Standard Setting with Innovative Measures of Early Literacy: Contrasting Groups

Michael C. Rodriguez, Anthony D. Albano, Scott McConnell,

Alisha Wackerle-Hollman, & Tracy Bradfield

University of Minnesota

Paper Presented at the Annual Meeting of the

National Council on Measurement in Education, New Orleans, LA.

April, 2011

Standard Setting with Innovative Measures of Early Literacy: Contrasting Groups

Effective response to intervention (RTI) requires information to make placement decisions. The use of IGDIs for assessing tier placement is an important goal. Consistent with the *Testing Standards* (AERA, APA, NCME, 1999), current standard setting practices were reviewed, with respect to available evidence supporting their use and appropriateness given the unique context of early childhood classrooms and the nature of the IGDIs. Contrasting Groups Design was selected as the most appropriate (see Cizek and Bunch, 2007). To facilitate the Contrasting Groups procedures, tier level descriptors (TLDs, similar to PLDs or performance-level descriptors used in NCLB assessments) were developed. These included elaborations of knowledge, skills or attributes of preschool children at each tier level. Initial drafts of TLDs were written by the lead investigators responsible for developing tier-specific interventions and curricula, and revised through iterative review by curriculum and IGDI developers. The paper will describe how the Rasch model was used to facilitate the development of score ranges, providing guidance for Tier-level placement, based on accuracy and precision of cut-scores. We will also briefly describe opportunities for validation and the kinds of validity evidence that may be available at each stage of the process.

An important functional step in the RTI framework is the identification of children who are likely to benefit from Tier 2 or 3 intervention. This identification process is improved through the collection and use of relevant information. That is, information relevant to the skills seen as important to successful literacy and language development through early childhood and early elementary school is important to assess and monitor to ensure successful progression. IGDIs 2.0 were developed with the intention of contribution to this information gathering process. The IGDIs are being evaluated as identification and progress monitoring tools. The first component, described here, includes the identification process.

To facilitate identification, cut scores or cut score ranges (as described below), need to be selected on each of the high performing IGDIs. To support this process, we used a standard setting process known as Contrasting Groups Design. Through this process, we are able to identify a point on the IGDI score scale that optimally distinguishes or contrasts known groups – groups of children likely to benefit from Tier 2 or 3 placement, based on teacher judgment. This process used in the CRtIEC initial standard setting study, based on end-of-year performance of 4 or 5 year old children prior to entering kindergarten.

Compared to more traditional Curriculum Based Measurement (CBM) of which IGDIs are an example, this approach is moving away from norm-referencing toward criterion referencing – as RTI becomes more successful, norm referenced systems become less meaningful. Typically, there is a normative decision making framework for identifying children as eligible for Tiered intervention, where 5% may need Tier 3 intervention and 10% may need Tier 2 intervention. In addition, through the construct map concepts of Wilson (2004), this approach is helping to connect measurement, instruction/ intervention, and learning in more coherent ways. Finally, through this integration, the CRTIEC effort is employing an iterative process for developing a sound assessment system to support RTI.

Several of the Testing Standards speak directly to the importance of establishing cut-scores and essential elements of reporting to support such decisions, including:

Standard 4.19

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

Standard 4.20
When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.

Standard 4.21
When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Recently the Contrasting Groups Design has been used successfully in the Kansas general Reading and Mathematics assessments standard setting process in 2006 and in the Nebraska 2006 standard setting process. When examining a very early application of this method in Kansas, Poggio (1984) reviewed the contrasting groups method and concluded:

1. the method is rather easily implemented;
2. teachers report little difficulty in following what is to be done in Contrasting Groups;
3. the public is both confused and tends to doubt the legitimacy of the standard when they (often) cannot understand the "statistical magic" which delivers the standard (commonly associated with more complex standard setting methods); and
4. the method gives support to the idea that "teachers can already tell us who is competent."

The Contrasting Groups Design

The process used in the CRtIEC standard setting study, based on end-of-year performance of 4 or 5 year old children prior to entering kindergarten, included these steps:

1. Teachers of children at the end of the year prior to entering kindergarten were invited to complete a child-performance survey, without information on performance from the assessments.
2. Teachers were asked to place children into a Tier level (1, 2, or 3), based on their understanding of the performance level from the tier level descriptors (TLDs). These assignments were made for each of the domains independently, including (a) oral language, (b) phonological awareness, and (c) alphabet knowledge.
3. Children were assessed on the IGDIs and the distributions on the actual measures for each performance level were compared.
4. The points (cut-scores) that discriminate among children between tier levels were estimated using multiple methods to assess agreement and sensitivity to method, including
   a. ROC analysis to achieve a balance between Sensitivity and Specificity, with a minimum Sensitivity of .70;
   b. A check on ROC analysis through logistic regression and classification accuracy;
   c. Analysis of the effect of the selected cut score on the distributions of scores.
5. This cut score estimation process was conducted on Wave 1 data directly based on the logit metric (measure from Winsteps) and the True Score (estimated from a test of 20 cards), based on Wave 4 Teacher Tier Placements of children. The process was replicated on Wave

4 data (setting the cut score at Wave 4), and then using the Wave 4 cut score to predict the Wave 1 cut score.

The following process provides the technical background and statistical results in the series of analyses employed to estimate the appropriate cut scores for the Identification Bundles to be used to inform Tier placement decisions.

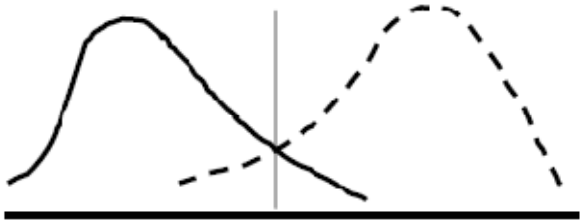
Two major steps to complete this process include:

1. Estimate the length of Identification Bundles and cut score ranges to secure number correct scores that will inform decisions such that:
   a. Below some number, we highly recommend Tier 2 placement;
   b. Within some range of scores we recommend additional information to make the placement decision; and
   c. Above some number, we recommend Tier 1 placement (business as usual).
   This could be achieved by constructing a test with the number of items that yields a confidence interval around the cut scores that is as wide as the range of scores below that confidence interval.
2. Recommend Identification Bundle length and appropriate cards to include.

Through the use of IGDI 2.0 Identification bundles, information is obtained that is construct and instructionally relevant given the RTI framework developed for early childhood education. IGDI results should be used within a clearly defined decision-making process as one source of information in a multiple-measures decision-making framework. No important educational decisions should be made with a single piece of information and certainly not with a single test score.

Note: An important limitation of these analyses should be taken into consideration. The teacher placement decisions were completed at the end of the year based on end-of-year skill-level expectations. The cut scores are being set for fall decision making. To reduce potential bias, IGDI scores at both waves 1 and 4 were scaled on the Wave 1 scale through Rasch equating.

A simple example is below, where the intersection of the two distributions is identified as the cut-point on the measure defining the difference between the two tiers.



Tier (Performance) Level Descriptors – Quality of PLDs Makes this Method Effective

To facilitate the Contrasting Groups standard setting procedures, tier level descriptors (TLDs) must be clearly defined. These must be elaborations of knowledge, skills or attributes of

individuals at each tier level. The elaborations should be relevant given the intended nature of the interventions in each tier, as they must identify children that are likely to benefit from tier level interventions. Initial drafts of TLDs were written by the lead investigators responsible for developing tier-specific interventions and curricula. They were given the general instruction: *Considering the content and procedures in your planned treatment, what are the characteristics of children who would benefit from Tier 2/Tier 3 [i.e., your!] intervention?*

Descriptive TLDs were written and are presented below. These more extensive TLDs were summarized and behavioral elements were highlighted in the directions to teachers for the Teacher Survey. The specific elements given to teachers are also provided in the tables below.

Vocabulary and oral language skills refer to a child's knowledge and use of words and grammatical sentence constructions to communicate verbally.

Advanced: Preschoolers with advanced vocabulary and oral language skills have a larger than expected vocabulary for their age and generally communicate in grammatically correct complete sentences, including some complex sentences (e.g., sentences that include two verb phrases or express complex ideas, such as “I went to the store because I needed milk”). They use a variety of words including nouns, verbs, adjectives, and adverbs to describe, tell similarities and differences, relate narrative events, and retell stories in a sequential and cohesive manner. They talk about people, places, things and events not present.

Competent: Preschoolers who are competent in vocabulary and oral language skills use a variety of words (i.e. nouns, verbs, adjectives, adverbs) to convey meaning in conversation and in most daily activities. They generally communicate in grammatically correct short sentences, and can describe concrete objects and people, places, and things that are in their immediate environment. They may use conjunctions, but have limited use of complex sentence structure with independent clauses. With adult support they can tell simple narratives and talk about people, places, things, and events not present.

Need Tier 2 Support: Preschoolers who need tier 2 support in vocabulary and oral language development use core vocabulary words consisting primarily of nouns and verbs in simple sentences in conversation and daily activities. These children may have sufficient oral language skills to engage in routine, everyday conversation, but may struggle to engage in academic discussion (“school talk”) or conversation about unfamiliar topics. They tend to use nonspecific words (e.g., "this, that, stuff") when describing objects people and places and generally have difficulty engaging in conversations about people, places and things that are not in their immediate environments, telling coherent narratives, or retelling stories in sequence.

Need Tier 3 Support: Preschoolers who need tier 3 support have limited verbal skills. These children use one- and two-word utterances and short phrases to communicate. They have difficulty describing objects, people, and places and do not engage in narrative discourse. They may exhibit frustration or challenging behavior related to limited communicative skill.

Oral Language TLDs provided to Teachers:

Tier 3	Tier 2	Tier 1
Describes a student that: • Has limited verbal skills • Uses primarily **1 to 2 word utterances** and short phrases to communicate • Does not tell or talk about stories. • May **exhibit frustration or challenging behavior** related to limited communicative skill.	Describes a student that: • Primarily uses **nouns and verbs** in simple sentences during conversation. • Tends to use **nonspecific words** (e.g., "this, that, stuff") when describing objects, people and places. • Struggles to engage in conversation about unfamiliar topics. • Struggles to engage in conversation about topics not in their immediate environment. • Struggles to tell or talk about stories.	Describes a student: • That does not meet criteria for Tier 2 or Tier 3. • For whom you have no concerns in this area.

Phonemic Awareness is the explicit awareness that spoken words are made up of individual sounds or phonemes. It is a metalinguistic skill that involves attending to, thinking about, and intentionally manipulating the individual phonemes within spoken words and syllables. For example, the knowledge that the word "dog" begins with the sound /d/ is phonemic awareness. The ability to replace the /d/ sound at the beginning of "dog" with the /h/ sound to make the word "hog" is also phonemic awareness. Phonemic awareness is part of a broader class of *phonological awareness* skills that involve attending to, thinking about, and intentionally manipulating phonological aspects of language, including units larger than phonemes, e.g., syllables, onsets, and rimes.

Advanced: Preschoolers who are advanced in phonological awareness skills have an understanding that words are made up of individual sounds (phonemes) and can segment single syllable words into their component phonemes. They may be able to perform tasks that require the manipulation of sounds in words (e.g., blending sounds to make words, clapping out sounds in words, and elision tasks, such as “say meat without saying /t/”).

Competent: Preschool children who are competent in phonemic awareness skills demonstrate an awareness that words are made up of sounds. They can match words that begin with the same sound and identify the first sound in words.

Need Tier 2 Support: Preschoolers who need support in acquiring phonemic awareness skills do not yet understand that words are made up of individual sounds. They may have an awareness of larger phonological units as evidenced by their ability to perform rhyming, blending, and

segmenting tasks at the level of syllables or words, but cannot perform these tasks at the phoneme level.

Need Tier 3 Support: Preschoolers who need Tier 3 support in acquiring phonemic awareness do not have an awareness of the phonological aspects of spoken language. They cannot perform rhyming, blending, and segmenting tasks at the level of words or syllables.

Phonological Awareness TLDs provided to Teachers:

Tier 3	Tier 2	Tier 1
Describes a student that: • Cannot rhyme, blend word parts into words, or segment words into syllables.	Describes a student that: • Has emerging ability to recognize and/or make rhymes at the word level and blend and segment at the word or syllable level.	Describes a student: • That does not meet criteria for Tier 2 or Tier 3. • For whom you have no concerns in this area.

Alphabet Knowledge is knowledge of the letters of the alphabet including the ability to recognize and name upper and lower case letters and the knowledge of the most common sounds for letters. The knowledge of the alphabet paired with the understanding that the alphabet represents the sounds of spoken language is known as the *alphabetic principle*.

Advanced: Preschoolers who have advanced alphabet knowledge can recognize and name all of the upper case letters and many lower case letters. They know that letters represent sounds in words and know the most common sounds of most letters. These children use this knowledge to read and write simple phonetically regular words.

Competent: Preschoolers who are competent in alphabet knowledge know most of their upper case letter names and some lower case letter names. They demonstrate the ability to say what sound typically goes with many letters and know that letters represent sounds in words. These children are beginning to use invented spelling to write words (e.g., write "car" as "kr" or write "is" as "iz").

Need Tier 2 Support: Preschoolers who need tier 2 support know some upper case letter names (e.g., letters in name, most commonly known upper case letters). They may be able to say what sound goes with some letters, but do not have an understanding that letters represent sounds in words and therefore do not apply letter-sound knowledge to write words using invented spelling.

Tier 3: Preschoolers who need tier 3 support know very few letter names and do not know letter sounds. They do not have an understanding that letters represent sounds in words.

Alphabet Knowledge TLDs provided to Teachers:

Tier 3	Tier 2	Tier 1
Describes a student that: • Knows **very few letter names.** • **Does not know letter sounds.**	Describes a student that: • Knows **some upper case letter names.** • Is able to say what sound goes with **some** letters. • Does not understand that letters represent sounds in words.	Describes a student: • That does not meet criteria for Tier 2 or Tier 3. • For whom you have no concerns in this area.

Sample Characteristics

Teachers in preschools in four states participated, including . For each teacher, the research team randomly selected about six students from their classroom to classify using the TLD process. In total, they classified 824 of their children for the purpose of the Contrasting Groups standard setting process.

Student Frequencies by State

	Frequency	Percent
State 1	185	32
State 2	159	27
State 3	116	20
State 4	127	22
Total	587	

Student Characteristics (*n*=587)

	Proportion
ELL	.23
IEP	.12
Speech & Language	.09
Emotional/Behavioral	.01
Developmental Delay	.03
Other	.01

Results

Based on teacher classification of student performance, we found that teachers classified approximately 33% of students into Tiers 1 and 2 on Oral Language and approximately 46% on Alphabet Knowledge. In total, approximately 62% of children were classified into the same Tier on both constructs. This indicates that teachers seemed to consider each construct independently as they classified children into a given Tier level.

One immediate implication regards the small number of children identified for Tier 3 placement, including only 5 in Oral Language and 22 in Alphabet Knowledge. These small samples will not support the identification of a separate cut score between Tiers 2 and 3. The analyses below consider children classified in Tiers 2 and 3 together, resulting in a single cut score between Tiers 1 and 2/3.

A second implication is the identification rate of children classified at the levels skill at Tiers 2 and 3. Because 33% to 46% of children are identified by teachers, optimal methods used

below to identify the maximally discriminating cut score (contrasting groups) will place approximately 35% to 45% of the children in Tiers 2 and 3.

Teacher Classification of Students into Tiers for Oral Language & Alphabet Knowledge

Oral Language Tier Placement		Alphabet Knowledge Tier Placement 1	2	3	OL Total
1	Count	269	102	22	393
	% of Total	46%	17%	4%	67%
2	Count	43	65	35	143
	% of Total	7%	11%	6%	24%
3	Count	5	18	28	51
	% of Total	1%	3%	5%	9%
AK Total	Count	317	185	85	587
	% of Total	54%	32%	15%	100%

Oral Language

ROC analysis employing the Picture Naming Wave 1 measure logit score (Test variable) in classifying children on Oral Language Tier Levels (State variable) as placed by teachers. The test statistic associated with a ROC analysis, signifying the statistical significance of classification, is based on the area under the curve. “Area” is the probability that a score for a randomly selected positive case (Tier 2/3) is lower than the score for a randomly selected negative case (Tier 1). The ROC analysis suggests successful classification based on PN, Area = .782, $p < .001$. The cut score that yields .70 sensitivity and .66 specificity is 1.90 logits. Although based on Figure 1, the maximal value on both metrics is about .68, which is associated with a score of 1.81.

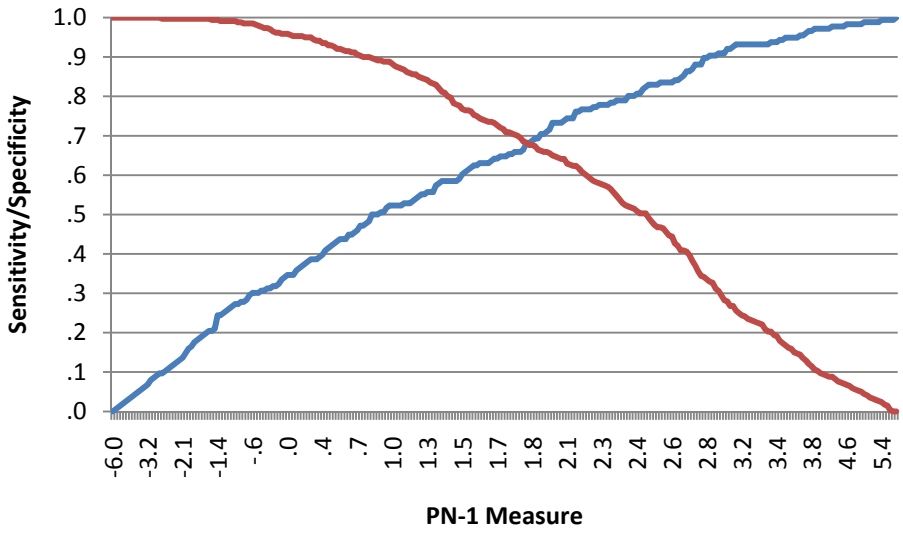


Figure 1. Illustration of the point of intersection of sensitivity (increasing line) and specificity (decreasing line) for PN Wave 1.

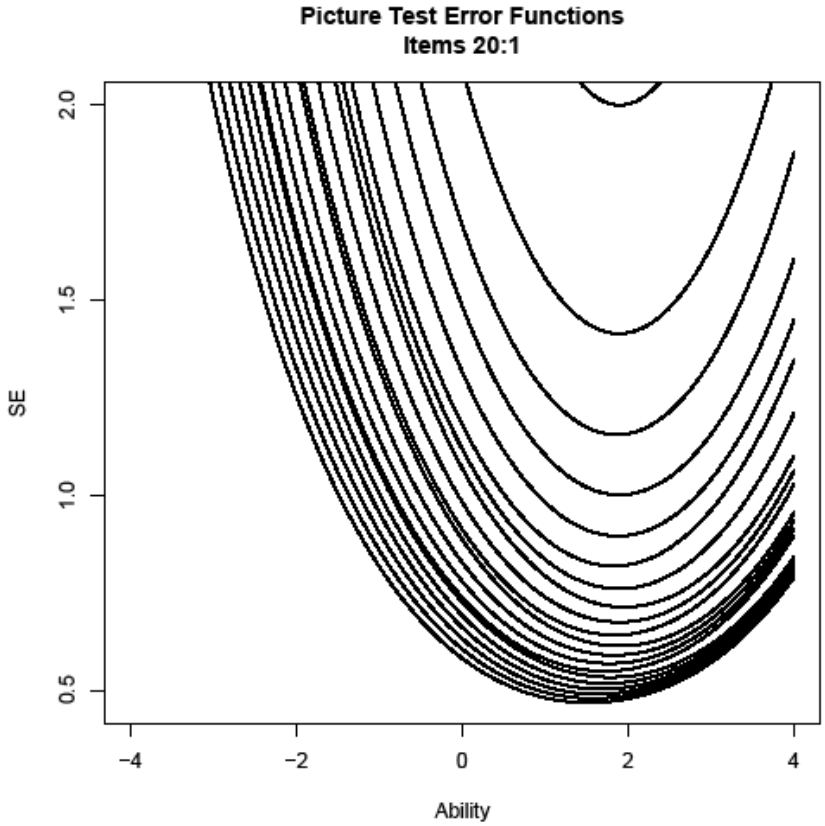


Figure 2. Picture Naming Identification Bundle measurement error functions given number of items.

In Figure 2, we illustrate the measurement error function across the ability range (logits) for every possible test length from 20 to 1 item. The lowest curve is based on 20 items and illustrates that the smallest SE is found within the range of approximately 0.5 to 2.5 logits. This is based on identifying 20 items most near the cut score (1.90), and successively removing the one item furthest from the cut point. You can see that tests of 4 or fewer items have SEs larger than 1.0.

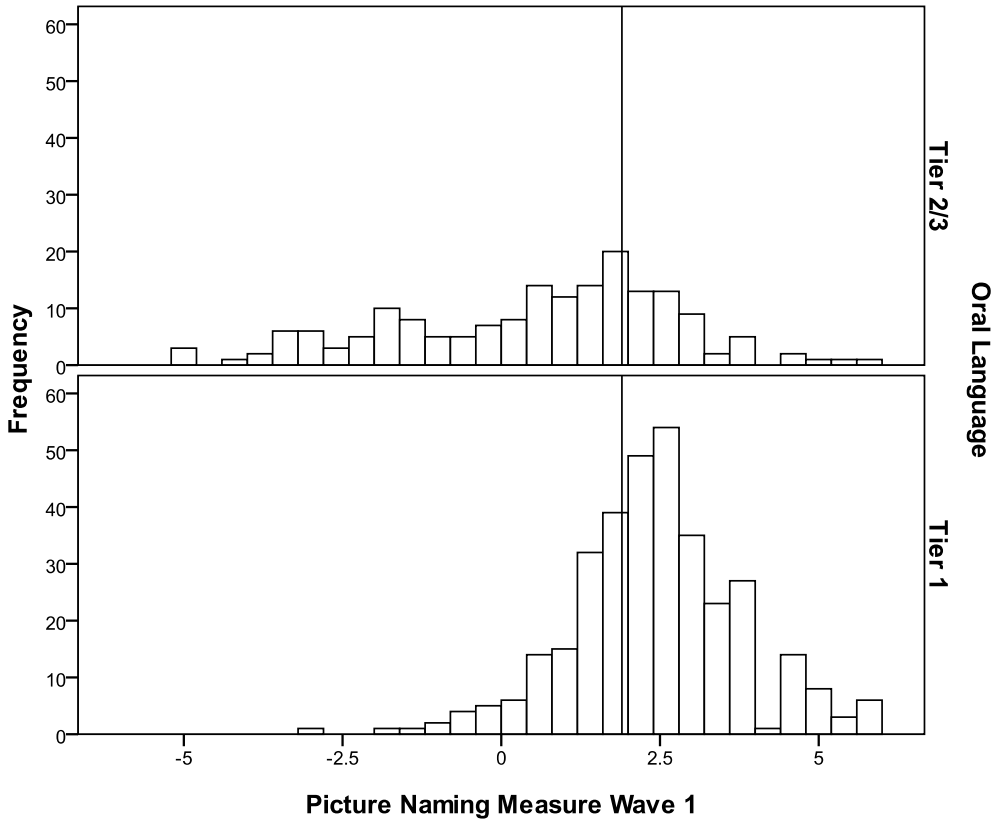


Figure3. Picture Naming measure histograms by Tier placement.

Figure 3 provides an illustration of the intersection (and overlap) of the score distributions of the two groups. The proposed cut score of 1.90 logits is marked on the graph.

Alphabet Knowledge

ROC analysis employing the Sound Identification Wave 1 measure logit score (Test variable) in classifying children on Alphabet Knowledge Tier Levels (State variable) as placed by teachers. The test statistic associated with a ROC analysis, signifying the statistical significance of classification, is based on the area under the curve. “Area” is the probability that a score for a randomly selected positive case (Tier 2/3) is lower than the score for a randomly selected negative case (Tier 1). The ROC analysis suggests successful classification based on SI, Area = .730, $p < .001$. See Figures 4 and 5 for ROC curves given fall performance (Figure 4) and spring performance (Figure 5). Note that spring performance yields stronger results since the teacher placements in tiers through the contrasting group standard setting was also conducted in the

spring. It appears that the spring performance is more sensitive to spring placements of teacher Tier classifications of students.

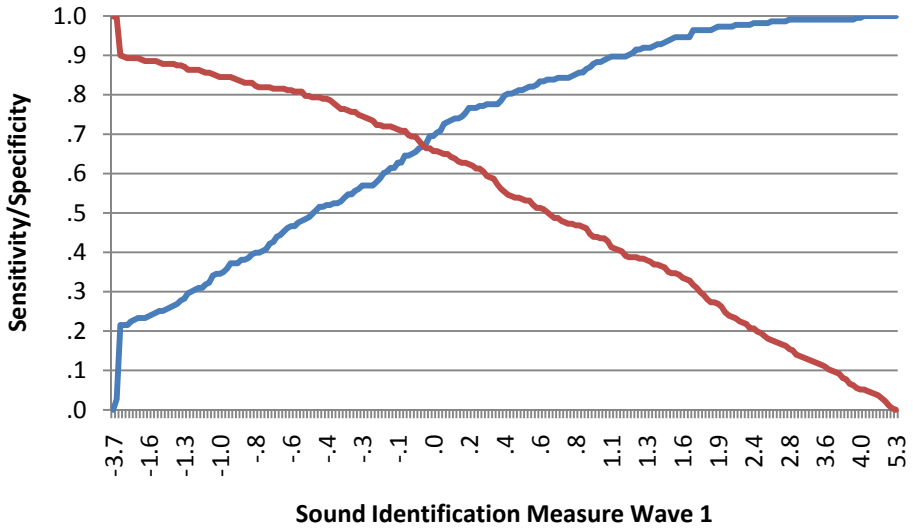


Figure 4. Illustration of the point of intersection of sensitivity (increasing line) and specificity (decreasing line) for SI Wave 1.

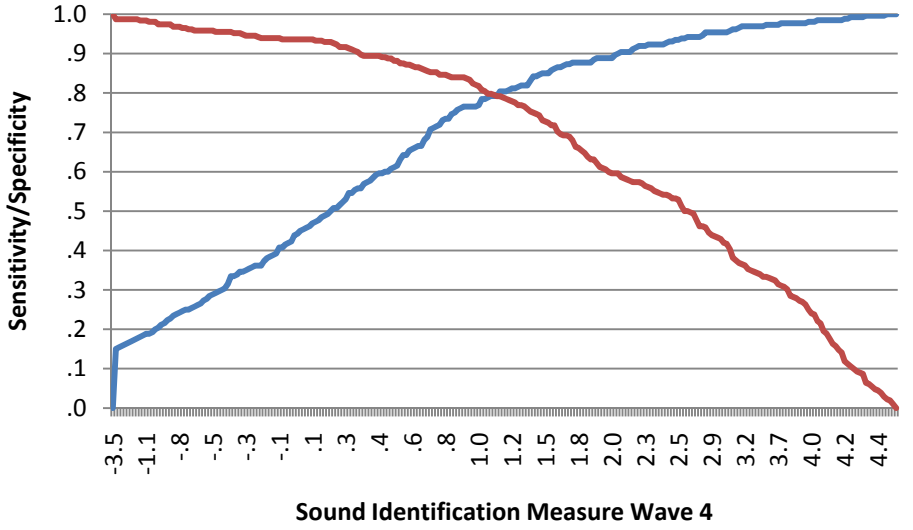


Figure 5. Illustration of the point of intersection of sensitivity (increasing line) and specificity (decreasing line) for SI Wave 4.

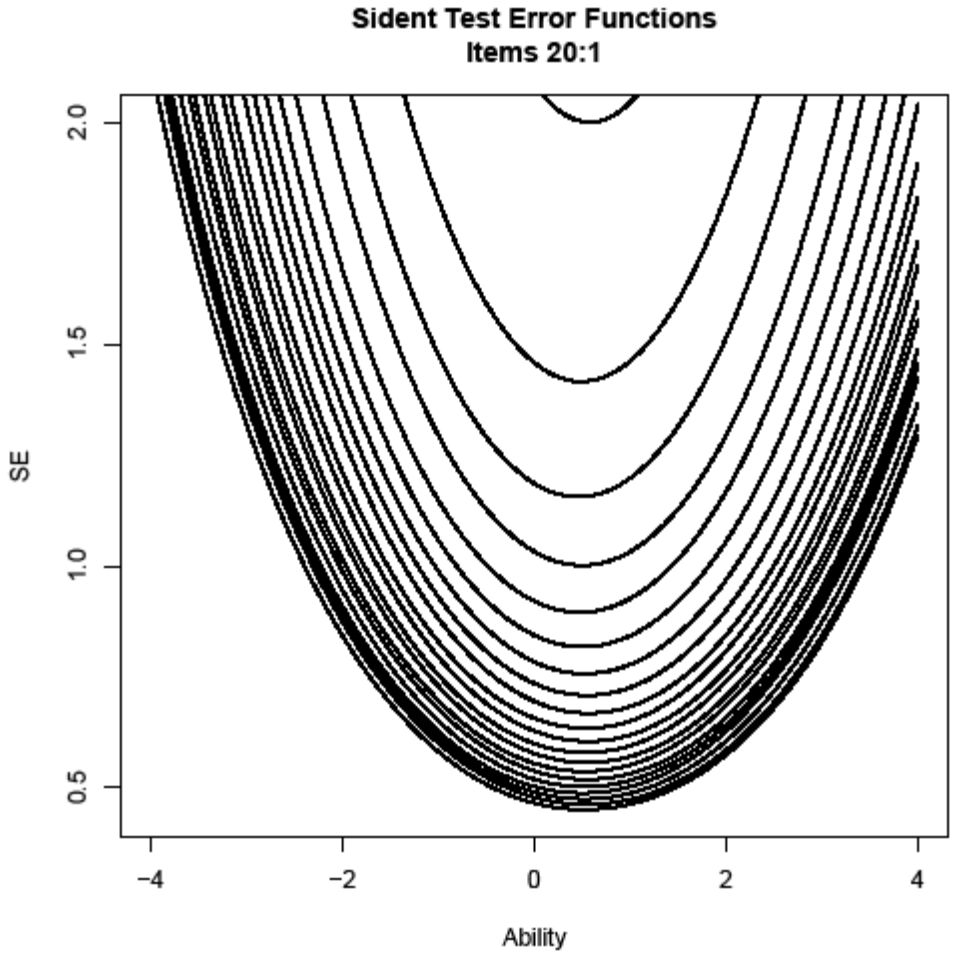


Figure 6. Sound ID Identification Bundle measurement error functions given number of items.

In Figure 6, we illustrate the measurement error function across the ability range (logits) for every possible test length from 20 to 1 item. The lowest curve is based on 20 items and illustrates that the smallest SE is found within the range of approximately -0.5 to 1.5 logits. This is based on identifying 20 items most near the cut score (0.05), and successively removing the one item furthest from the cut point. You can see that tests of 4 or fewer items have SEs larger than 1.0.

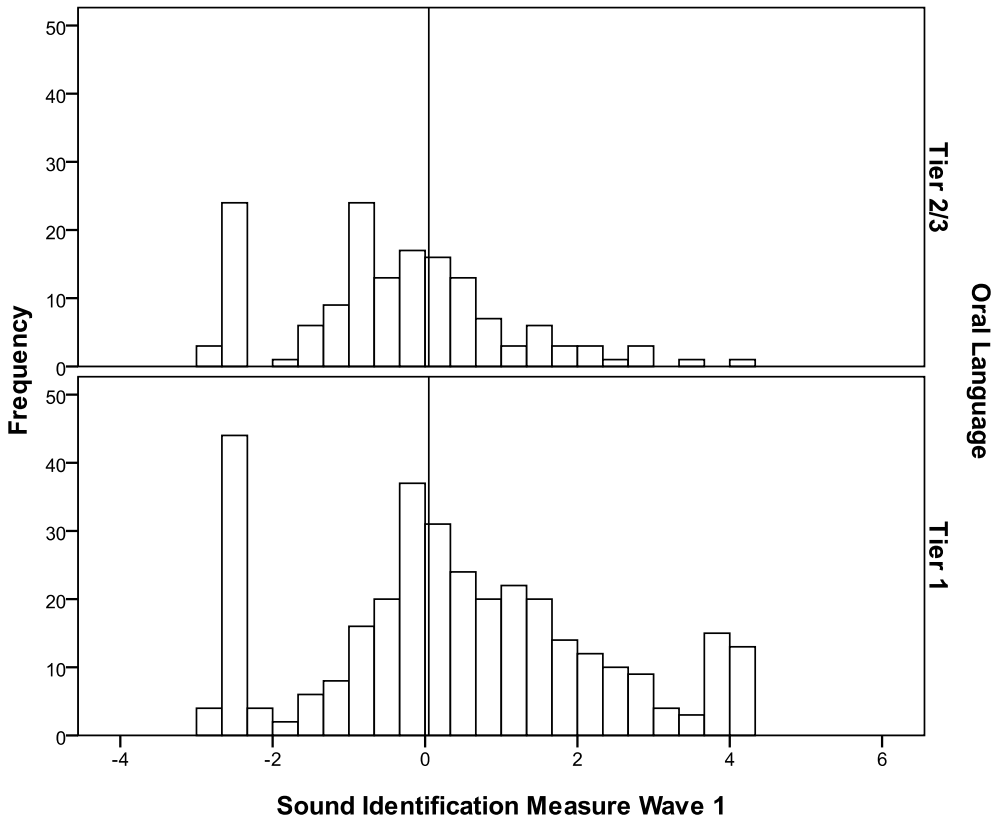


Figure 7. Sound Identification measure histograms by Tier placement.

Figure 7 provides an illustration of the intersection (and overlap) of the score distributions of the two groups. The proposed cut score of 0.05 logits is marked on the graph.

Validation Considerations

Perhaps the ultimate validation criterion is the performance of students given their Tier placement. Children placed in Tier one should grow and develop normally, ultimately becoming a successful reader at grade 3 (as one plausible criterion). Children placed in Tier 2 are likely in need of focused intervention to remediate a specific early language delay or limited skill area, such that Tier 2 intervention responsiveness will allow the child to re-enter Tier 1 settings (typical classrooms). Children placed in Tier 3 are in need of more intense intervention.

There are a number of ways to capture and utilize this information in an effective manner over time. An initial possibility is to assess Tier placement with a number of IGDIs. Although each IGDI is designed to capture specific subskills within each domain, there are common components across subskills that should yield some consistency in Tier placement across a domain of interest.

More directly, the amount of time a child spends within a Tier (as an indication of responsiveness to intervention) could serve as a signal regarding appropriateness of Tier placement. In addition, once a child transitions out of a Tier, reassessment with IGDIs will provide follow-up evidence about performance following intervention. If the IGDI correctly identifies children in need of Tier intervention and that intervention appears successful (given mastery monitoring assessment results or progress monitoring assessment results), the reassessment with IGDIs should reflect that change in performance.

Examining Teacher Tier Placements vis-à-vis Criterion Measures

Two criterion measures were available to assess the functioning of the Picture Naming IGDI and to evaluate the teacher Tier placement decisions. These included the CELF Expressive Vocabulary Test and the Peabody Picture Vocabulary Test (PPVT).

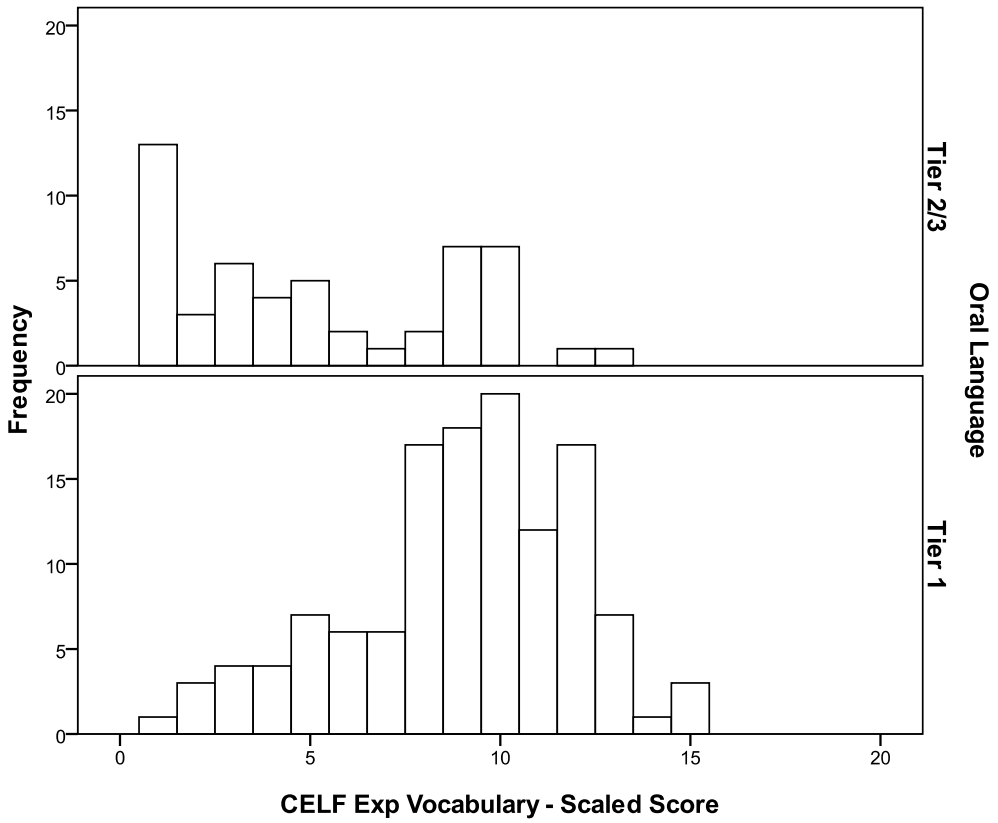


Figure 8. Score distributions of the CELF Exp Vocabulary test divided by Tier placement as determined by teacher placement.

Here we see the distributions of scores on the CELF Expressive Vocabulary scale for the two levels of Tiers based on Teacher placement.

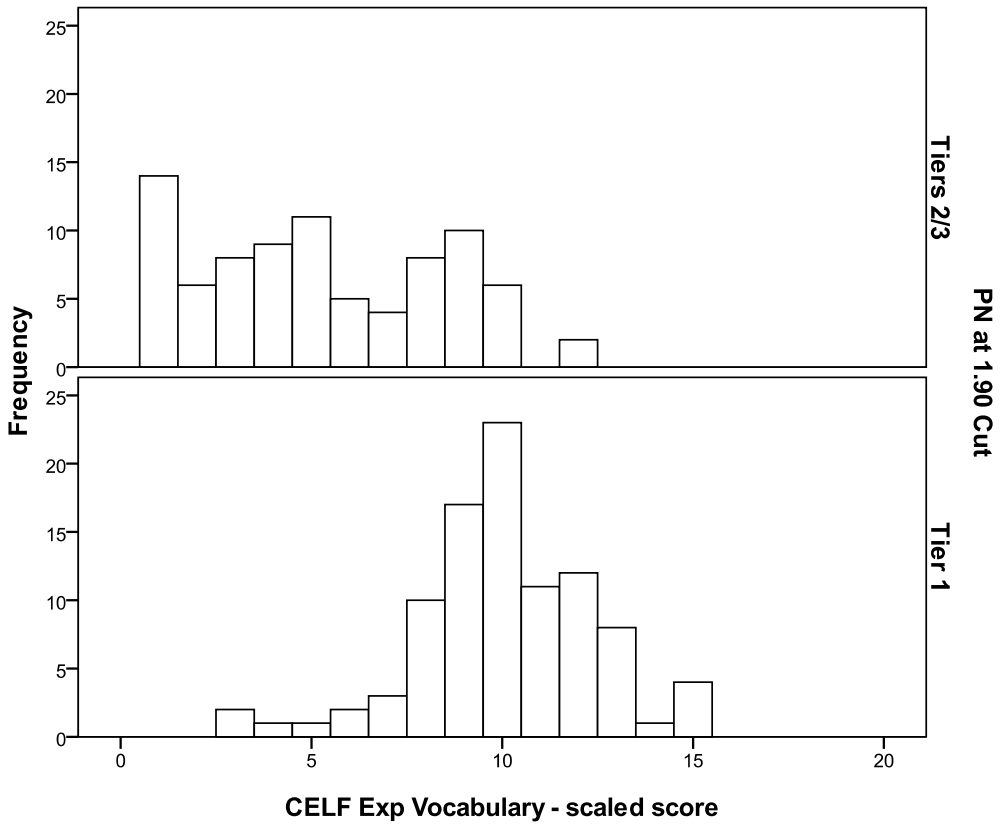


Figure 9. Score distributions of the CELF Exp Vocabulary test divided by Tier placement as determined by the cut score on Picture Naming.

Here we see the distributions of scores on the CELF Expressive Vocabulary scale for the two levels of Tiers based on identification from the Picture Naming measure with the cut score set at 1.90 on the logit scale.

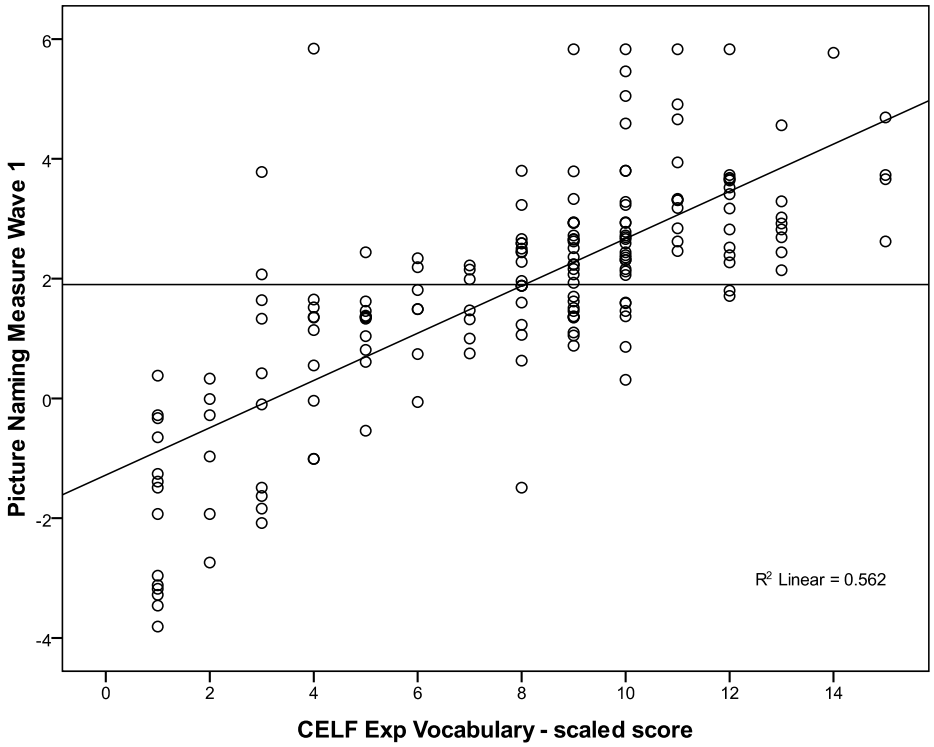


Figure 10. Scatterplot of the Picture Naming Rasch scale score (with cut score of 1.90 marked) and the CELF Expressive Vocabulary scale score, and the best fitting line (R^2 = .562).

The correlation between the CELF EV and PN Measure is approximately .75. Here we see the intersection of the best-fit line (regression line) and the cut score on PN of 1.90. The point of intersection occurs at a score of 8 on the CELF. A CELF score of 8 is associated with the 25th percentile performance. In this data set, 37% of the sample scored below 8 (10% obtained a score of 8).

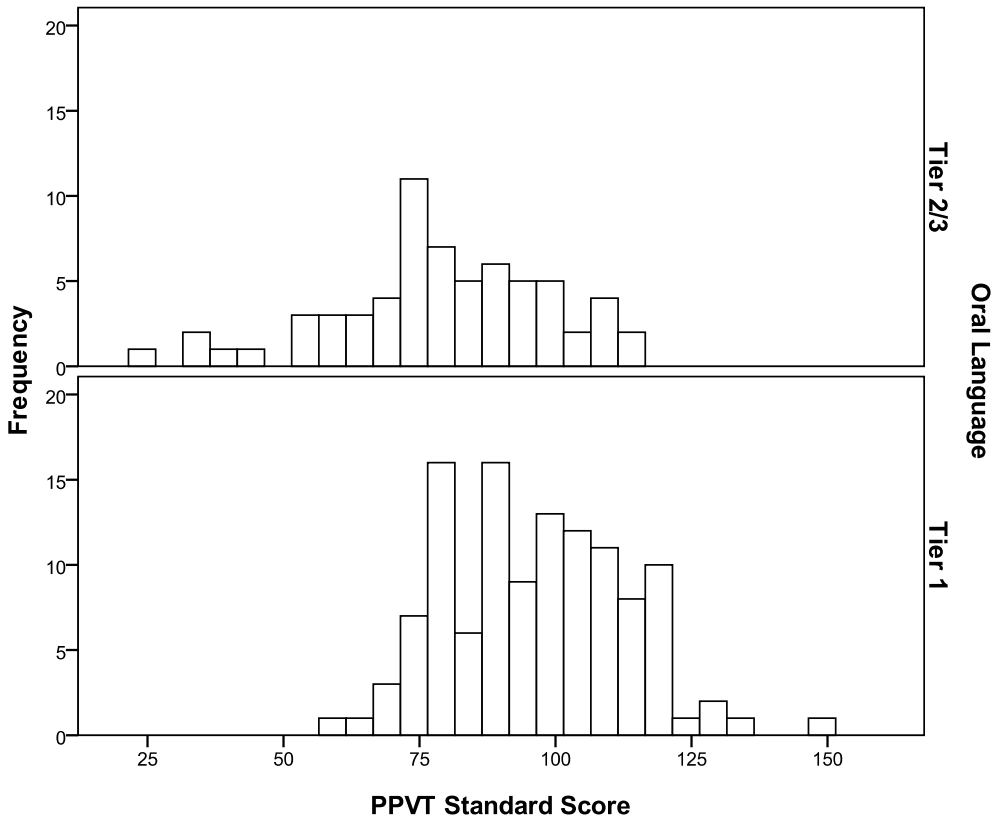


Figure 11. Score distributions of the PPVT divided by Tier placement as determined by teacher placement.

Here we see the distributions of scores on the PPVT scale for the two levels of Tiers based on Teacher placement. The Tier distributions have much heavier overlap with the PPVT than with the CELF.

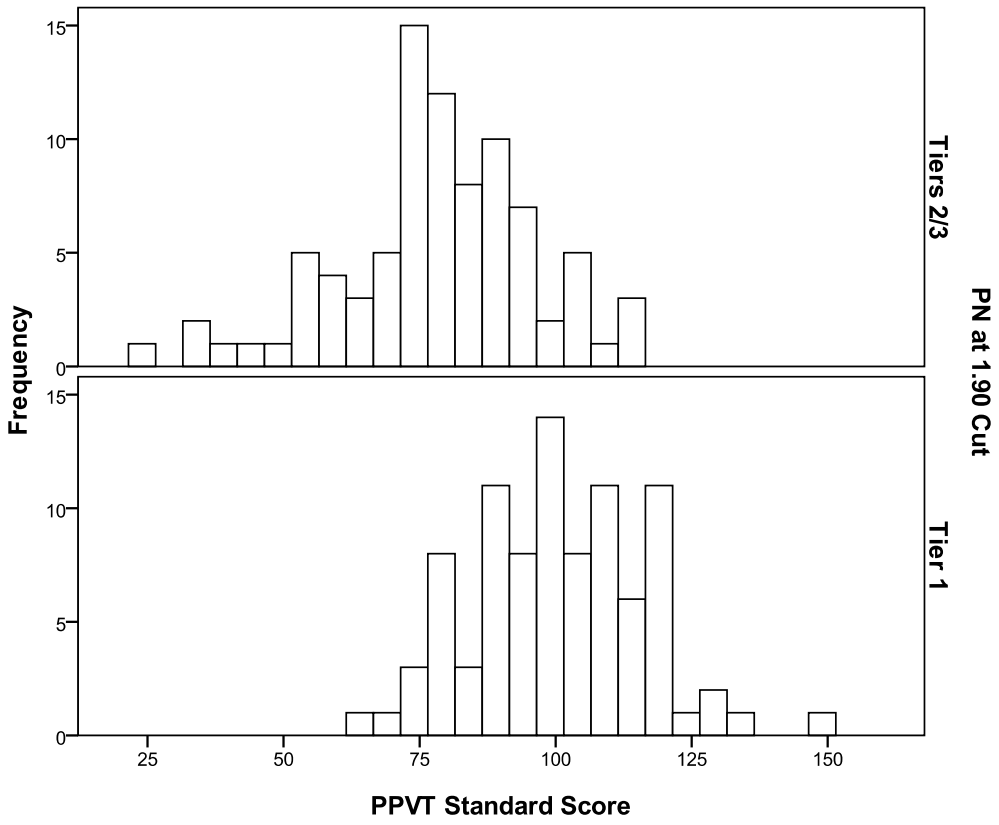


Figure 12. Score distributions of the PPVT divided by Tier placement as determined by the cut score on Picture Naming.

Here we see the distributions of scores on the PPVT scale for the two levels of Tiers based on identification with Picture Naming with a cut score set at 1.90 logits. Here we notice greater separation between the Tier distributions, when defining the Tiers based on Picture Naming cut-score performance.

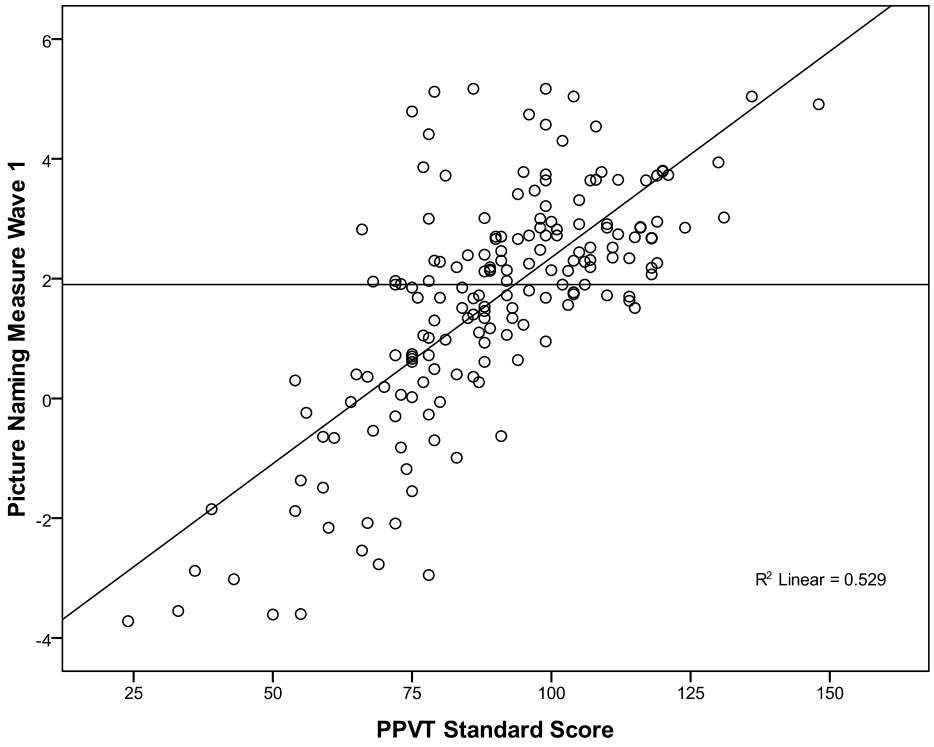


Figure 13. Scatterplot of the Picture Naming Rasch scale score (with cut score of 1.90 marked) and the PPVT standard score, and the best fitting line (R^2 = .529).

Here we can observe the association between PPVT and Picture naming, with a correlation of .73. The point of intersection between the best-fitting line (regression line) and the PN cut score is approximately 93 on the PPVT scale. The PPVT score of 93 is associated with a percentile of 32. In this sample, 56% scored below a 93 on the PPVT.

References

AERA, APA, NCME. (1999). *Standards for educational psychological testing*. Washington DC: AERA.

Cizek, G.J., & Bunch, M.B. (Eds.). (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Cizek, G.J., Rosenberg, S.L., & Koons, H.H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397-412.

Downing, S.M., & Haladyna, T.M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82.

Poggio, J.P. (1984, April). *Practical considerations when setting test standards: A look at the process used in Kansas*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved through ERIC, Document No. ED249267.

Scriven, M. (2002). Assessing six assumptions in assessment. In H.I. Braun, D.N. Jackson, & D.E. Wiley (Eds.) *The role of constructs in psychological and educational measurement* (pp. 255-275). Mahwah, NJ: Lawrence Erlbaum.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.