Limitations of Norm-Referenced Tests

This document was created as a reference document to accompany the presentation by Dr. John Heilmann on "The Limitations of Norm-Referenced Tests". Additional information related to this topic is also provided in the Wisconsin DPI Significant Discrepancy and Consideration for Speech or Language Impairment document.

Limitation of Norm-Referenced Tests #1: Many Potential Biases

Norm-referenced tests have potential biases, some that are blatant and others that are more subtle. More blatant biases include the fact that English tests are not appropriate for students with limited English proficiency.

Norm-referenced tests also may be biased based on the speaker's dialect spoken. Many norm-referenced tests examine inflectional morphology (e.g., walk/ing; talk/s). Zero marking of third person singular and present progressive is a systematic rule within African American English, making these tests that assess these aspects of language biased if used with speakers of dialects other than General American English.

More subtle biases also exist for norm-referenced tests, such as format biases. The format of tests are based on traditional Western values (e.g., individual achievement, competitiveness, value of speed, adhering to time schedules, emphasis on logical thought and objectivity, belief that performance can be objectified). If a student does not share the same values, they will struggle with the format of the test. For example:

- A student from the nondominant culture may have less experience completing the tasks on a given test. Less experience means poorer performance and scores that are artificially low.
- Most tests assume that children share common beliefs/values. If a student does not share beliefs/values, tasks will be unfamiliar, and the test is biased.
- Among norm-referenced tests, there is the assumption that children tolerate a certain level of frustration to achieve an ending point, or ceiling. Most students from the dominant culture have experience and know this expectation. This assumption can be problematic if this is something new for a student.

What can an evaluator do to combat this bias? In order to account for the biases of norm-referenced tests, it is important to make adjustments to the assessment plan and take into consideration the recommendations from the comprehensive assessment framework.

- The new rule for Speech or Language Impairment is consistent with the comprehensive assessment model. If norm-referenced tests are administered to a

student, another SLP probe should be used to determine whether any identified need on the norm-referenced test is noted on other measures.
- If you have a student who is an English Learner, tests for monolingual English speakers are <u>not</u> appropriate (student would not be represented in comparison sample). Other assessments that would be appropriate to conduct include (but are not limited to) dynamic assessment, observations, interviews, review of academic tests and writing samples.
- Test scores should not be reported for students who are culturally and linguistically diverse if the student is not represented in the normative sample.

Limitation of Norm-Referenced Tests #2: Not as Accurate as We Think

It is important to review the psychometric properties of norm-referenced assessments before delving into this limitation. The following terms are important when considering the accuracy and appropriateness of individual norm-referenced tests during assessment.

- The *reliability* of a test indicates that data obtained are stable and repeatable (e.g., results are similar when someone takes a test on Monday & Friday; results are similar across odd and even questions).
- The *validity* of a test indicates the test measures what an evaluator believes it is measuring. For example, test developers will often show that performance on their test has a strong relationship with performance on other tests.
- The diagnostic accuracy of a test assists the evaluator with knowing whether the test data result in an accurate diagnosis.
    - The *sensitivity* of a test indicates how accurately the test diagnoses impairment (true positive).
    - The *specificity* of a test indicates how accurately the test diagnoses that no impairment is present (true negative).
- The *cut-off score* of a test is the score at which the test distinguishes typically developing students from students with an impairment.
    - Performance above the score = typically developing
    - Performance below the score = impairment/disorder.

Obtaining a standard score below the cut score for a norm-referenced test is indicative of performance below the expected range on that measure when compared with the normative sample. Results from norm-referenced tests should be one piece of information when considering an impairment and should be compared with other assessment information obtained about the student.

Why do we see differences across tests?
Evaluators see differences in scores across tests administered due to differences in each test's normative sample (i.e., race, ethnicity, SES, language spoken, dialect spoken,

inclusion of children with disabilities). Different samples have different performance *especially noticed with culturally and linguistically diverse students.*

Why is performance so different for some students?
Performance varies across tests due to the biases in individual tests. If a student is not familiar with a test (or task within test), they will perform worse.

What can an evaluator do to overcome this limitation?
SLPs and other evaluators must review test manuals to learn about the tests used. Most manuals now include psychometrics (e.g., reliability, validity, sensitivity, specificity, representation of your student in the normative sample).
- [VDOE SLP Test Comparison Chart](#)
- [VDOE Four Additional Tests](#)

If an evaluator chooses to use a norm-referenced test, they must account for the culture, language, and dialect of the student being assessed.

Limitation of Norm-Referenced Tests #3: Decisions Aren't Always Clear Cut

The primary purpose of many norm-referenced tests is to show an individual's *relative standing* compared to a reference group (i.e., rank-order a group of individuals) and to identify at what point performance is significantly different from peers. The general research criterion is -1.25 SD. Historically in Wisconsin, the criterion has been -1.75 SD. The common criterion utilized across states in recent years has been -1.5 SD. It should be noted that *all of these are arbitrary and may not be accurate* because of the many potential biases (discussed above) as well as due to the acknowledged imprecision in tests (which is why confidence intervals, and a standard error of measurement are ways to report the data). In addition, the standardized procedures and format of norm-referenced tests do not provide information about a student's functional communication. Test items are chosen because they differentiate kids, not because of educational impact. Most items are *decontextualized* and removed from any communicative intent.

What can an evaluator do to overcome this limitation?
- Evaluators may utilize the comprehensive assessment model above as well as DPI adopted rubrics to review data and to determine what the preponderance of evidence is indicating for an individual student. Any concerns indicated by a score on a norm-referenced test should be consistent with evidence from other data sources that consider a student's skills in more functional contexts (e.g., performance in classroom activities or during language samples).

Limitation of Norm-Referenced Tests #4: Functional Communication is Not Assessed

The standardized procedures and format of norm-referenced tests does not provide information about a student's functional communication. Test items are chosen because they differentiate kids and have the ability to distinguish between high and low performers and do not assess educational impact. Most items are *decontextualized*, meaning taken away from actual communication. Some tests focus on skills that are more functional than others (e.g., Test of Narrative Language assesses a student's ability to understand and tell stories).

What can an evaluator do to overcome this limitation?
- Pair data from norm-referenced tests with other data from other probes. SLPs should utilize functional assessments including dynamic assessment, play-based assessment, and language sample analysis and look for patterns of language strengths and difficulties.

Conclusion

Norm-referenced tests have a prominent place in assessment practices. The biases inherent in these tests should give evaluators pause and consider other assessment methods and practices that are more culturally responsive and that assess how well a student is able to use language functionally across tasks and environments. Norm-referenced tests have a place within the comprehensive assessment model framework when included as one of many sources of data for individual students.