Center for Response to Intervention in Early Childhood Michael C. Rodriguez
University of Minnesota 2010

Building a Validity Framework for IGDIs 2.0

The validity framework for Individual Growth and Development Indicators (IGDIs, as redesigned through the work of CRtIEC, the Center for Response to Intervention in Early Childhood) is one focused on supporting the *local* meaning of scores. Validation efforts are intended to support score meaning in a way that is separate from score utility, although we recognize that score utility is important, the key to utility is meaning. IGDIs could potentially become important tools for decision making, program evaluation, and accountability, and so must meet technical quality standards. Because of the unique nature of IGDIs, the use of standard psychometric criteria may not be appropriate, but there is no reason that IGDIs must necessarily then result in information of lower technical quality. In the contexts of RTI placement and progress monitoring, the validity framework must target the immediate inferences we hope to defend in these contexts.

Wilson (2005) described an item response model approach to constructing measures based on four building blocks, including the construct map, item design, outcome space, and measurement model. The construct map takes construct definition one more step, including describing the form of the construct. To Wilson, the form, to support strong measurement, should be simple; that is, the measure should enable inferences about the amount of the trait an individual possess, from more to less or high to low. Item development should be supported by and enhance construct definition via the construct map. A helpful idea is conceptualizing items as realizations of the construct. Then the outcome space needs to be specified, defining the aspect of responses we value. The outcome space also specifies the scoring rules acknowledging construct-relevant features of responses. Finally, we need a measurement model that relates scores to constructs. In this context with IGDIs, we will use the Rasch model.

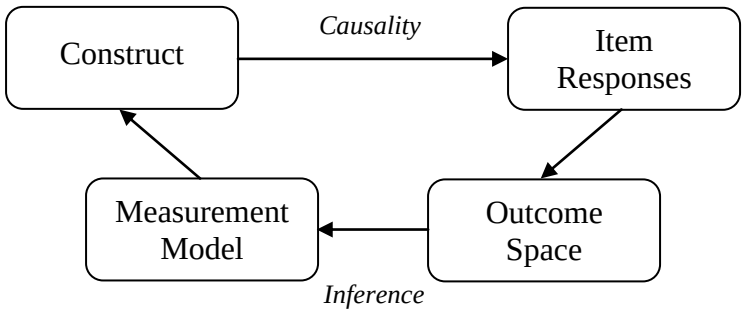


Figure 1. The four building blocks of an item response model approach to measurement construction (Wilson, 2005).

These four building blocks are more generally illustrated in Figure 1. An important element of the process is to note that the construct as it is manifested in an individual *causes* the observed item responses. The item responses constitute a sample of observations of the trait of interest (the construct). The features of the item responses that we value are present in the outcome space where we apply rules of scoring responses. These scores are then related to the construct through the application of the Rasch measurement model. The elements of the process moving from item responses to outcome space to measurement model back to the construct involve inferences; whereas the process involving the construct and item responses is a causal one.

This document will provide an evidence-based description of the elements of this measurement model, with focused attention on the construct map. It will also provide a description of several sources of validity evidence for the IGDIs, including:

I. Evidence related to domains, tasks, and intended inferences
II. Evidence related to scores
III. Evidence related to administration and scoring
IV. Evidence related to placement standards and decisions

For the purpose of this validity framework, we will focus attention on the construct map. The construct map fully describes the construct and provides a strong interpretation guide. It enables the design of tasks that will lead children to give responses that inform important levels of the construct map, by specifying relevant task features. It provides a basis to develop criterion to analyze responses and scores regarding degree of consistency with construct map. It provides the basis of establishing meaningful and appropriate administration, scoring, and the establishment of cut-scores (standards) for tier placement and progress monitoring.

Validity and Validation

Current conceptualizations of validity vary across the field; however most agree with the framework described in the *Standards for Educational and Psychological Testing* (hereafter referred to as *Testing Standards*; AERA, APA, NCME, 1999). "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, NCME, 1999). Validation is the process of gathering evidence to achieve these goals, including construct-related evidence, and evidence related to content, response processes, internal structure, relations to other variables, and the consequential bases of validity. In all cases, the most important sources of validity evidence are those that are most closely related the nature of the inferences we draw regarding scores. What evidence do we need to support the intended meaning of IGDI results?

Not all sources of validity evidence are equally important and some may not be relevant at all. Some have suggested that the consequential basis of test score use is not relevant in the process of validation (Cizek, Rosenberg, & Koons, 2008). Scriven (2002), among others, argued that issues related to test score use are important, but are more about the utility of test scores rather than the meaning of test scores. Utility of test scores are important, but do not provide evidence regarding score meaning. Our focus here will be on establishing the strongest framework possible to support score meaning.

I. Sources of Validity Evidence Related to Domains, Tasks, and Intended Inferences

Here we name and define the four domains of interest. It will be important to describe the following:

1. domain definition
2. sources of information used to define the domains and elements of the domain
3. aspects of the domain that are testable with intended age groups
4. design criteria (principles used to identify options)
5. item review criteria
6. editing criteria

Intended inferences and uses of the IGDIs include inferences to the domain and their relevance to future literacy skills and ability; dual uses in RTI settings for placement and progress monitoring.

Phonological Awareness:

Construct Definition	The ability to detect and manipulate the sound structure of words independent of their meanings (Phillips, Clancy-Menchetti, Lonigan, 2008), which develops along a continuum of complexity from identification to synthesis to analysis.
Measures of identification level:	***Rhyming*** and ***Alliteration***
Measure of synthesis level:	***Sound Blending***

Oral Language:

Construct Definition	The ability to use words to communicate ideas and thoughts and to use language as a tool to communicate to others (Dunst, Trivette, Masiello, Roper, & Robyak, 2008; Morgan & Meier, 2008). 1. Expressive language: the use of words to express meaning. 2. Receptive language: the ability to listen, process, and understand the meaning of spoken words.
Measures of expressive language:	***Picture Naming***
Measure of receptive language:	***Which One Doesn't Belong*** and ***Definitional Vocabulary***

Alphabet Knowledge:

Construct Definition	Knowledge about the names and sounds of the 26 letters of the alphabet (McBride-Chang, 1999)
Measures of aural identification:	***Sound Identification***
Measure of visual identification:	***Letter Orientation***

Comprehension:

Construct Definition	Text Comprehension: the ability to understand and interpret text as a whole (Storch & Whitehurst, 2002); it includes the "recognition of pictures and symbols in books and the ability to interpret and infer meaning from what is seen" (Dunst, Trivette, Masiello, Roper, & Robyak, 2006, p. 4). Listening Comprehension: the ability to understand and interpret spoken phonemes, words, phrases, sentences, narratives, and stories (Dickinson & Smith, 1994; Skarakis-Doyle, Dempsey, & Lee, 2008).
Measures of text comprehension:	***Picture Sequencing*** and ***Picture Comprehension***
Measure of listening comprehension:	***Sentence Comprehension*** and ***Story Comprehension***

Sources of Validity Evidence Related to Tasks

In addition, out of the work of test development more generally, there are multiple sources of validity evidence that can be collected to bear on test score meaning. Much of the work in the development of the second generation IGDIs focused on task or item design. Downing and Haladyna (1997) provided guidance for gathering item validity evidence, noting the importance of supporting score meaning at the item level. Among their advice, they suggested gathering evidence about:

a. task developer training (documentation of training materials and methods),
b. task development principles used (evidence of compliance with rules and item review processes),
c. cognitive behavior and depth of knowledge (the cognitive classification system and documentation of its use),
d. task editing procedures (including the credentials of editors and results of item review),
e. task tryout results (including item field tests, cognitive lab, and experimental test results), and
f. key validation (documentation of the process to verify that the correct responses have been identified and are included in the scoring).

II. Sources of Validity Evidence Related to Scores

Although these sources of evidence are not intended to provide a cookbook approach to validation, they do provide some utility in supporting the intended inferences, interpretations and uses of IGDI scores. Here, five common sources of evidence as defined by the *Testing Standards* are briefly described.

Construct-related evidence

- largely comes from the descriptions of process and evidence presented above (domains and intended inferences)

Content-related evidence

- again, supported from prior information presented in the domains above
- could include expert review

Response processes

- this is more difficult with younger children
- might consider qualitative review of the kinds of "talk" kids do during the administration – questions they ask or comments they make
- do children understand what they are doing??

Internal structure

- typically includes confirmatory factor analysis
- can also use evidence of score quality from the Rasch analyses

Relations to other variables

- correlations among the IGDIs and with criterion measures

III. Sources of Validity Evidence Related to Administration and Scoring

The use of the Rasch measurement model provides a number of options for administration and scoring. These options can be selected based on a number of criteria, including improved precision and efficiency, requirements for technical assumptions, and ease and interpretability. Different methods of administration provide for different scoring options. The degree to which necessary assumptions are met provides important validity evidence to support any particular method of administration and scoring. The intent is to support a system that closely resembles current IGDI practices and provides a simple and meaningful way for teachers to employ IGDIs in the classroom for important placement and monitoring purposes.

Here a number of useful considerations are addressed, relevant to the use of a measurement model and in support of meaningful interpretation and use of IGDIs 2.0.

Administration in Linear Forms

The alternative is to administer IGDIs in a linear form. Linear forms can be created in a number of ways. They can be preset bundles or the bundles can be created during administration through random selection. The advantage the Rasch analyses provides is to have item information to help create "exchangeable" bundles of items, or essentially parallel forms of predefined bundles. Administration could proceed with actual cards as in current practice, or via electronic devices such as computers or handheld devices.

Predefined Bundles for Screening/Placement

Predefined bundles could be created that contain items targeting the cut score identified through standard setting. These Bundles can be used for screening and placement because they contain tasks at the decision point, maximizing the information available for identifying the ability of students relative to the cut score. This is the most efficient approach to screening.

Scoring is based on number correct. From the Rasch analysis of items, a "test characteristics curve" can be used to identify the number correct associated with the cut score selected through standard setting. The number correct that places a student within a specific tier will have been preselected. If a student achieves a score in the predefined range, placement can be made (given the full set of criteria for placement decision making).

Predefined Bundles for Progress Monitoring

To facilitate progress monitoring within Tier 2 or Tier 3, several bundles will be needed that are balanced in terms of the range of abilities covered by the items within a bundle. This maintains the GOM aspects of the measures – they do not change in difficulty over the course of administering multiple bundles. The bundles could be created by balancing the item locations and including items around the cut score. These bundles could be administered in any order, as they are exchangeable. Items could contribute to more than one bundle.

Scoring is based on number correct. These scores are exchangeable because the bundles are designed to be exchangeable, so they are appropriate for progress monitoring. IGDIs have two features relative to number correct: (a) how many tasks can be correctly solved of those that

are presented and (b) how many tasks can be reached within the given time limit. Both features contribute to the total number correct; both features can increase or decrease the total number correct.

Randomly Selected Bundles for Progress Monitoring

If an assistive device is used for administration, items could be randomly selected for administration during progress monitoring assessments. The range of item locations could be restricted depending on the Tier in which a student is at the time, where items within that range are randomly selected by the device – administration occurs until the time limit is reached. The range of items should also include the cut score at which placement decisions occur, to facilitate interpretation of progress.

Scoring is based on number correct and follows the same reasoning as the scoring of Predefined Bundles for Progress Monitoring.

Administration in CAT Systems

Having the Rasch location parameter (item difficulty) on each item makes it possible to use the item information during administration, as in computer adaptive testing (CAT). Using CAT administration routines can maximize efficiency and precision in scores, if certain elements are in place. Typically, a test is administered in a CAT to avoid having examinees taking items that are too easy or too difficult for them – these items provide no information about their ability level. We can do a better job of estimating ability by targeting item selection – providing greater efficiency. The stopping rule for most CATs is a function of number of items, administration time, and measurement precision (standard error of measurement). The CAT administers items until these stopping criteria have been met. To facilitate this, the CAT requires a large number of items within small ranges of ability to do the targeted item selection. The efficiency and precision are only gained when testing time is long enough to maximize precision through the selection of enough items at a specific location on the ability scale.

Scoring would be based on the Rasch information based on the item responses, and would be on a logit metric, typically from -4 to +4. These scores are scalable to any scale of interest (for instance to a scale of mean of 50 and SD of 10). These scores are quite different than the typical number correct scores of original IGDIs and would require an interpretation guide for users.

With respect to IGDIs, in the tradition of the 1 to 2 minute assessment tasks, the additional technology and programming required to design the CAT would not be worth the effort given the nature of the results; CAT that reflects the current typical IGDI administration would not achieve the efficiency and precision gains obtained in full CAT routines.

Preliminary Thoughts on Administration to Achieve a Time Limit or a Number of Tasks

Traditionally, IGDIs are administered within specific time limits. The time limit is part of the construct and scoring is a direct function of the time limit. The time limit contributes to a form of standardization. In prior research (e.g., MELT results analyses), corrects for characteristics of scores that reflect time have not improved the validity of number correct scores. For instance,

Number Correct minus Number Incorrect, Number Correct divided by Total Number Attempted, and Total Number Attempted do not correlate with criterion measures as well as Number Correct. These analyses (making adjustments for the total number of tasks attempted) suggest that time is not a significant factor in the validity of total number correct scores.

There is a significant practical appeal to creating assessments that can be administered in one or two minutes, particularly with young children. The question remains, though, regarding the best design of IGDIs in terms of a fixed number of items per administration or as they are now, a fixed time limit for administration. Logic may suggest that the time limit is appropriate, because it equalizes the main feature of assessment experience across children – that being time duration of the assessment. For children who struggle more, they may attempt fewer tasks, but the duration of their struggling through tasks is limited. For children who find the tasks easier, they may attempt many more tasks, but the duration of their experience is similarly limited.

There are other models of assessment that are primarily time-restricted (most large-scale high-stakes assessments like SAT, GRE, etc.). Most K-12 tests are based on a fixed number of items, making interpretation and scalability of scores easier. In the context of predefined bundles, each bundle will be based on a fixed set of tasks, although there could be more tasks than the typical child might answer within a time limit. However, a fixed number of tasks could be bundled so that most children finish them within a set time limit. Then the administration process is to simply administer the entire bundle of tasks. This simplifies administration to some extent and standardizes the experience across children.

There are no purely psychometric criteria to define the administration context in terms of fixed tasks for fixed time. Both are used successfully in large-scale settings. Theoretically, CAT produces greater efficiency by achieving a criterion level of precision with fewer items in less time. Precision, however, is typically a function of the number of items (available information) and not the amount of testing time.

IV. Sources of Validity Evidence Related to Classification and Placement Standards

A primary requirement for RTI is predicated on the availability of information to make placement decisions. The use of IGDIs for assessing tier placement is an important goal, requiring evidence to support such use. Consistent with the *Testing Standards*, current standard setting practices were reviewed, with respect to the available evidence supporting their use and appropriateness given the unique context of early childhood classrooms and the natures of the IGDIs. The method of Contrasting Groups Design was identified as the most appropriate; see Cizek and Bunch (2007) for a complete description of this method.

Several standards speak directly to the importance of establishing cut-scores and essential elements of reporting to support such decisions.

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.

We note that current uses of the Contrasting Groups Design include the Kansas general Reading and Mathematics assessments standard setting process in 2006 and similarly in the Nebraska 2006 standard setting process. When examining a very early application of this method in Kansa, Poggio (1984) reviewed the contrasting groups method and concluded:

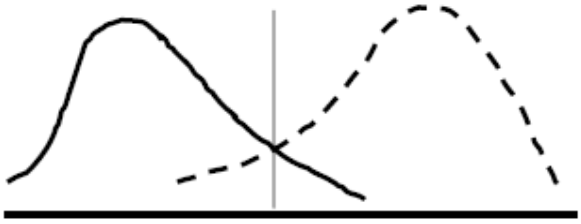
1. the method is rather easily implemented;
2. teachers report little difficulty in following what is to be done in Contrasting Groups;
3. the public is both confused and tends to doubt the legitimacy of the standard when they (often) cannot understand the "statistical magic" which delivers the standard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

The 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, included the following 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, 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. Visual inspection of the intersection of the distributions;
   b. Computation of the midpoint between tier means and modes;
   c. Estimation of logistic regression to predict the .50 probability point of placement in the next tier.

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.

Approaches to secure evidence of validity of tier placement inferences include:

- After transition out of a Tier, reassess with IGDIs.
- Confirmation among individual IGDIs; convergence of scores.
- Examine length of time in a Tier as an indication of responsiveness to intervention (appropriateness of Tier placement).

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.

Comprehension refers to a child's understanding of spoken language as demonstrated through ability to answer questions about narratives and stories and follow multi-step instructions. Question answering includes literal questions that require one to describe what is seen or recall what has occurred and inferential questions that require one to infer what has happened, what is likely to occur, what a character knows, or how someone is feeling based on understanding a story and applying background knowledge.

Advanced: Preschoolers with advanced language comprehension skills comprehend stories and narratives at a level beyond what would be expected for their age. They appropriately answer literal and inferential questions about above grade level stories and can follow complex, multi-step instructions.

Competent: Preschoolers who are competent in their language comprehension skills comprehend age appropriate stories and narratives. They appropriately answer literal and inferential questions about grade level stories and can follow two- and three-step instructions with ease.

Need Tier 2 Support: Preschoolers who need tier 2 support in developing language comprehension can respond appropriately to literal questions about simple short stories and narratives; however, they may have difficulty making inferences and predictions. Direction following tends to breakdown when instructions require more than two steps.

Need Tier 3 Support: Preschoolers who need Tier 3 support to develop language comprehension display difficulty responding appropriately to literal questions about simple short stories and narratives, especially if the answer is not visually available. They may have difficulty following simple instructions of one or two steps.

The comprehension measures were not included in this first phase of standard setting.

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.