity.** To consider the degree of overlap between the test and other standard speech sampling tools that measure the same or similar abilities. Of central importance in criteria-related validity is the choice of criterion measure, which has to be a well-established valid measurement in itself (Gay, 1985; Messick, 1993) (relevant statistical analyses include: correlation).

(c) **Construct validity.** To demonstrate that the test measures the construct of interest from a number of perspectives and sources of evidence. The central theme is to examine how well the obtained test scores reflect the unobservable attribute. It is the most difficult and yet the most important form of validity to obtain. For test scores to be meaningful, it has to be shown that all the individual test items relate to others in some way so that they are measuring the same rather than diverse dimensions. Intercorrelations among test items, factor analyses and item response theory are often used to determine construct validity (AERA, APA, NCME, 1999; Benson & Clark, 1982; Bubela et al., 1990; Clark & Watson, 1995; Hambelton, 1993; Hasselkus & Safrit, 1976; Nunnally & Bernstein, 1994). Other commonly used approaches to evaluate the strength of the score interpretation include the known groups method, convergence and discrimination (relevant statistical analyses include: correlation, factor analysis, Rasch analysis).

(3) **Item analysis.** To determine item difficulty and item discrimination scores. James (2001) provided an example of undertaking an item analysis when developing a speech sampling tool and determined that polysyllabic words were better than monosyllabic words for discriminating between English-speaking children.

(4) **Sensitivity and specificity.** To describe the proportion of cases classified correctly. Measures with high sensitivity are able to accurately identify children with speech sound disorder (indicated as the percent identification of true positives) and measures with high specificity are able to accurately identify typically developing children (indicated as the percent identification of true negatives) (relevant statistical analyses include: linear discriminant function analysis).

(5) **Standardization.** Testing the sampling tool on a large normative sample (who represent the population of children for whom the test is intended) results in a wide range of raw
scores. These scores can be converted into percentiles or standard scores to construct a normal distribution or test norms. Typical and unusual error patterns can also be documented. Identification accuracy can be increased if the normative sample only includes typically developing children (Peña et al., 2006).

The often-cited paper by McCauley and Swisher (1984) discusses the importance of establishing psychometric properties for speech and language assessments. The Contextual Probes of Articulation Competence – Spanish (Goldstein & Iglesias, 2009) provides an example of how to use McCauley and Swisher’s guidelines when operationalizing a speech sampling tool and Friberg (2010) provides an updated version of these criteria.

Finally, Stockman (1996: 355) offers the following words of caution ‘an adequate assessment is not insured simply by creating new norm-referenced standardized tests’. Assessment using a speech sampling tool should be complemented by other techniques, including criterion-referenced tools and dynamic assessment, as well as listening carefully to the children, their parents and teachers (see McLeod, Chapter 13).

Acknowledgment
Preparation of this chapter was supported by an Australian Research Council Future Fellowship (FT0990588) and helpful insights from Kathy Yuet Sheung Lee are also acknowledged.

References


