Evaluation Criteria Checklist (December 2016)

Purpose of document: To guide SLPs in choosing reliable and valid assessment tools with adequate psychometric properties when initially qualifying students for Speech/Language Impairment (SLI) in the schools. Current Minnesota eligibility criteria for SLI is available on the State of Minnesota's website (Minnesota Administrative Rule 3525.1343). *This checklist is intended to be considered as the examiner reads the entire test manual, including the administration section.*

The Evaluation Criteria Quick Checklist is available to assist SLPs in their decision making regarding whether an assessment has adequate psychometric properties to be used clinically. Below are detailed descriptions of many of the checklist criteria to help guide decision making. The checklist is based on largely on Friberg (2010), which was based on McCauley and Swisher (1984).

❑ **1. Is this the most recent version of the test or was it published within the last 10 years per MDE best practice guidance?**

❑ **2. Is the purpose of the assessment tool identified in the manual?**

"The purpose of a test is an important component of any assessment tool, as testing is often completed for very different diagnostic reasons (Hutchinson, 1996; Peña, Spaulding, and Plante, 2006). For instance, while an assessment tool might be administered to diagnose the presence or absence of a disorder it might also be used to determine the severity level of a known disorder or to establish treatment goals and/or objectives. This information indicates that clinicians need to be cognizant of the purpose of a given test in order to collect data reflecting their diagnostic needs. Additionally, clinicians need to be aware that assessment tools might purport to serve a specific purpose, but offer no data to substantiate the validity of using a test for that rationale. Further, clinicians should have the awareness that any given standardized testing battery 'may not be able to support multiple diagnostic purposes' (Peña et al., 2006: 252)." (Friberg, 2010, p. 81)

❑ **3. Are the qualifications for the person administering the test explicitly stated?**

"For an assessment tool to demonstrate this criterion, it need[s] to specify any special training/qualifications potential necessary to administer and score the test in question. This information is considered to be essential to the validity of a test, as any data collected cannot be considered valid if it is administered and/or interpreted by an unqualified individual." (Friberg, 2010, p. 82)

- ❑ **4. Are the testing procedures sufficiently explained? Will the examiner know how to administer the tool within the testing time frame suggested in the manual? Can the results be duplicated when given by another examiner?**

  "For this criterion to be considered present, sufficient detail must (have) be(en) provided within the examiner's manual to allow for test administration in a manner duplicating the conditions and procedures present at the time the test was standardized. Without this information, clinicians cannot be confident that they are administering the assessment tool in a way that matches the presentation of the test to those in the standardization sample. Any differences in how standardized assessment tools are administered yields scores that cannot be reliably compared to the normative sample. Thus the quality of the data collected can be compromised, rendering test scores unusable for the purpose(s) they intended to fulfill." (Friberg, 2010, p. 82)

- ❑ **5. Are standardized scores reported that can be applied to the category of eligibility being assessed?**

- ❑ **6. Is there an adequate standardization sample size (> 100)?**

  "For an assessment tool to have an adequate sample size, it needs to have a normative sample of 100 or more children per subgroup within the standardization sample. The inclusion of fewer children in a subgroup decreases the validity of test results, as the consistency of test scores is questionable. Test scores that are compared to larger groups of children are more stable, and thus can be used more dependably in the clinical decision-making process. Smaller sample sizes can also be indicative of a less representative sample for comparing scores, as with a small group of children included in the standardization pool it becomes doubtful that all possible subgroups of children (e.g. ethnicity, socio-economic status) have been included in a satisfactory manner, thus rendering the assessment tool in question unusable in many clinical settings." (Friberg, 2010, p. 82)

- ❑ **7. Are the characteristics of the standardization sample clearly defined? How well does the sample represent the population to which the child will be compared?**

  For an assessment to be used by a particular SLP, the geographic representation, socio-economic status/parent education representation, gender distribution, ethnic background, presence/absence of impairment(s), and age distribution of the normative sample should be reflective of the population served.

- ❑ **8. Does the manual provide evidence of item analysis?**

"Item analysis is used to maximize both the reliability and quality of questions included within a particular test battery by looking at the content of individual questions, screening items for inclusion in the assessment tool, and ensuring that tests target the skills they purport to measure. If an assessment tool lacks item analysis, it is possible that questions might be included that are too difficult or fail to access the skills in question. Thus, use of an assessment tool that fails to report data relative to item analysis could lead to clinical judgments being made on the basis of test questions that were poorly constructed." (Friberg, 2010, p. 83).

A variety of methods may be acceptable for item analysis evidence, including the two most common forms: Classical Test Theory, which looks to improve the reliability of standardized assessment tools, and Item Response Theory, which reflects the probability of performance as a function of a particular level of functioning (Fan, 1998).

❑ **9. Does the assessment have adequate measures of central tendency?**

The mean and standard deviation of all subtest scores for all groups of the normative sample must be reported to meet this criterion. "As these measures are the basis for other scores that are derived for comparison of performance, an assessment tool that fails to report these scores lacks flexibility in the use and interpretation of test its scores, which can impact the validity and reliability of the scores derived from a given testing instrument." (Friberg, 2010, p. 83)

❑ **10. Does the assessment report a reasonable level of error reported (e.g., standard error of measurement, confidence intervals)?**

"Test manuals for norm-referenced tests should provide the examiner with the standard error of measurement (SEM), which takes into account the variability that is inherent in human behavior." (Kennedy, 2007, p. 53). The SEM is used to determine the confidence interval around a person's test score.

❑ **11. Is there clear and supportable rationale for test content? Are there directions for alternate administration and/or scoring for the test?**

"Does it assume a certain cultural background, degree of world knowledge, or proficiency in a language the child does not have? Are the justifications for the definitions and selections of the test items clear enough so that a user could generate additional items or exercises to fit the test model?"

❑ **12. Is reasonable concurrent validity documented?**

"To demonstrate this criterion, the examiner's manual of each test needs to provide verification of concurrent validity; specifically, evidence demonstrating a correlation between results obtained from the test in question as well as other, similar assessment tools in indicating the presence or absence of a communication disorder. Concurrent validity is important because it demonstrates that results from a given assessment tool are more likely to be valid if a tool that assesses a similar construct has yielded analogous results." (Friberg, 2010, p. 83)

❑ **13. Is reasonable predictive validity documented?**

"To possess predictive validity, the examiner's manual for each test needs to provide evidence that performance on a given test is predictive of performance observed in a more functional setting through direct observation or other, less formal measures (e.g., student observed using specific language skills within a classroom environment). Absence of predictive validity leads to uncertainty as to how assessment tools and real-life tasks can be compared. Further, decisions related to intervention planning could be compromised as a result of a lack of reliability evident in test scores collected from such instruments." (Friberg, 2010, p. 83)

❑ **14. Is internal consistency adequate?**

Internal consistency measures man include split-half, Kuder-Richardson, or coefficient alpha, among others. Reliability coefficients should be greater than .90 (McCauley, 2013).

❑ **15. Is test/re-test reliability adequate?**

"For a test to demonstrate this criterion, values for test–retest reliability must be reported in order to ensure that scores attained on a given test are stable over time. A correlation coefficient of greater than .90 is required for an assessment tool to satisfactorily meet this criterion (McCauley and Swisher, 1984). It should be noted that for tests to possess acceptable levels of test–retest reliability, a short test–retest interval should be observed, as longer intervals between testing could lead to an inflated reliability coefficient that reflects not the actual test–retest reliability of a given test, but spontaneous recovery or maturation that could naturally occur outside a testing situation. A test with that lacks test–retest reliability might yield scores that would fluctuate over time, thus compromising the reliability of reported results." (Friberg, 2010, p. 84)

❑ **16. Is inter-rater reliability adequate?**

"Evidence of inter-examiner reliability must be reported in the examiner's manual for this criterion to be present. Inter-examiner reliability ensures that test scores do not fluctuate

when different clinicians administer the test battery. A correlation coefficient of .90 is required for a test to meet this criterion (McCauley and Swisher, 1984). A score lower than this cut-off value demonstrates a lack of reliability, as significantly different scores could be observed if the same child was administered the same test by two different clinicians." (Friberg, 2010, p. 84)

❑ **17. Are sensitivity and specificity measures included? If yes, consider questions below. The questions are based on Dollaghan's CADE: Critical Appraisal of Diagnostic Evidence (Dollaghan, 2007)**

1. Was the target assessment compared to a reference standard (e.g. gold standard)?

2. Was the reference standard valid, reliable, and reasonable?

3. Were measures administered independently?

4. Were measures administered with blinding?

5. Were methods and participants specified prospectively?

6. Were participant recognizable and representative of the actual diagnostic task?

7. Were the reference standard and the target assessment both administered to all participants?

8. Was LR+ (sensitivity/1-specificity) ≥ 10.0?

9. Was LR- (1-sensitivity/specificity) $\leq .10$?

For more information on school evaluations please see the ASHA School Evaluation Guide.

References:

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*: Paul H Brookes Publishing Company.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92.

Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. *Language, Speech, and Hearing Services in Schools, 27*(2), 109-121.

Kennedy, M. (2007). Principles of Assessment. In R. Paul & P. W. Cascella (Eds.), *Introduction to Clinical Methods in Communication Disorders* (Second ed.). Baltimore: Brookes Publishing Co.

McCauley, R. J. (2013). *Assessment of language disorders in children*: Psychology Press.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*(1), 34-42.

Pena, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech Language Pathology, 15*(3), 247-254. doi: 10.1044/1058-0360(2006/023)