Technical report 7: Standard setting for IGDI language and early literacy measure for 3-year-old children

Alisha Wackerle-Hollman, Ph.D.

Anthony Albano, Ph.D.

Marianne Elmquist

Kelsey Will

Scott McConnell, Ph.D.

Erin Lease, Ph.D.

Spring 2021

COI statement

Technical report 7: Standard setting for IGDI language and early literacy measure for 3-year old children

Many children experience their first educational environment at age three, when they begin preschool or other early childhood programming. It is during the third year that many children experience a structured curriculum with intentional instruction in service of identified academic domains, such as early literacy and language development. The settings for such programming vary to include formal classrooms in academic settings, such as public schools, to family, friend and neighbor settings, to Head Start and other targeted programs. Regardless of the service delivery location, early educators must be prepared to understand the skills that 3-year-old children bring to their first academic experience and meet them with instruction that facilitates learning new content.

One model used to determine what skills children bring to the classroom is the multi-tiered systems of support (MTSS) framework. MTSS is designed to facilitate an instructional match between the child's individual skill level and instructional need through screening, progress monitoring and targeted intervention. In the MTSS model, children are assessed with universal screening tools to determine intervention candidacy. Intervention dosage is then prescribed based on the child's performance level, where children in need of moderate to high levels of support receive Tier 2/3 intervention and children who are successful in the current curriculum continue to receive the Tier 1 or current practice (Carta & Miller-Young, 2018).

Screening is a critical component of MTSS models, and to adequately identify which children are candidates for intervention, the assessments used must demonstrate adequate sensitivity, specificity and predictive power (Carta & Miller-Young, 2018). Measures that accurately identify children who are in need of intervention have high instances of true positives

(++) and true negatives (--), and low instances of false positives (+-) or false negatives (-+). One set of measures specifically designed for use in MTSS models are the Age 3 IGDIs (McConnell et al., 2019). At present, few studies are available in the research literature that illustrate standard setting processes and their impact on MTSS candidacy. Here we address the process of setting performance standards, or cut scores, for the Age 3 IGDIs measures using a contrasting groups approach. As a result of these cut scores, we provide preliminary information about frequency rates of tier-level candidacy produced within a MTSS. Therefore, the purpose of this paper is to describe the standard setting process used to establish screening benchmarks with the 3-year-old IGDI measures.

Age 3 IGDIs

The Age 3 Individual Growth and Development Indicators (IGDIs) are designed to measure language and early literacy abilities in 3-year-old preschoolers (McConnell et al., 2019). Age 3 IGDIs assess three domains: oral language, phonological awareness, and alphabet knowledge. We have defined these domains and provided a comprehensive literature review on each construct in other resources (see tech reports 1-3). The three Age 3 IGDI measures (described below) are screening tools that provide data to inform decisions about instructional practices that support early literacy and language skills. However, to appropriately screen students and maximize predictive power, we must establish a threshold or performance criterion regarding adequate performance for each measure. With performance criteria identified, Age 3 IGDI scores can be used to determine intervention candidacy within MTSS or other similar frameworks.

The Age 3 IGDIs consist of three measures: Oral Language (OL), Alphabet Knowledge (AK), and Robot Blending (PA). The Robot Blending measure is abbreviated as PA because it

measures phonological awareness. The measures are untimed and administered on tablet devices in an assessor-student dyad. Each measure has 25 items and takes 2-5 minutes to administer per student. All Age 3 IGDI items were previously calibrated through a comprehensive item design process in prior work to confirm target selection and produce item-response theory difficulty values (see technical report 6).

Oral Language

The OL Age 3 IGDI measure includes two item types: Point to Picture (P2P) and Picture Naming (PN). Items for P2P are receptive and each item includes 2-3 images. Once an item's images appear on the student's iPad screen, the assessor verbally prompts the student to "point to the (target image label)." (See Figures 1 & 2). The student then touches their response on the tablet device, where the embedded code automatically scores the item as "correct" or "incorrect." If a student does not select a response, the item is scored as "no response."

PN items are expressive, with presentation of a single isolated image on the student's tablet screen. The assessor verbally prompts the student by asking "What is this?" (See Figures 3 & 4). Students' verbal responses are scored by the assessor as "correct" if the response is an exact match or a correct extension of the correct response (e.g., if a student says, "brown bear" when the correct response is "bear" and the image is of a brown bear. An answer such as "polar bear" would not be a correct extension). Students' verbal responses are scored by the assessor as "incorrect" if the response does not match the correct response or "no response" if the child does not respond to the item or says, "I don't know."

Alphabet Knowledge

The AK Age 3 IGDI measure includes four item types: Letter Find (LF), Letter Orientation (LO), Point to Letter (P2L), and Letter Naming (LN). Items for LF are receptive,

where two symbols or shapes (e.g., circle, question mark) are presented with one letter (for a total of 3 stimuli per item) (See Figures 5 & 6). Once the symbols and letter appear on the students' screen, the student is verbally prompted by the assessor to "point to the letter." After the student touches their response on the tablet device, where the embedded code automatically scores the item as "correct" or "incorrect." If a student does not select a response, the item is scored as "no response."

Items for LO are receptive, where three images of the same letter oriented in different directions are presented to the student (see Figures 7 & 8). Once the letters appear on the student's screen, they are verbally prompted by the assessor to "point to the letter that's the right way." After the student touches their response on the tablet device the embedded code automatically scores the item as "correct" or "incorrect." If a student does not select a response, the item is scored as "no response."

Items for P2L are receptive, where three letters are presented to the student (see Figures 9 & 10). Once the letters appear on the student's tablet device, the student is verbally prompted by the assessor to "point to (target letter)." After the student touches their response on the tablet device, the embedded code automatically scores the item as "correct" or "incorrect." If a student does not select a response, the item is scored as "no response."

Items for LN are expressive, where letters are presented individually on the tablet device (see Figures 11 & 12). Once the letter appears on the student's screen, student is verbally prompted by the assessor asking, "What letter is this?" The student's verbal response is scored by the assessor as "correct" or "incorrect." If the child does not respond to the item or says, "I don't know," the item is scored as "no response."

Phonological Awareness (Robot Blending)

The PA measure (Robot Blending) has one item type. The measure begins by introducing *Igdi* the robot. An image of a robot is displayed on the student's tablet device while a recorded voice introduces the robot and explains that Igdi "talks funny." Igdi speaks words in segments to prepare students for the speech format they will hear throughout the measure. For each item, the assessor begins by touching and labeling each image as it appears on the student's screen (e.g., rake, shovel, road; see Figures 13 & 14). Each item incudes 2 or 3 images, and the assessor verbally prompts the student by asking "What is Igdi trying to say?" The assessor then pressed a "play" button on their device to start a recorded clip (e.g., r/a/ke). After the student touches their response on the tablet device, the embedded code automatically scores the item as "correct" or "incorrect." If a student does not select a response, the item is scored as "no response."

IGDI Validity Claims

Validity is a process of collecting evidence to support claims or inferences made about a measure (Kane, 2013). Kane notes validity must be gathered to support an interpretation and use argument (IUA) which specifies a priori the intended ways a measure should be used and for whom the resulting scores should be interpreted. For the purposes of this report, we specified two claims for the PK3-IGDIs IUA.

(1) Age 3 IGDIs will discern between children who are candidates for Tier 1 and Tier 2/3 intervention with adequate levels of sensitivity and specificity.

(2) Age 3 IGDIs can be used to support classroom level interactions with data by examining base-rates of Age 3 IGDI scores to examine tier level candidacy and discern necessary intervention levels that the classroom would benefit from. This information will provide practical levels of expected distributions in a tiered model.

Standard Setting Procedures

To set standards to identify children who are candidates for intervention, we first established a criterion of performance that represents the underlying meaning of the cut score. There are two models in the field frequently used to establish performance standards and related cut scores: criterion-referenced standards and norm-references standards.

In norm-referend standard setting, child performance is measured against a normative distribution, where the cut score is set at an identified percentage of the population. For example, many norm-referenced standards use quartiles and set the cut score at the first quartile or 25%. This means that for all children tested, those with scores between the 1st and 25th percentile are identified as candidates for intervention, regardless of the absolute skill level of the group. Thus, in a group that is composed of children already identified at-risk, such as in Head Start programs, when a norm-referenced standard is used, those with performance below the 25th percentile would be candidates for intervention, even if those between the 26th and 99th percentile showed very low skill levels. In this way, the normative group used to establish the cut scores significantly impacts the likelihood of over or under-identification of children who may need tiered interventions. Many standardized norm-referenced tests use large samples to gather a normative group that is nationally representative to limit challenges with norm-referenced standards. However, some test designers instead focus on localized norming groups so that local populations directly inform inferences. This model can be useful when resource allocation is limited and programs are required to determine how best to devote time and energy to the children who are the most in need.

In contrast, criterion-referenced standard setting focuses on an absolute standard. In this model child performance is measured against an identified criterion that represents mastery of a specific skill set. For example, if the underlying domain is early literacy skills, and content

experts, existing studies, and theory align to suggest children should be able to identify at least 75% of uppercase letters by the end of their third year, then the criterion established for letter names would constitute an identified set of letters that represents most of the alphabet. In this way, any child who is unable to correctly identify 75%, or 20 letters, would be identified as in need of intervention, regardless of how many children are included in the pool of candidates. This model can be useful to examine mastery of a skill set based on an absolute standard but can also be challenging when resource allocation is limited. This is because it is difficult to determine how best to allocate intervention services when there are more children who are intervention candidates than resources to support them.

The IGDI model employs a criterion-referenced standard setting model to gather information on to what degree children have met absolute standards of language and early literacy performance in the preschool year two years before kindergarten. To set the criterion-referenced standard, the model requires an anchoring variable that defines a "true positive." Typically, an anchoring variable would be modeled from predictive validity studies by using longitudinal data to define what score or criterion is necessary on the measure to obtain probability of success on a longitudinal outcome. However, when new measures are constructed, predictive validity models are frequently not yet available, and preliminary data must be used to establish criterion-referenced standards.

One approach to preliminary standard setting is the use of contrasting groups design. In contrasting groups design two independent variables are identified and supporting data are collected in service of the same outcome or dependent variable. The independent variable data are contrasted to find the location that best maximizes the information from each data source. The Age 3 IGDIs model used IGDI scores and teacher use of performance level descriptors

(PLDs) as the independent variables in the contrasting groups design to examine to what degree teachers identified children as intervention candidates for each domain based on performance level descriptor surveys, and compared those ratings to IGDI scores (thus tier level candidacy is the dependent variable).

Method

Sample

Students and teachers were recruited from Local Educational Agencies (LEA) and a Head Start center in Minnesota as well as a charter school in Washington, D.C. We obtained passive parental consent for 479 potential student participants across 35 classrooms and 12 schools (356 students in Minnesota and 123 students in Washington, D.C. participated). At the start of the fall timepoint there were a total of 465 students enrolled and consented in participating classrooms (346 in Minnesota and 119 in Washington, D.C.). Between recruitment and the start of the fall assessments 10 students were removed from the study due to the following reasons: parents opted student out of the study, student moved or disenrolled from the program, or a teacher requested the student be removed from the study. Demographic information for the participating children came from school administration records. Teacher and classroom information came from teacher surveys.

Teachers

A total of 20 teachers participated in the current study. Most teachers were recruited from our Minnesota sites; a) suburban school districts (n = 10), b) urban Head Start program (n = 3),

and urban school district (*n* = 1). The remaining six teachers worked in a Charter school in Washington, D.C.

Teachers had a mean of 12.2 years of teaching experience (range 1-37 years) and teaching assistants had an average of 12.45 years of experience (see Table 1 for additional teaching experience). Over half of participating teachers (*n* = 12) had a Bachelor's degree and eight had a Master's degree. Three teachers had a Child Development Associate. Table 2 provides information on the majors of teacher's degrees. Teaching assistant's highest educational level included: Bachelor's degree (*n* = 8), Associates degree (*n* = 5) and some college or high school (*n* = 3). Table 3 describes the educational backgrounds of the teaching assistants.

Children

There were 451 students that participated in the current study. Table 4 describes student demographics, this information was obtained from school administrative records. Students had a mean age of 3;6 years (SD = 0.30; range = 3;0-4;4; missing data *n* = 2). The sample included 198 boys and 221 girls (missing data *n* = 32). English was the home language for the majority of children (*n* = 378), other home language included: Bosnia, Cantonese, Enbosh, English & Oromo, Hausa, Hindi, Hmong, Japanese, Mandarin, Marathi, Nepali, Oromo, Russian, Somali, Spanish, Telugu, Tibetan, Vietnamese, Wolof, and Yoruba (see table 4 for n's). The sample included 165 Black and 178 White children, see Table 4 for a full breakdown of child race/ethnicity. Around 16% of children received free and reduced lunch, however it is important to note that this information was unavailable for 38.56% of students. About 12% of students had an Individualized Education Plan (IEP).

Setting

Participating classrooms varied from full-day Monday–Friday sessions, to half-day, or split week session types. The majority of classrooms were providing a School Readiness program (n = 8). Most teachers delivered instruction in English (n = 18) and one teacher delivered instruction in English and Spanish. Teachers had an average of six dual language learners (DLL) in their participating classrooms. Eight participating teachers used IGDIs in their classroom prior to the current study (note that teachers who had prior experience using IGDIs would have used the IGDIs measures for 4-year-olds rather than the new Age 3 IGDIs). Published early literacy and language curricula were used by 18 teachers in their classrooms, with 11 using Creative Curriculum for Preschoolers by Teaching Strategies (See Table 5 for a full description of curricula used).

Measures

Performance Level Descriptors

To collect teacher judgements on participating student's literacy development, the researchers developed the performance level descriptor (PLD) survey (see Figure 15). Teachers, with help from teaching assistants or other school/program staff, rated each participating student's skill level in OL, AK, and PA based on operational definitions for each domain. After reviewing each domain definition, teachers and their assistants collaboratively rated each child's performance on three tiers; a) Tier 1: child has little or no difficulty and needs no special intervention, b) Tier 2: child has moderate difficulty and needs supplemental intervention, c) Tier 3: child has significant difficulty and needs intensive intervention.

Teacher Classroom Survey

To collect information on participating students' classrooms, the researchers utilized IGDIlab's Teacher Classroom Survey. The survey provided information regarding teachers' and

teaching assistants' teaching experience and educational background as well as type of program, class size, schedule for when students attended the classrooms, adult to student ratio, language of instruction, number of dual language learners (DLLs) and use of literacy curricula.

Age 3 IGDIs

The Age 3 IGDIs were used to measure student performance on three early literacy domains; OL, AK, and PA, see the introduction for detailed information on each domain. Assessments were given during three timepoints: fall, winter, and spring.

Procedures

Age 3 IGDIs

Training and Fidelity. Graduate research assistants trained assessors to administer all Age 3 IGDIs (i.e. OL, AK, and PA) during a single three-hour training session. The training session was a combination of PowerPoint lecture describing the project and measures, administration video models, and practice opportunities. At the end of the training sessions assessors demonstrated fidelity of administration on each measure.

Administration and Scoring. Students were assessed by trained assessors during the regular school day. Assessments took place in hallways or empty rooms near students' classrooms. When available, tablets were set up on desks or tables, otherwise assessments were completed on the floor. The OL measure was always given first to prevent children from learning some labels of images in other tasks that might be presented in the OL Picture Naming task.

Performance Level Descriptors

Teachers completed PLDs once during each data collection season. Surveys were emailed to teachers and they were asked to complete the survey within two weeks of the assessment window. Teachers completed the PLDs without knowledge of IGDI scores to prevent prior

knowledge of scores from contaminating their ratings. For each IGDI domain (i.e. OL, AK, and PA) teachers were asked to mark the response that most accurately described the student's skills based on the three tier definitions. Teachers were told they could consult with teaching assistants or school/program staff familiar with the student and their performance for the three domains.

Teacher Classroom Survey

The teacher classroom survey was included in the Spring PLD survey. Teachers were asked to complete questions on teacher and classroom characteristics.

Analysis

To produce cut scores for screening, we employed a contrasting groups design to identify Age 3 IGDI scores that optimally differentiate children in need of intervention (Tier 2/3 candidates) from those children who are successful in the current curriculum (Tier 1 candidates). Analyses were based on receiver-operating characteristic curves (ROC) which compared the proportions of children judged by teachers as having some need for instructional intervention (Tier 2/3) vs not (Tier 1), across the ability scale and IGDI scores. ROC analysis provides statistical indices for evaluating classification into groups. ROC analysis was used here to compare the proportions of children within the two overlapping curves across a series of theta values, so as to inform the decision of a cut score that differentiates best between the two curves. In the absence of any practical guidance, the optimal cut score is often defined as the one that separates the curves as much as possible based on a statistic called Youden's *J*, which is calculated as

$$J = \text{sensitivity} + \text{specificity} - 1.$$

When using only *J* to select a cut score, no constraints are placed on the sensitivity and specificity, and one of the two values can be significantly lower than the other while still

producing the largest *J*. Contrasting groups analyses from previous IGDI development projects required a target sensitivity of at least 0.70. This guideline was combined here with consideration of *J*. ROC results were first subset to values with sensitivity of 0.70 or larger, and then the cut score was chosen to have maximum *J*, thereby maximizing specificity within the subset of theta. We also examined the area under the curve or AUC, as described below, from ROC analysis. Once we obtained optimized cut score values, we examined the base rates of Age 3 IGDIs scores above (Tier 1), below (Tier 2/3), and within a standard error of measurement (Tier M) for each Age 3 IGDI measure.

Results

We calculated cut cores for each of the Age 3 IGDI measures using the contrasting group method for each domain.

Oral Language

To compute cut scores, we first examined the distributions of children's performance based on teacher ratings using the PLDs for each season. Figures 16-18 depict children's performance by teacher PLD tier level candidacy (Tier 3 vs Tier 1) for the fall, winter, and spring administrations. We used visual overlap in the distributions to evaluate the effectiveness of ability, based on theta scores, as a predictor of tier classification. In fall and winter, plots depict Tier 2/3 and Tier 1 to have significant distinction, while the plot for spring overlaps significantly.

We examined ROC based on seasonal administrations. Resulting ROC curves for each season are provided in Figures 19-21. Large deviations of each bold ROC line from the gray diagonal line indicate larger area under the curve (AUC) and more effective classification based on ability. In contrast, as the ROC line gets closer to the diagonal line overall, the AUC

approaches 0. When interpreting AUC as an overall index of the effectiveness of ability as a predictor of tier classification, values above 0 indicate that classification is better than chance. Ideally, AUC will be 0.70 or higher. Lower values indicate theta is a less effective predictor of tier level, and additional research and development may be needed. These plots confirm visually that classification was strongest in the fall and winter, with smaller AUC in the spring.

Cut scores were established for oral language as described above, first by targeting a sensitivity level of at least 0.70, and then selecting the ability level that maximized *J*. The resulting cut scores were -0.27, 0.34, and 0.67 for fall, winter, and spring, respectively. Sensitivity, specificity, AUC, and *J* corresponding to these cut scores are provided in Table 6. AUC was above 0.70 for fall and winter administrations, but slightly below for spring at 0.696.

Alphabet Knowledge

Parallel to oral language, we computed cut scores for alphabet knowledge by examining distributions of the teacher PLD performance ratings for each seasonal screening. Figures 22-24 depict the ability distributions for children by tier level candidacy (Tier 3 vs Tier 1) for the fall, winter, and spring administrations. In fall, winter, and spring plots depicted Tier 2/3 and Tier 1 with significant distinction.

Results for alphabet knowledge are presented in Table 7. Fall, winter, and spring administrations yielded cut scores of 0.02, 0.37, and 0.93, respectively. AUC were all above 0.70. AUC curves are depicted in Figures 25-27.

Phonological Awareness

We computed the cut scores for phonological awareness, yielding cut scores of 0.40, 0.16, and 0.40 for each season, respectively. When the phonological awareness measure was included in the contrasting groups analyses and the target sensitivity was established, the

specificity values were lower than desired, as noted in Table 8. Figures 28-30 depict score distributions. AUC curves are depicted in Figures 31-33. With low specificity and AUC all below 0.70, cut scores are not recommended for use with phonological awareness at this time.

Summarized cut scores

We used the contrasting groups results to inform final cut score selection, which incorporated sensitivity and specificity values from each domain analyses, as well as the Rasch standard error of ability estimates. We calculated the average standard error by measure by taking the root mean square of the reported standard errors SE, labeled RMSE, across all test takers as

$$RMSE = \sqrt{\frac{SE^2}{N}}.$$

Table 9 depicts the RMSE per measure, original cut scores, and final cut score ranges that account for RMSE. We calculated the final cut score range as the original cut +/- RMSE for each measure. This range was used to define an intermediate tier, labeled Tier M, wherein additional information would be needed to determine whether a child's performance should be categorized into Tier 1 or Tier 2/ 3. The remainder of the table below shows percentages of children in each tier based on their ability estimates.

Base Rates

With cut scores identified, we examined the base rates of each measure by Tier level and data source. First, we examined PLD base rates, followed by base rates using the identified cut scores.

Table 10 shows base rate frequencies and mean thetas for each domain, season, measure, and tier status levels based on PLD responses. Teachers who rated children's performance as

needing support or intervention are noted as Tier 2/3, and teachers who rated children's performance as appropriate given current classroom instruction, are noted as Tier 1. In addition, children's average scores (mean thetas) represent the average ability score for each measure, season, and tier. Finally, we reported the mean difference between the average score between Tier 1 and Tier 2/3. Given the design of the cut scores, we expected the difference to be positive indicating children with performance classified as Tier 1 had a higher mean score than children with performance classified as Tier 2/3.

Results indicate that for oral language, the majority of children's performance was rated as Tier 1, with percentages of children's performance in Tier 2/3 ranging from 41% to 46%. For alphabet knowledge, the performance of fewer children was rated as Tier 1 than Tier 2/3 in the fall and winter, but percentages of performance had nearly balanced out by spring. Finally, for phonological awareness, children's performance was consistently larger in Tier 2/3.

Following PLD base rates, we examined the percentage of children who had performance characteristic in each Tier based on the IGDI scores relative to cut score values (Table 11). Results indicated the proportion of children with performance characteristic of Tier 1 tended to remain stable or improve across seasons, from fall to winter and from winter to spring. For alphabet knowledge, performance characteristic of Tier 2/3 went down from fall, to winter, to spring. For oral language and phonological awareness, Tier 2/3 proportions went down from fall to winter, but appeared to maintain or show a minor increase from winter to spring.

Discussion

This report described the standard setting process for the PK3-IGDIs and related results regarding sensitivity, specificity, AUC and base rates. This report focused on supporting two validity claims:

(1) Age 3 IGDIs will discern between children who are candidates for Tier 1 and Tier 2/3 intervention with adequate levels of sensitivity and specificity.

(2) Age 3 IGDIs can be used to support classroom level interactions with data by examining base-rates of Age 3 IGDI scores to examine tier level candidacy and discern necessary intervention levels that the classroom would benefit from. This information will provide practical levels of expected distributions in a tiered model.

To address the first claim we used a contrasting groups design to produce empirical cut scores that maximized the differentiation between performance levels characteristic of Tier 1 and Tier 2/3 based on Age 3 IGDI scores and teacher PLDs. Results indicated that for oral language and alphabet knowledge, we can reasonably discern between Tier 1 and Tier 2/3 performance. More research is needed with phonological awareness measures, where specificity was lower and AUC were under 0.70. Based on these results, we recommend delaying the use of the phonological awareness measures until additional analyses are available.

To address the second validity claim we examined the base rates of student performance classified in Tier 1 and Tier 2/3. Pragmatic constraints limit the percentage of student performance that can be classified in Tier 2/3 because early childhood classrooms are equipped with limited resources. If the Tier 2/3 categorization yields a large percentage of students, teachers will be forced to allocate resources to some students, while others will not benefit from the same level of intervention, even though they are flagged for intervention. As such, it is important to balance the empirical cut score with practical information on base rates. If high base rates are established, the validity claims would be violated and the intention and use of the measure should be revisited.

Our results indicated that the base rates across each measure using the established cut scores were high, but rarely exceeded 50%. By spring, the oral language measures achieved 36% at Tier 2/3, alphabet knowledge achieved 43%, and phonological awareness achieved 46%. These rates are absent explicit intervention and have the potential for significant reduction in a model where appropriate intervention is put in place to meet the students' needs. These data suggest that there are many children in the age 3 preschool year who can benefit from language and early literacy intervention and classroom teachers may need to evaluate existing Tier-1 practices designed to support these emerging skills.

Conclusion

Understanding 3-year-old preschooler's early literacy and language performance is an important component of ensuring they are on track for later reading success. To evaluate to what degree children are on track for later reading success, milestones, or cut scores, are necessary to ensure appropriate progress is made. This report examined the Age 3 IGDI measure's validity in setting cut scores to examine if students are achieving appropriate early literacy and language milestones (Tier 1), if they need more intensive support (Tier 2/3), or if more information is needed (Tier M). Our results indicated the Age 3 IGDIs could meaningfully differentiate performance in the OL and AK domains for fall, winter, and spring, as well as in the PA domain for the winter and spring of the academic year.

References

Carta, J. J., & Miller-Young, R. M. (2018). *Multi-Tiered Systems of Support for Young Children: Driving Change in Early Education*. Brookes Publishing.

Kane, M. (2013). Articulating a validity argument. In *The Routledge handbook of language testing*. Routledge.

McConnell, S., Lease, E., Wackerle-Hollman, A., Albano, A., Schardt, A., Schuster, K., & Cai, L. (2018). Testing initial items for alphabet knowledge, phonological awareness, and oral language for three-year-old children: Report of phase 2 results. Technical Report #6, IGDILab, University of Minnesota.

McConnell, S., Wackerle-Hollman, A., Albano, A., Lease, E., Will, K., & Elmquist, M. (2019). Age 3 Individual Growth & Development Indicators (IGDIs). Unpublished measure.

Schardt, A., Brandes, D., Wackerle-Hollman, A., & McConnell, S. (2017). Defining a language and early literacy domain for assessment of three-year-olds: Alphabet knowledge. Technical Report #2, IGDILab, University of Minnesota.

Schardt, A., Schuster, K., Brandes, D., Wackerle-Hollman, A., & McConnell, S. (2017). Defining a language and early literacy domain for assessment of three-year-olds: Phonological awareness. Technical Report #1, IGDILab, University of Minnesota.

Schuster, K., Cai, L., McConnell, S., & Wackerle-Hollman, A. (2017). Defining a domain of assessment for language and early literacy in three-year-olds: Oral language. Technical Report #3, IGDILab, University of Minnesota.

Table 1

Teaching Experience of Teachers and Teaching Assistants as Reported From Teacher Classroom Survey (N =20)

Survey Question	Years
Mean (*SD*) number of years teaching – Teacher	12.2 (10.43) range: 1-37
Mean (*SD*) number of years teaching – TA	12.45 (9.82) range; 0-30
Early childhood program experience - Teachers	*n*
≤ 5 years	5
6-10 years	6
11-19 years	4
≥ 20 years	5

Note. SD = Standard deviation. TA = Teaching Assistant.

Table 2

Educational Background of Participating Teachers as Reported From Teacher Classroom Survey (n =20)

Survey question	*n*
Highest education level	
Associate degree	0
Bachelor's degree	12
Master's degree	8
Bachelor's degree major	
Early childhood	2
Early childhood and family education	1
Early childhood education	2
Early childhood studies	1
Elementary education w/ PreK endorsement	1
Elementary education, early childhood, reading	1
Elementary education	1
MN teaching degree Elementary Ed (k-6) with early childhood minor (birth -k)	1
K-8 Education	1
Pre-k – grade 6	1
Master's degree major	
Early childhood education	2
Educational leadership	1
Special Education	1
Elementary education	1
Early Childhood Education Early Childhood Education + licensure	1
Parent and family education	1
Not reported	1
Have you completed a CDA?	
Yes; No	3;16
Not reported	1
Are you currently working on a CDA?	
No	16
Not reported	1

Note. CDA = Child Development Associate.

Table 3

Educational Background of Participating Teaching Assistants as Reported From Teacher Classroom survey (n =20)

Survey Question	*n*
Highest education level	
Associate Degree	5
Bachelor's Degree	8
High school or equivalent	2
Some college	1
Not reported	4
Associate Degree – major	
Associates of arts	1
Early childhood	2
Not reported	17
Bachelor's degree - major	
Child psychology	1
Education	2
Elementary education (K-6)	1
Journalism	1
Social work	2
K-6 teacher	1

Table 4

Student Demographics (n = 451)

Variable	*n* (%)	*Missing n (%)*
Sex		32 (7.10)
Male	198 (43.90)	
Female	221 (49)	
Race/Ethnicity		35 (7.76)
Asian Pacific	35 (7.76)	
Asian Pacific & White	8 (1.78)	
Asian Pacific, White & Native American	1 (0.22)	
Black	165 (36.59)	
Black, Asian Pacific, & White	1 (0.22)	
Black & Latino	1 (0.22)	
Black & Native American	2 (0.44)	
Black & White	4 (0.89)	
Latino	9 (2.00)	
Native American	3 (0.66)	
White	178 (39.47)	
White & Latino	8 (1.78)	
White & Native American	1 (0.22)	
Home Language		36 (7.98)
Bosnian	1 (0.22)	
Cantonese	1 (0.22)	
Enbosh	3 (0.67)	
English	378 (83.81)	
English & Oromo	1 (0.22)	
Hausa	1 (0.22)	
Hindi	1 (0.22)	
Hmong	4 (0.89)	
Japanese	1 (0.22)	
Mandarin	1 (0.22)	
Marathi	1 (0.22)	
Nepali	1 (0.22)	
Oromo	1 (0.22)	
Russian	1 (0.22)	
Somali	5 (1.12)	
Spanish	6 (1.33)	
Telugu	1 (0.22)	
Tibetan	1 (0.22)	
Vietnamese	3 (0.67)	
Wolof	2 (0.22)	
Yoruba	1 (0.22)	
Child has an IEP on record	55 (12.20)	29 (6.43)
Child receives free & reduced priced lunch	71 (15.74)	173 (38.36)

Note. IEP = Individualized education plan.

Table 5

Description of Participating Classrooms as Reported from Teacher Classroom Survey (N =20)

Survey Question	Classroom *n* (unless noted otherwise)
Type of program [a]	
Head Start	3
Private preschool	0
ECFE	5
School readiness	8
State funded preschool	5
Public Charter school	3
Number of DLL students – Mean (SD) range	6.13 (8) 0-30
Language of instruction	
English	18
English & Spanish	1
Not Reported	1
Used IGDIs before	
Yes; No	8; 12
Early literacy & language curriculum use	
Using a curriculum	
Yes; no	18; 2
Published curriculum	
Yes; No	18; 0
Name & Publisher of Curriculum [b]	
Creative Curriculum for Preschoolers – Teaching Strategies	11
Big Day in PreK - Scholastic	5
Handwriting without tears – Learning without Tears	3
Jolly Phonics	1
eZwrite	1
Incredible Years	1
Scholastic	1
Gold – teaching strategies	1
Days per week child is exposed to curriculum	
Daily	11
4 days	2
3 days	2
2 days	3

Note. [a] Some classrooms fall into multiple program types so *n* is greater than 20. [b] Some classrooms used multiple curricula, so *n* is greater than 20. SD = standard deviation. DLL = Dual language learner

Table 6

Oral Language ROC with Updated Sensitivity Criteria

Season	Sensitivity	Specificity	Cut	*J*	AUC
Fall	0.756	0.603	-0.271	0.359	0.756
Winter	0.752	0.545	0.339	0.297	0.703
Spring	0.710	0.603	0.666	0.313	0.696

Note. Updated sensitivity criteria targeted at 0.70.

Table 7

Alphabet Knowledge ROC Results with Targeted Sensitivity

Season	Sensitivity	Specificity	Cut	*J*	AUC
Fall	0.738	0.616	0.015	0.354	0.733
Winter	0.700	0.759	0.371	0.459	0.781
Spring	0.741	0.591	0.926	0.332	0.712

Table 8

Phonological Awareness ROC Results with Targeted Sensitivity

Season	Sensitivity	Specificity	Cut	*J*	AUC
Fall	0.896	0.250	0.397	0.146	0.597
Winter	0.740	0.586	0.160	0.326	0.694
Spring	0.726	0.571	0.394	0.297	0.663

Table 9

Final Cut Scores for Each Domain, Measure, and Season

Measure	Season	Original cut	RMSE	Final cut: (low boundary)	Final cut: (high boundary)
OL	Fall	-0.27	0.45	-0.72	0.18
OL	Winter	0.34	0.45	-0.11	0.79
OL	Spring	0.67	0.45	0.22	1.11
AK	Fall	0.02	0.41	-0.39	0.42
AK	Winter	0.37	0.41	-0.04	0.78
AK	Spring	0.93	0.41	0.52	1.33
PA	Fall	0.40	0.43	-0.03	0.82
PA	Winter	0.16	0.43	-0.27	0.59
PA	Spring	0.39	0.43	-0.03	0.82

Note. OL= oral language, AK= alphabet knowledge, PA= phonological awareness, RMSE= root mean square of the reported standard errors.

Table 10

PLD Counts, Mean Theta, and Proportion by Season and Measure

Measure	Season	Tier 1 *n*	Tier 2/3 *n*	Tier 1 %	Tier 2/3 %	Tier 1 *M*	Tier 2/3 *M*	*M* diff
OL	Fall	234	164	0.59	0.41	-0.10	-0.87	0.77
OL	Winter	233	153	0.60	0.40	0.44	-0.27	0.71
OL	Spring	219	183	0.54	0.46	0.81	0.10	0.71
AK	Fall	112	279	0.29	0.71	0.47	-0.56	1.03
AK	Winter	116	270	0.30	0.70	1.17	-0.19	1.37
AK	Spring	198	205	0.49	0.51	1.07	0.14	0.93
PA	Fall	124	259	0.32	0.68	-0.21	-0.50	0.30
PA	Winter	145	231	0.39	0.61	0.36	-0.28	0.64
PA	Spring	177	219	0.45	0.55	0.49	-0.07	0.57

Note. OL= oral language, AK= alphabet knowledge, PA= phonological awareness. *M*= mean theta.

Table 11

Proportion of Children in Each Tier Relative to Cut Score Values Based on PK-3-IGDI Scores

Measure	Season	Tier 2/3 prop.	Tier M prop.	Tier 1 prop.
OL	Fall	0.37	0.32	0.31
OL	Winter	0.34	0.38	0.27
OL	Spring	0.36	0.34	0.30
AK	Fall	0.53	0.21	0.26
AK	Winter	0.44	0.22	0.34
AK	Spring	0.43	0.26	0.31
PA	Fall	0.67	0.26	0.07
PA	Winter	0.45	0.28	0.27
PA	Spring	0.46	0.27	0.27

Note. OL= oral language, AK= alphabet knowledge, PA= phonological awareness, prop.= proportion.

Figure 1

Point to Picture Student Screen

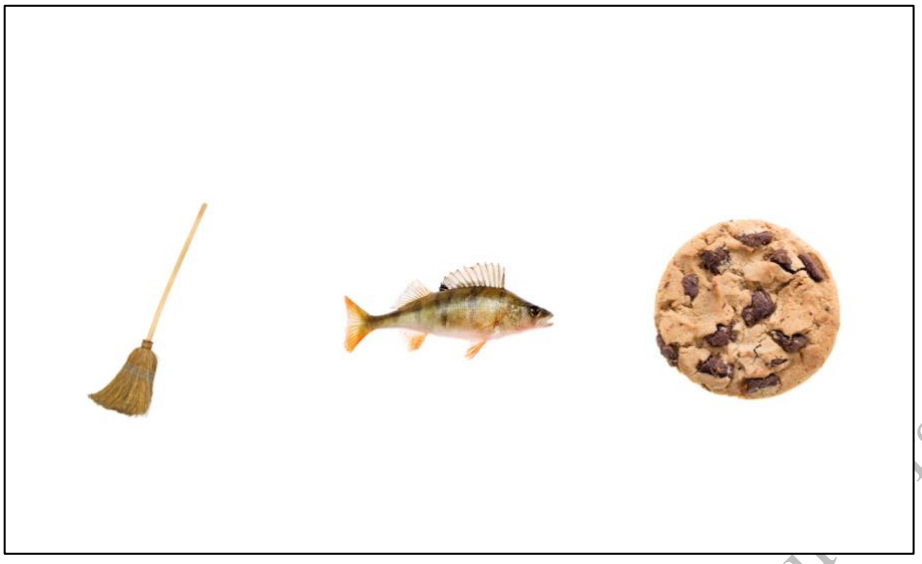

Figure 2

Point to Picture Assessor Screen

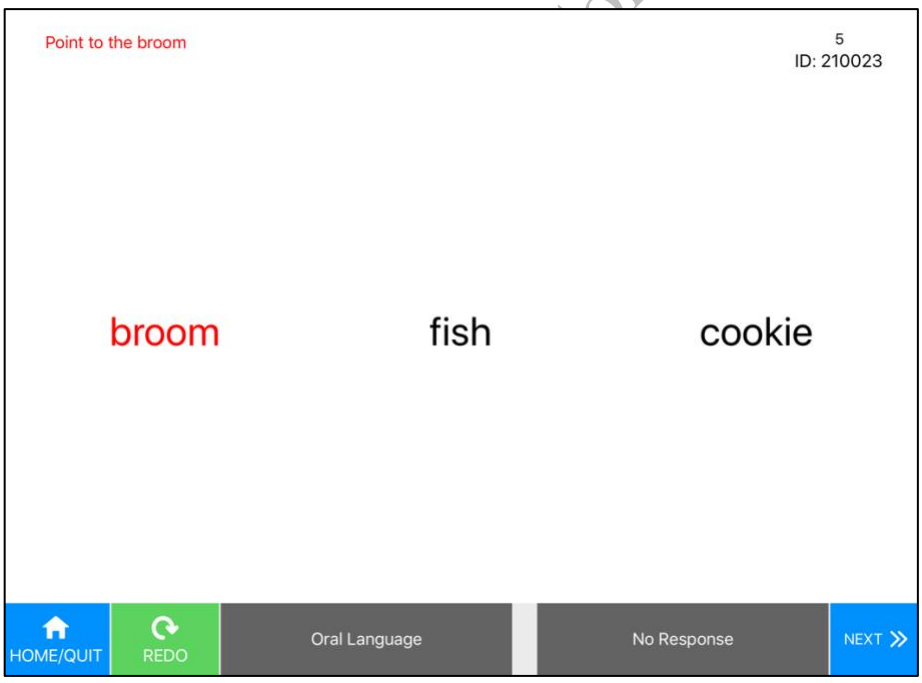

Figure 3

Picture Naming Student Screen

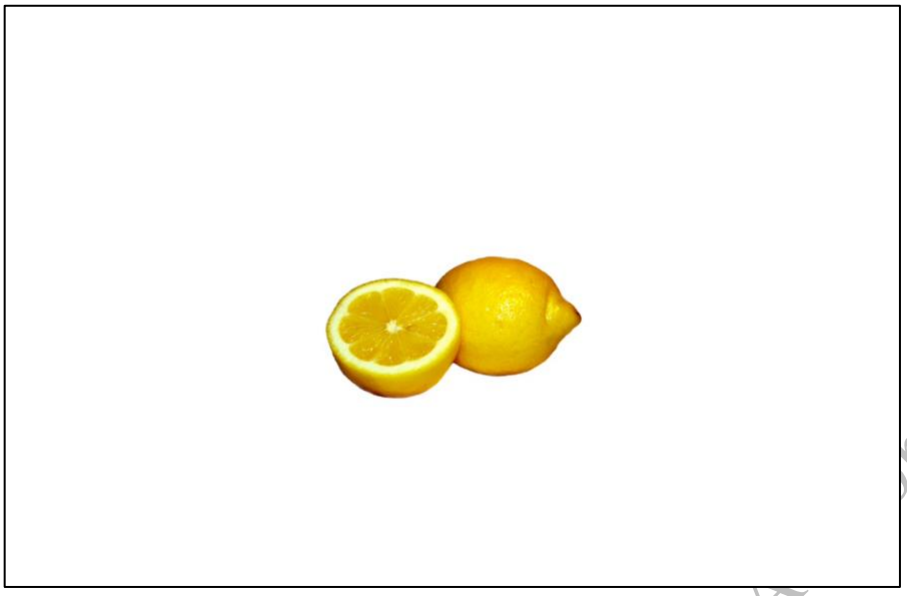

Figure 4

Picture Naming Assessor Screen

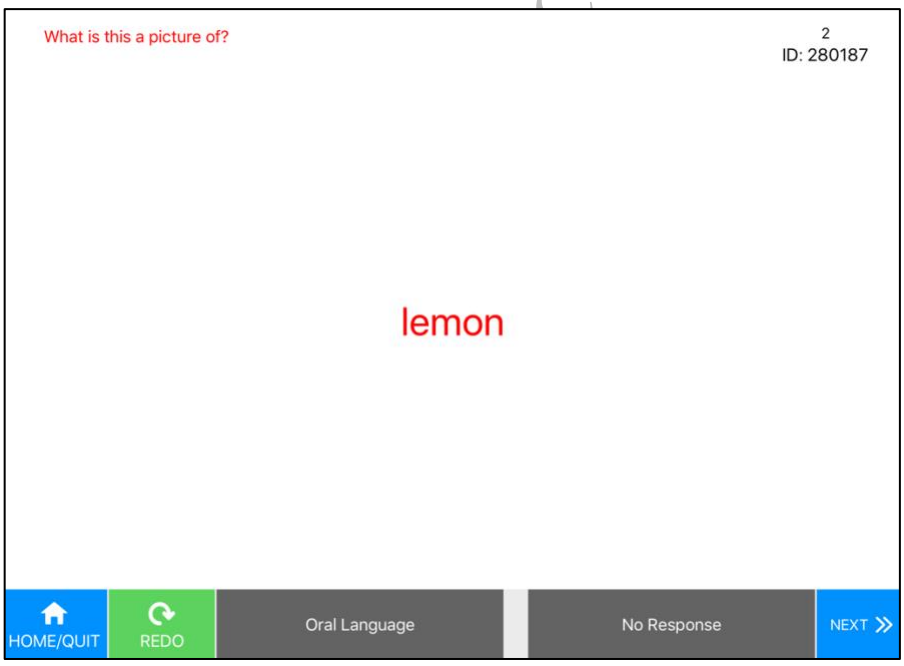

Figure 5

Letter Find Student Screen

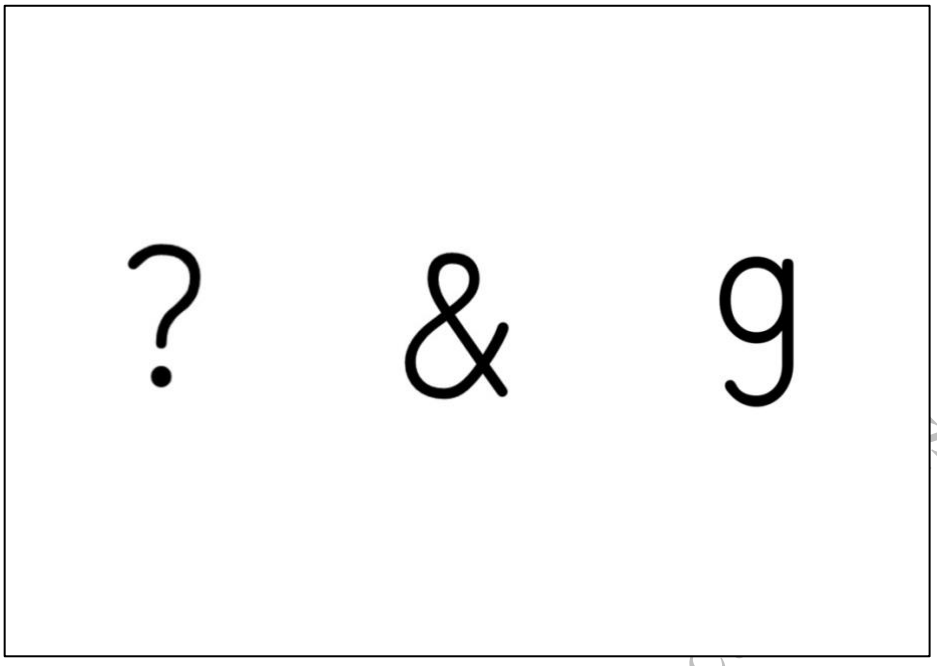

Figure 6

Letter Find Assessor Screen

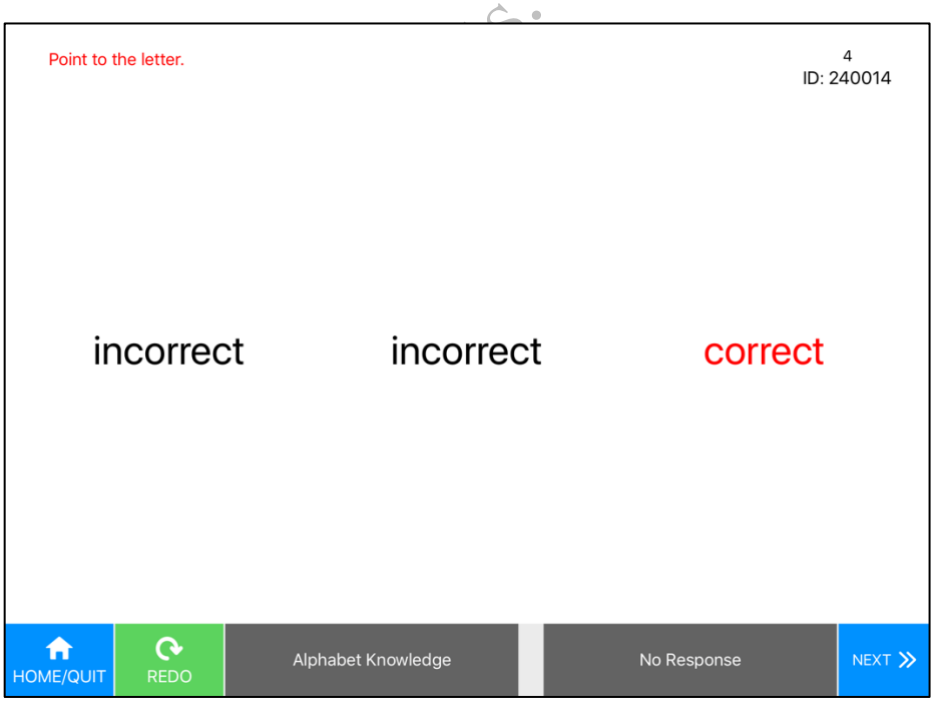

Figure 7

Letter Orientation Student Screen

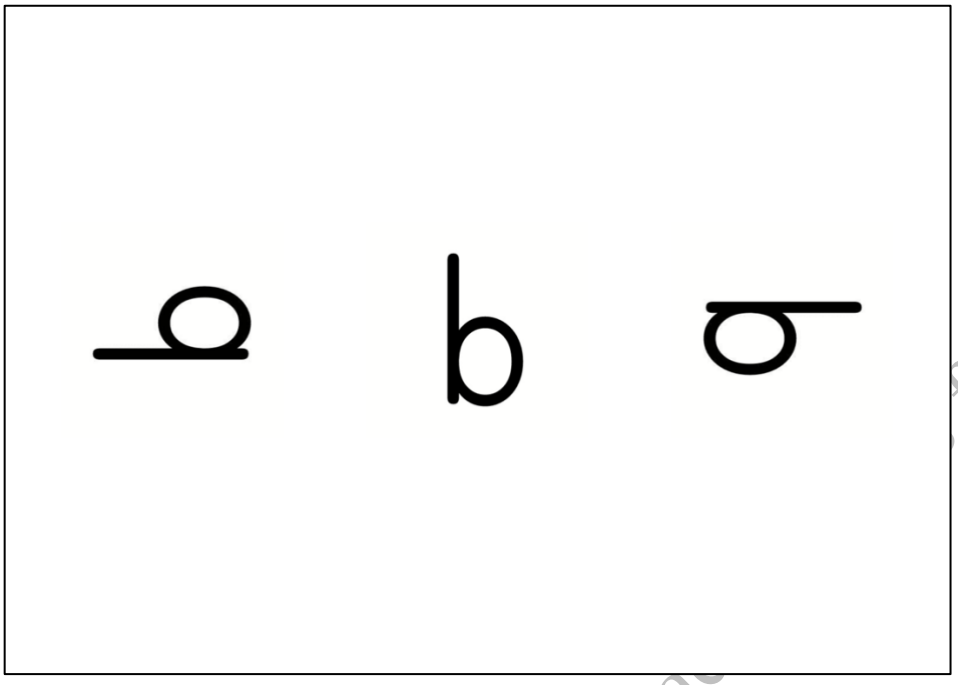

Figure 8

Letter Orientation Assessor Screen

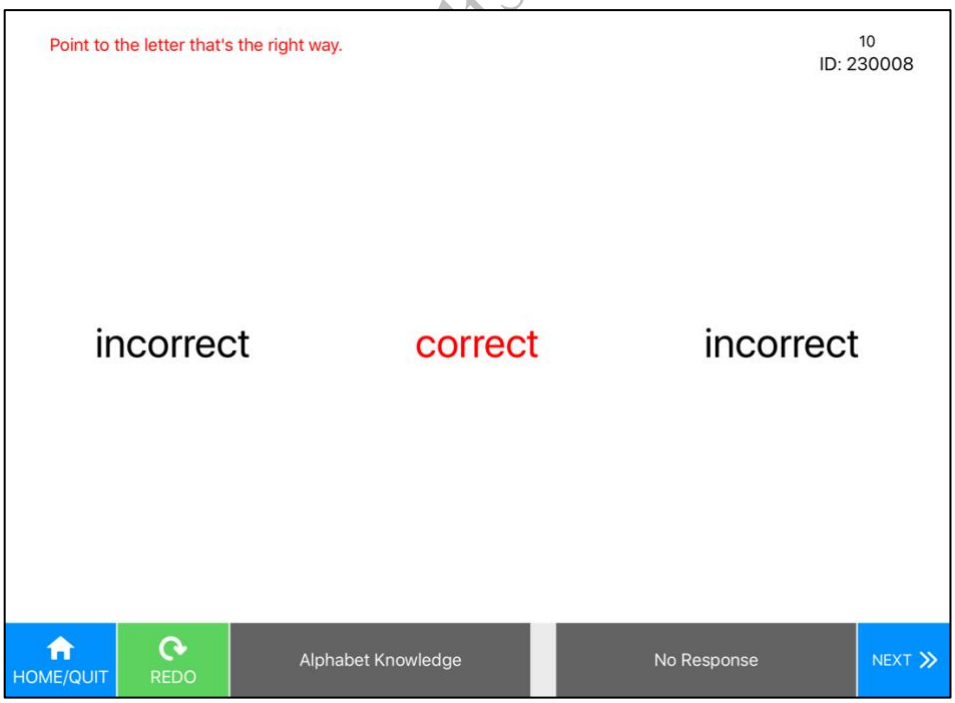

Figure 9

Point to Letter Student Screen

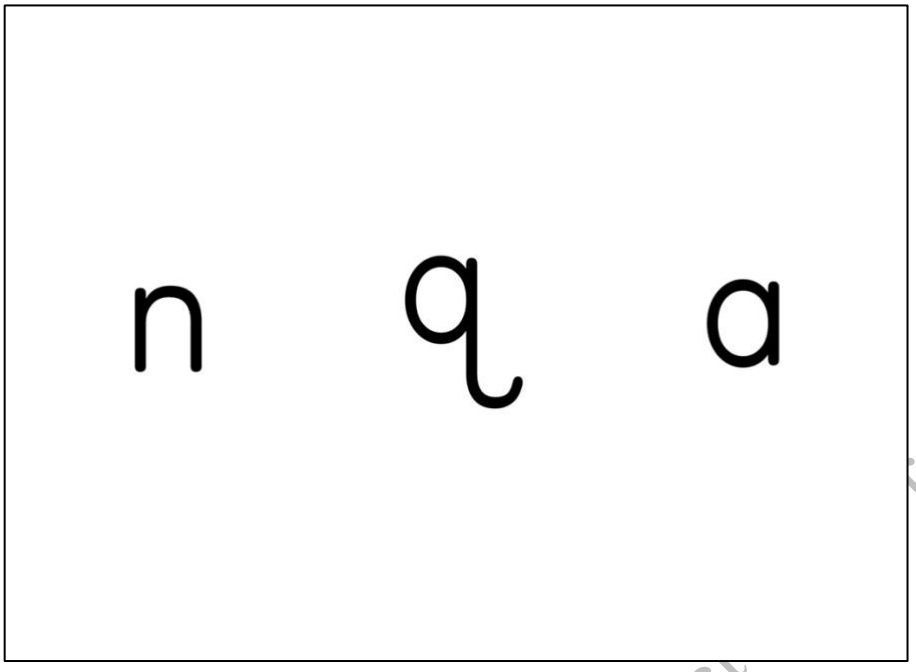

Figure 10

Point to Letter Assessor Screen

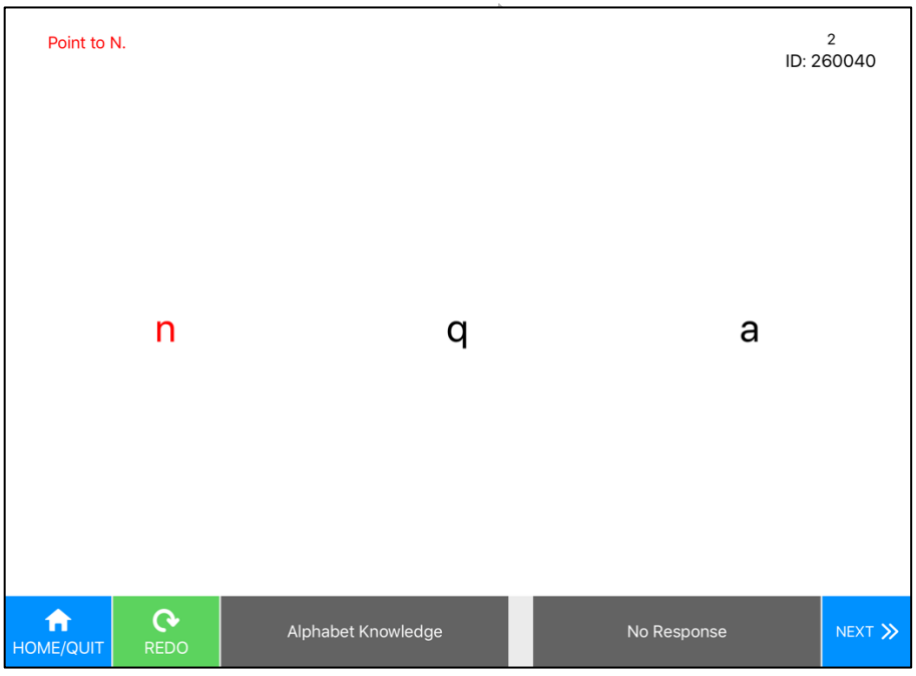

Figure 11

Letter Naming Student Screen

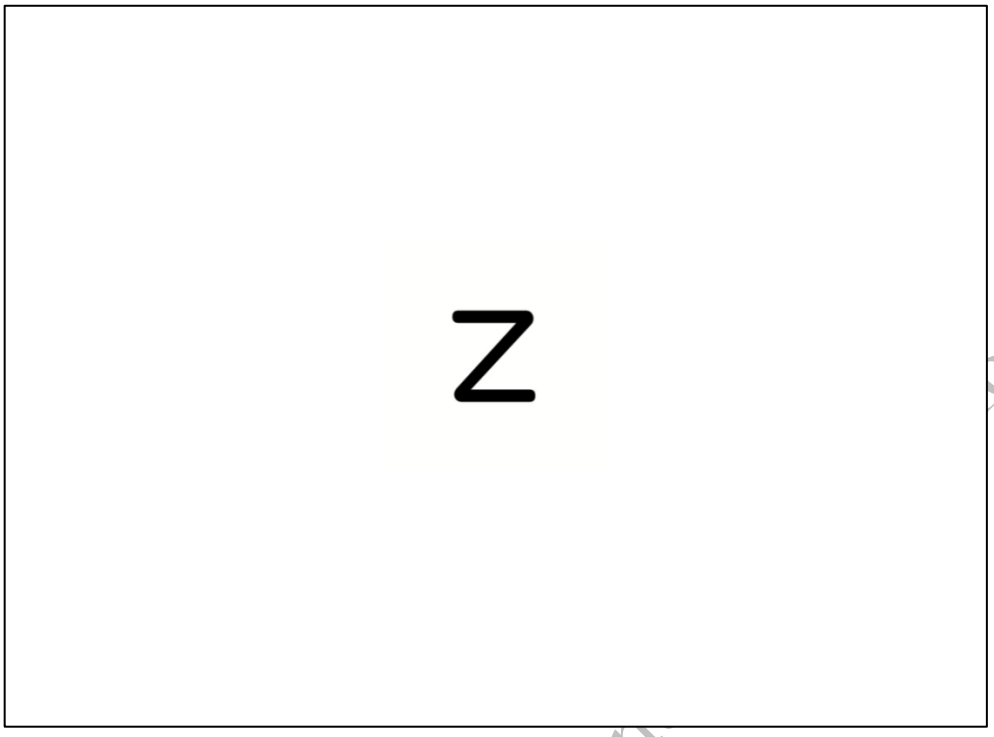

Figure 12

Letter Naming Assessor Screen

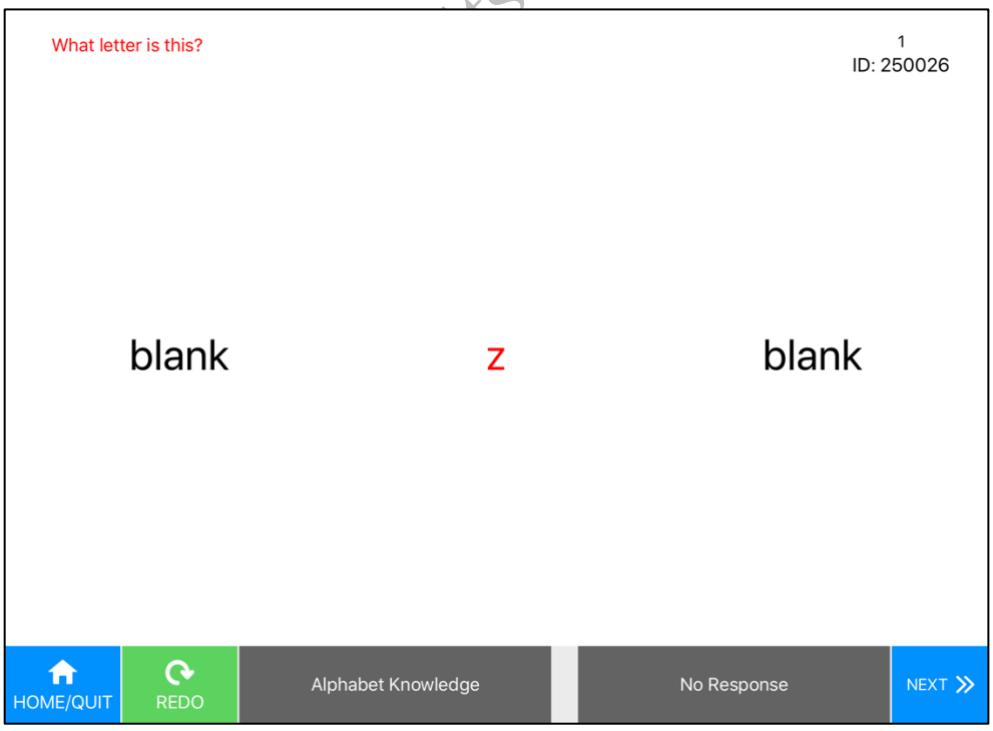

Figure 13

Phonological Awareness (Robot Blending) Student Screen

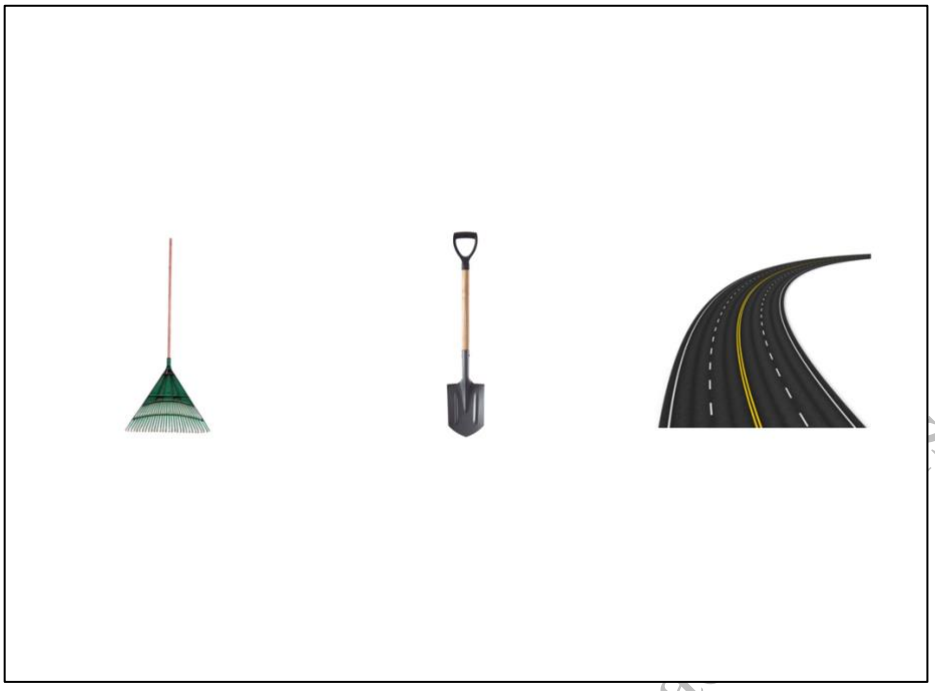

Figure 14

Phonological Awareness (Robot Blending) Assessor Screen

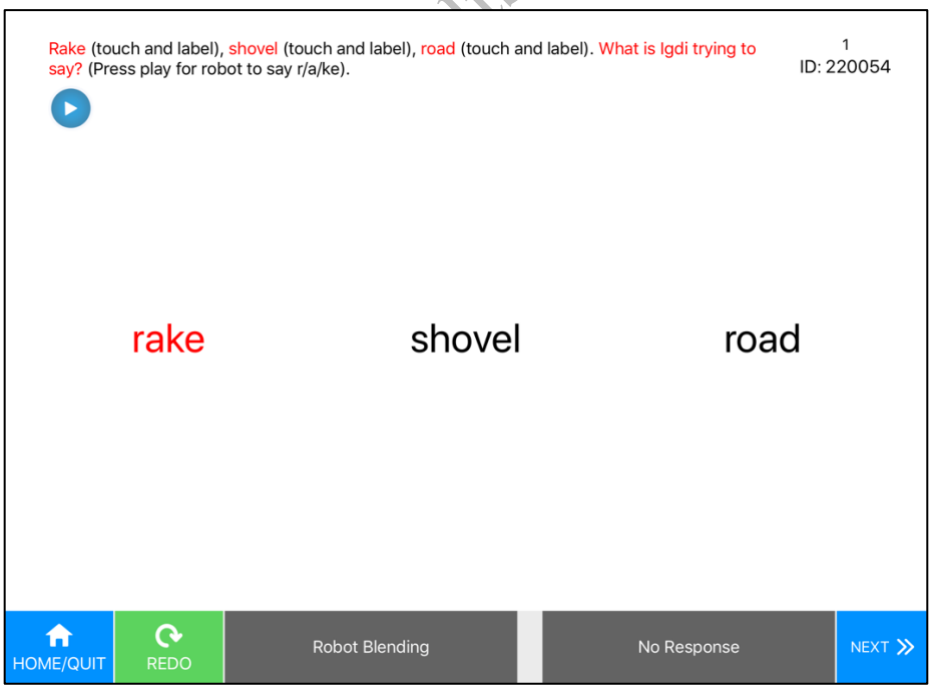

Figure 15

Performance Level Descriptors (PLD) Survey Example: Alphabet Knowledge

Alphabet Knowledge

DESCRIPTION: Please rate each student's ***alphabet knowledge*** – their ability to **know a variety of the names and sounds of the letters** of the alphabet. For example, knowing the name of the letter "M" and the sound ("mmm") that it makes.

Tier 3	**Tier 2**	***Tier 1***
Child has ***significant difficulty*** and needs ***intensive*** intervention	Child has ***moderate difficulty*** and needs ***supplemental intervention***	Child has ***little or no difficulty*** and ***needs no special intervention***

Place mark **the response** that **most accurately describes** each student's Alphabet Knowledge ability.

	Tier Assignment				How certain are you of this rating?		
	Tier 3	Tier 2	Tier 1	N/A - child no longer enrolled	Not very certain	Very certain	N/A - child no longer enrolled

Figure 16

Fall OL Ability Distributions by Teacher PLD Tier Level Candidacy

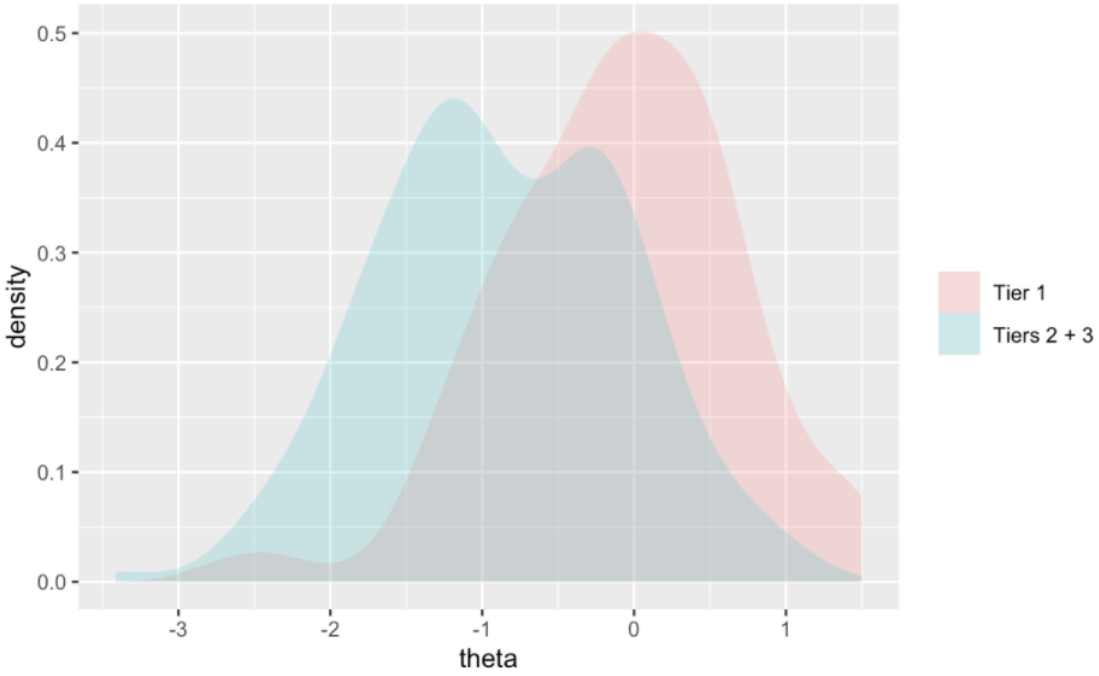

Figure 17

Winter OL Ability Distributions by Teacher PLD Tier Level Candidacy

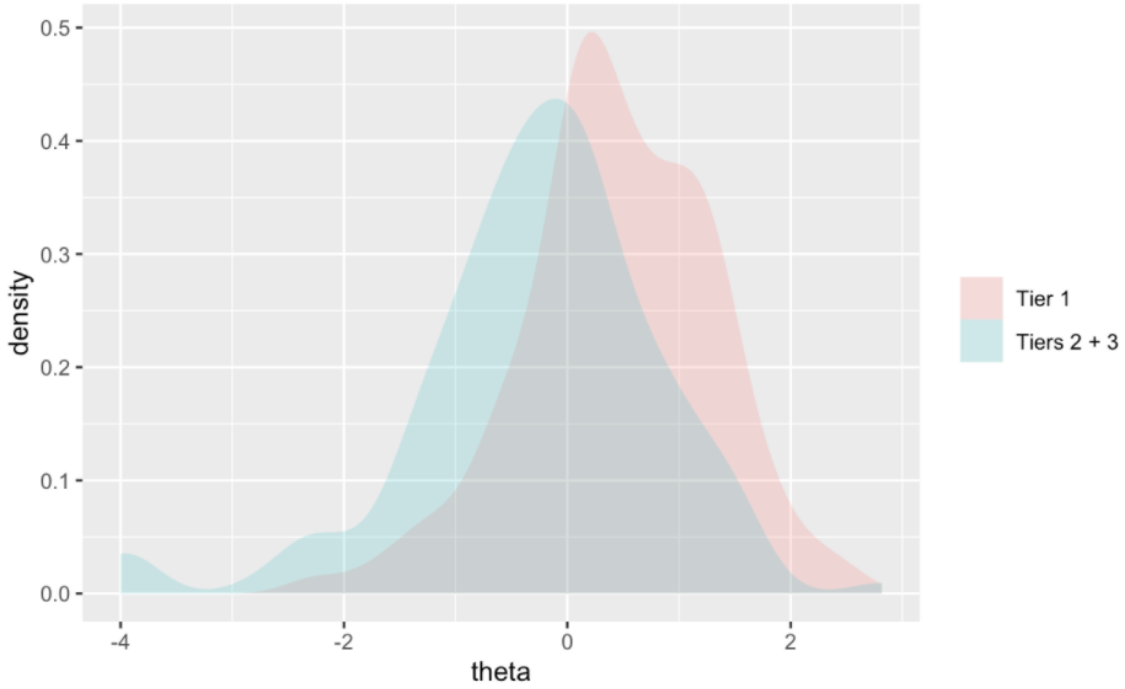

Figure 18

Spring OL Ability Distributions by Teacher PLD Tier Level Candidacy

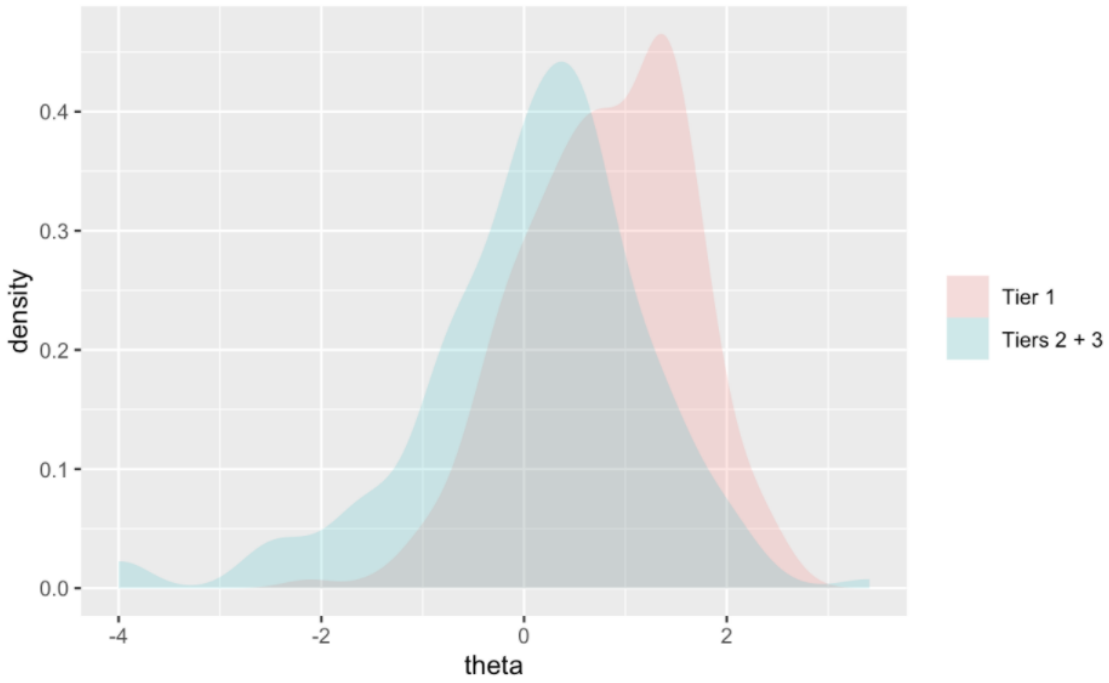

Figure 19

Oral Language AUC for Fall

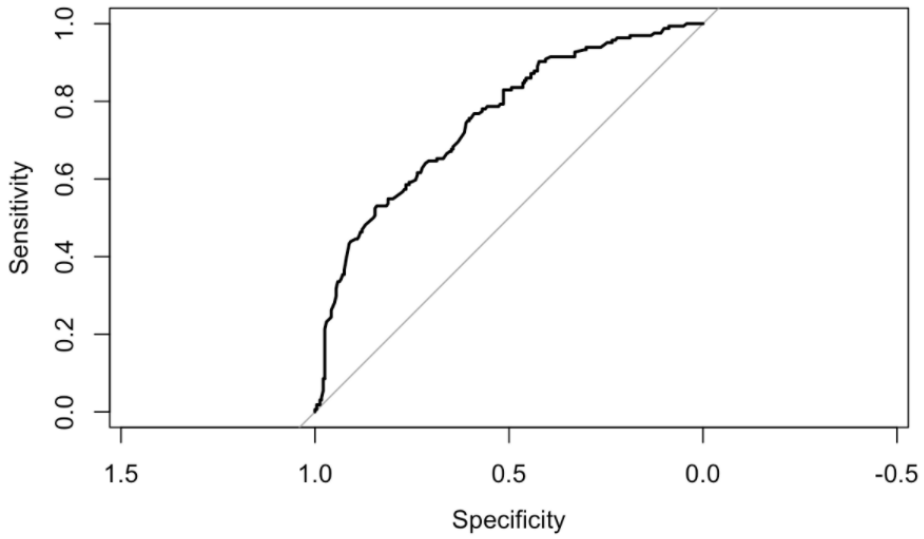

Figure 20

Oral Language AUC for Winter

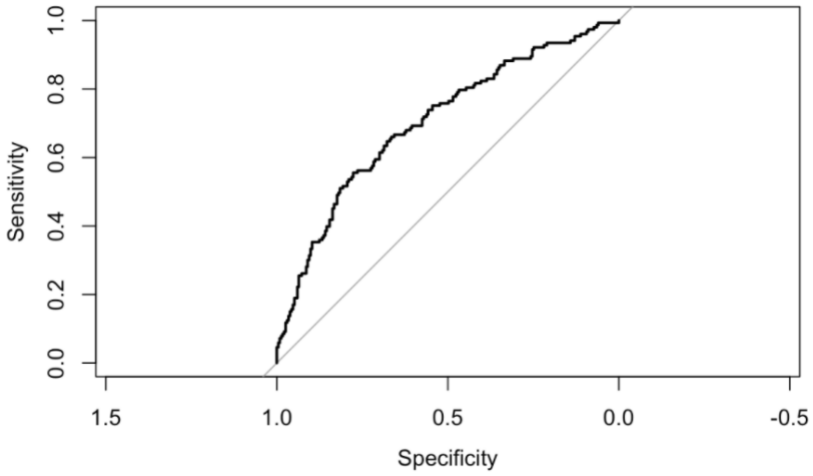

Figure 21

Oral Language AUC for Spring

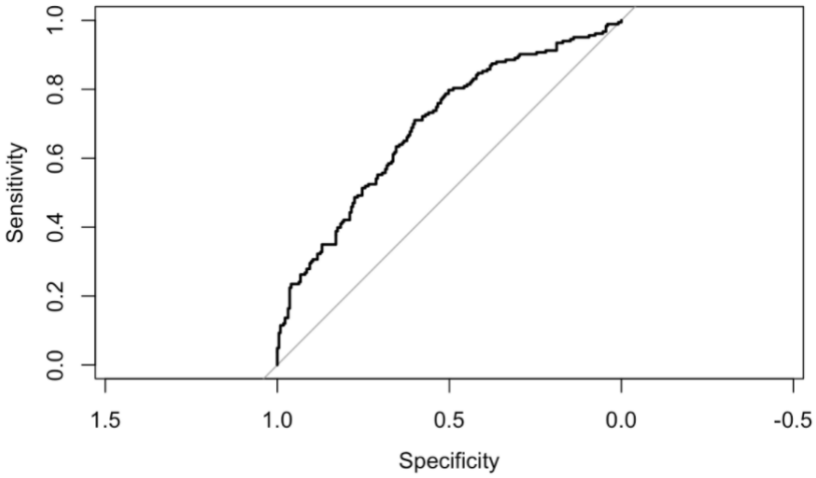

Figure 22

Fall AK Ability Distributions by Teacher PLD Tier Level Candidacy

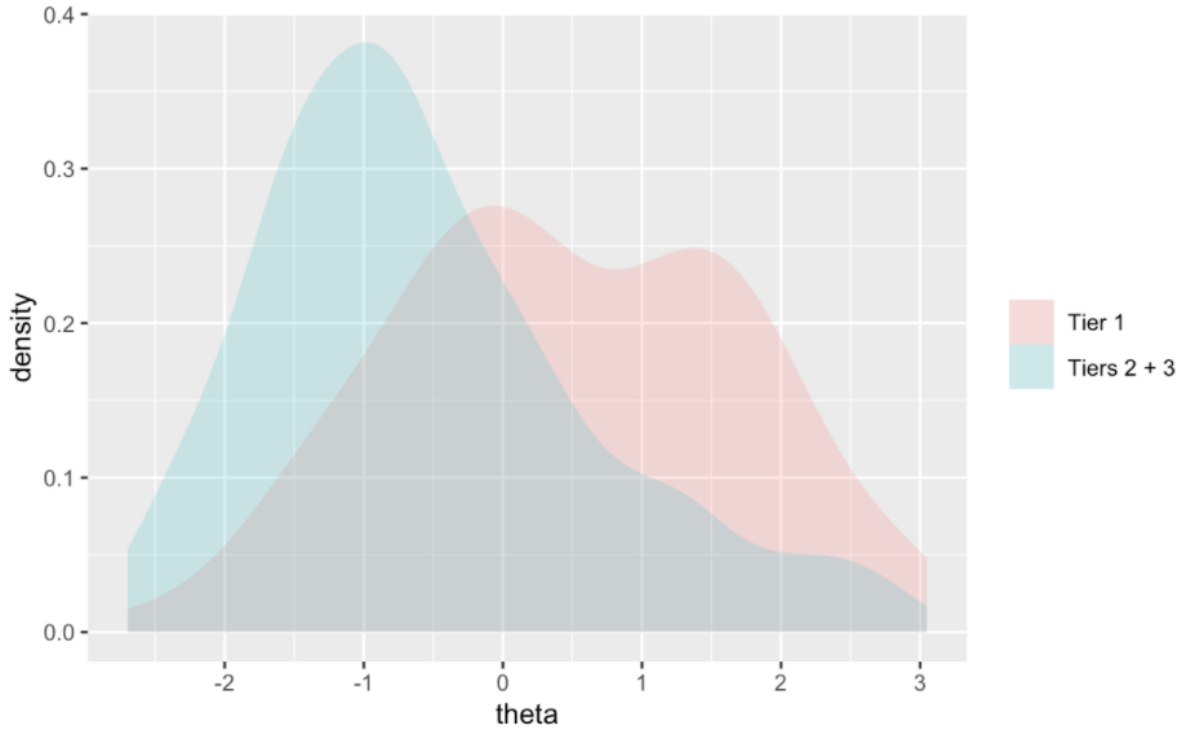

Figure 23

Winter AK Ability Distributions by Teacher PLD Tier Level Candidacy

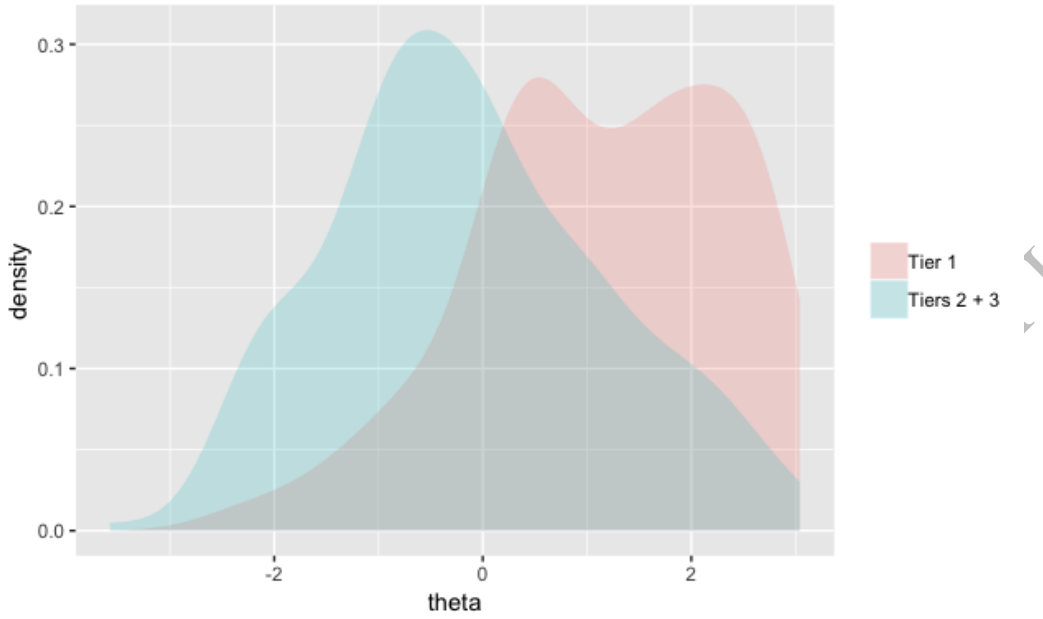

Figure 24

Spring AK Ability Distributions by Teacher PLD Tier Level Candidacy

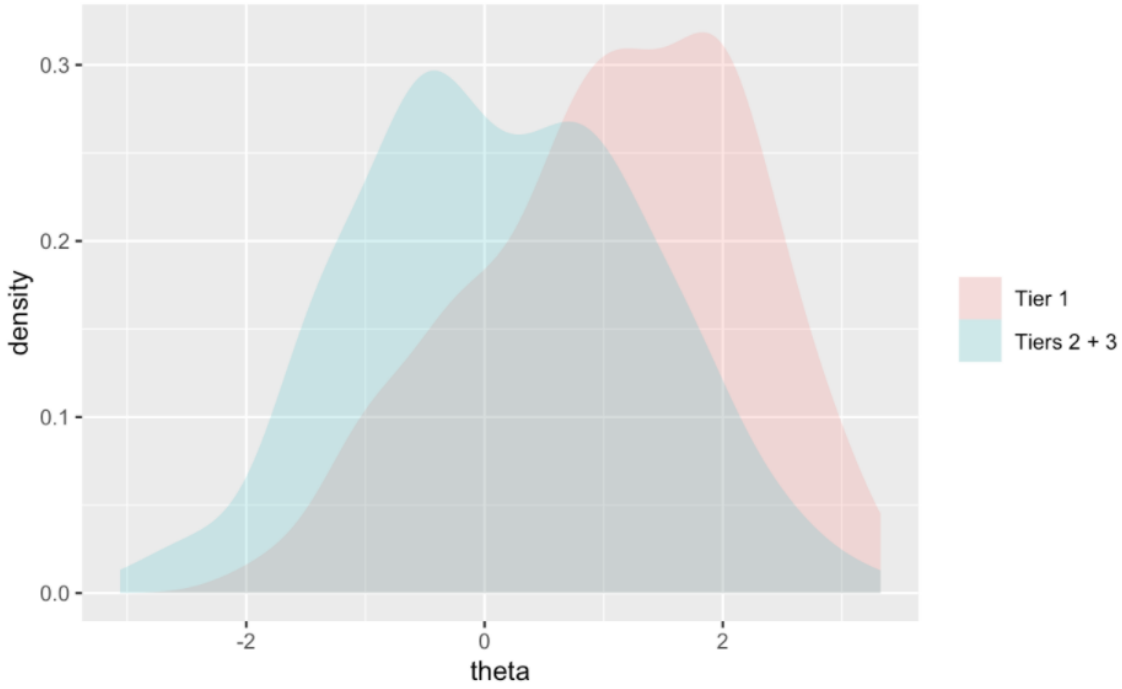

Figure 25

Alphabet Knowledge AUC for Fall

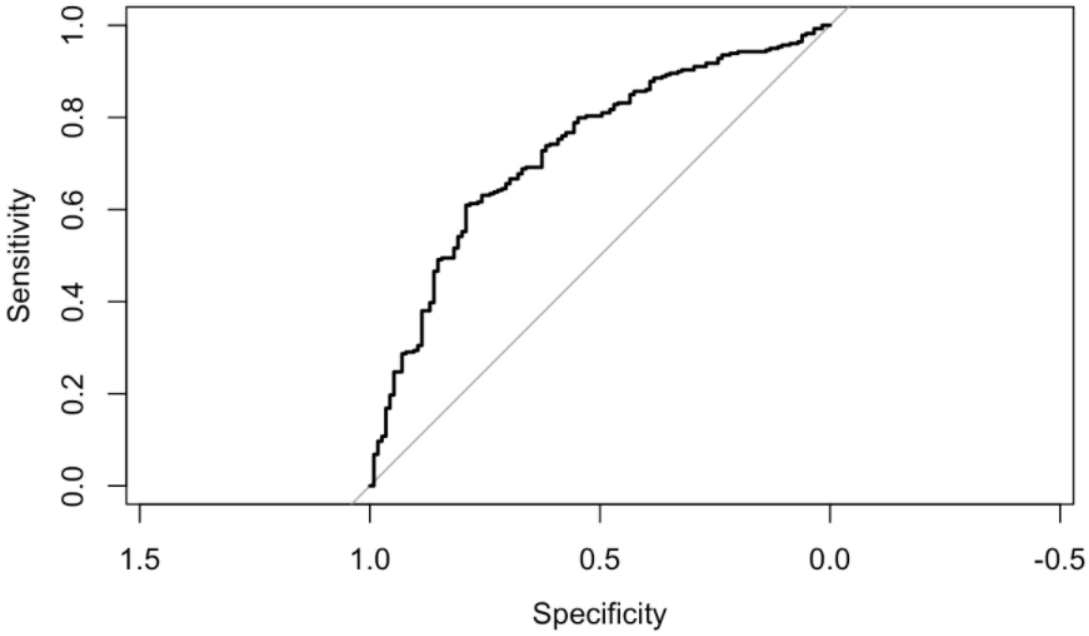

Figure 26

Alphabet Knowledge AUC for Winter

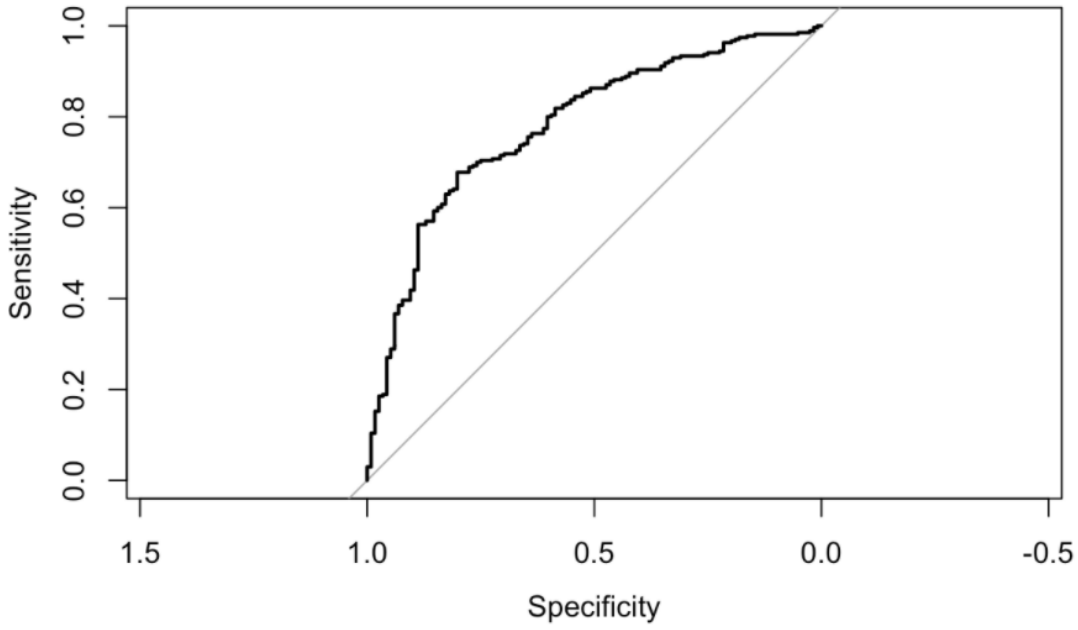

Figure 27

Alphabet Knowledge AUC for Spring

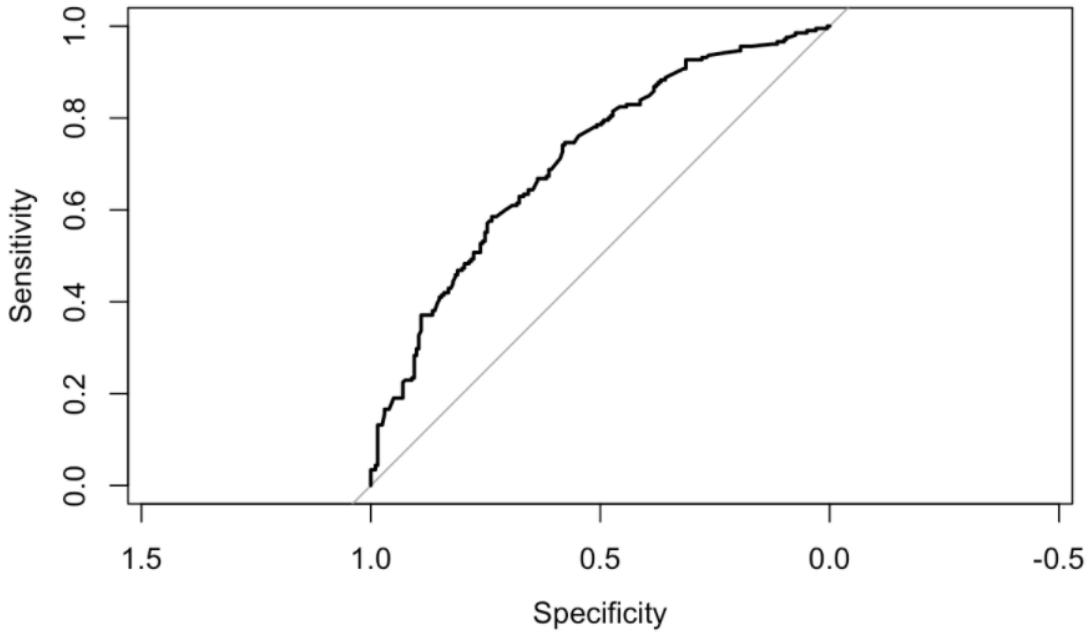

Figure 28

Fall PA Ability Distributions by Teacher PLD Tier Level Candidacy

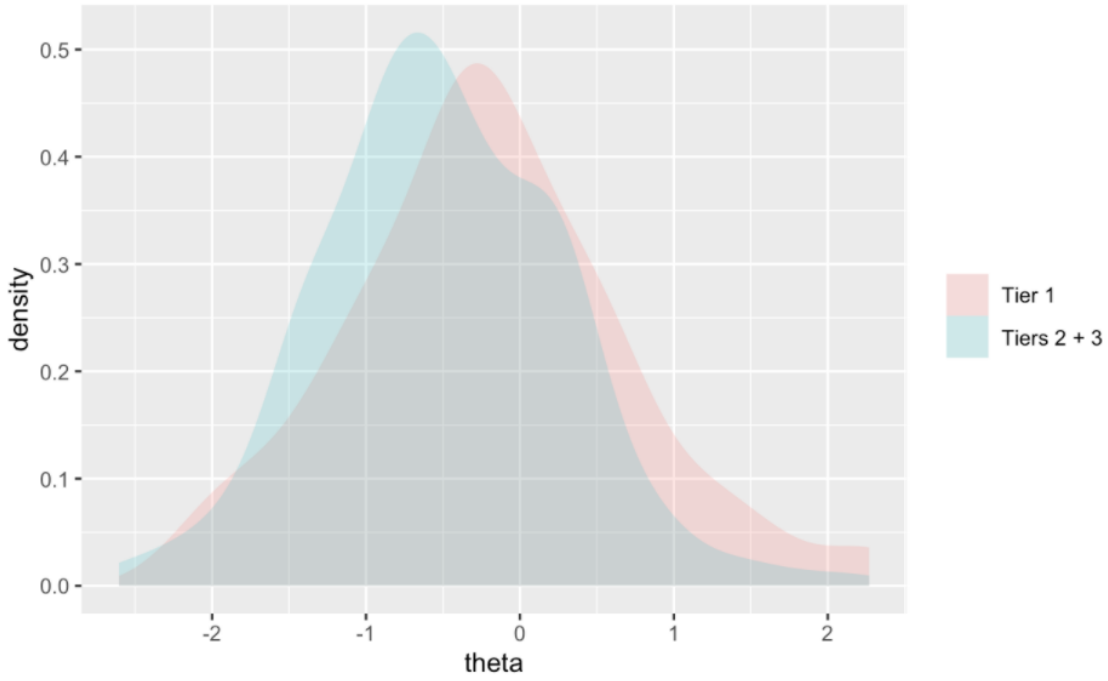

Figure 29

Winter PA Ability Distributions by Teacher PLD Tier Level Candidacy

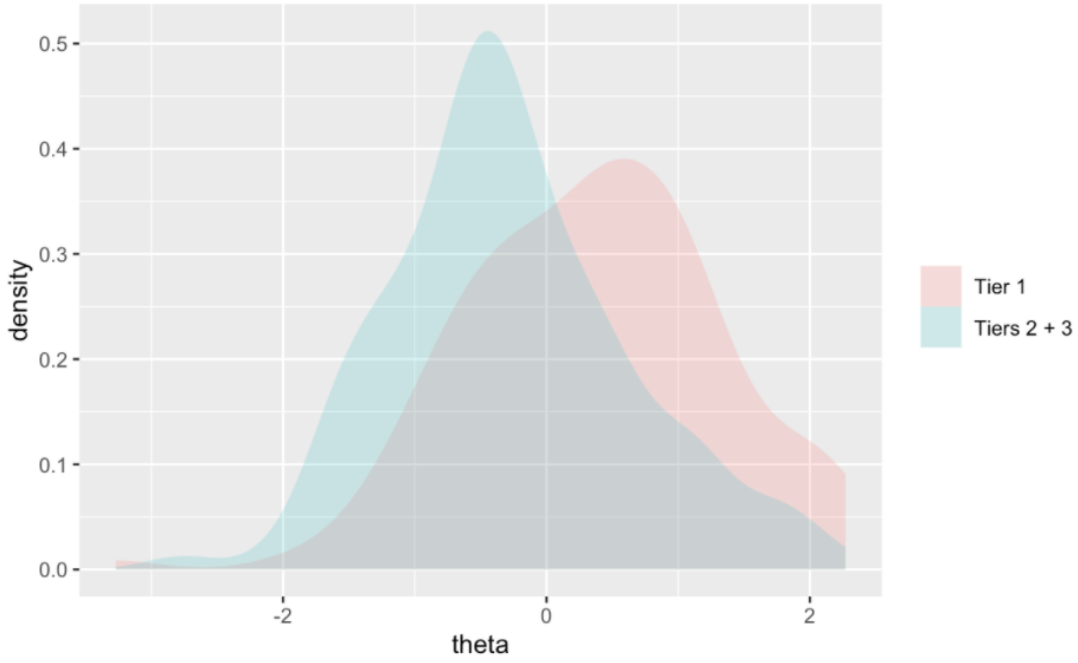

Figure 30

Spring PA Ability Distributions by Teacher PLD Tier Level Candidacy

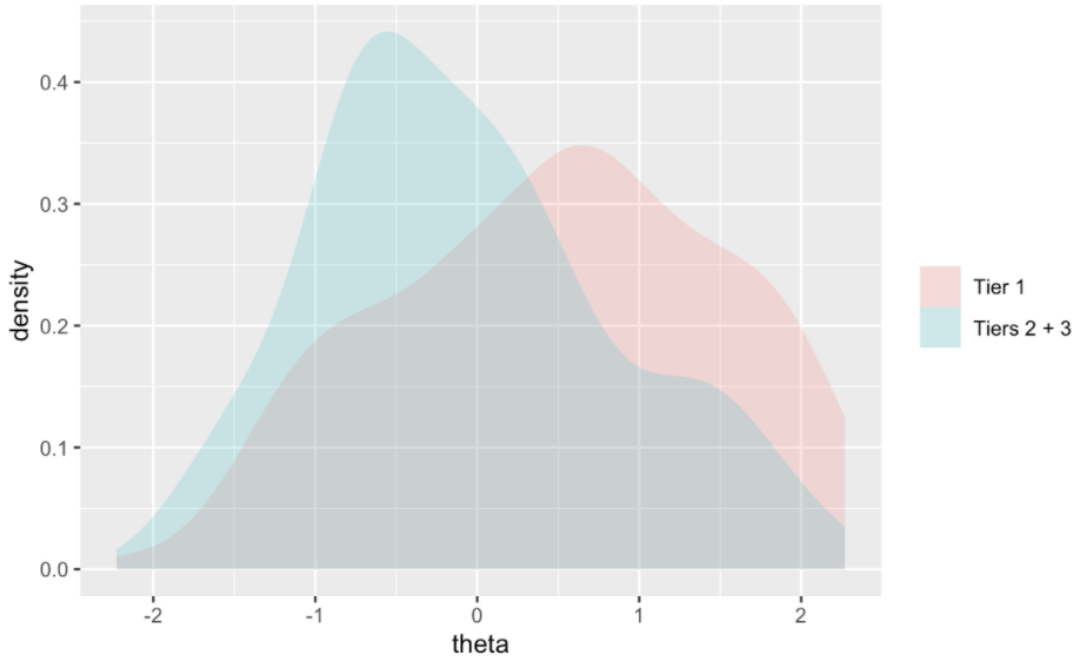

Figure 31

Phonological Awareness AUC for Fall

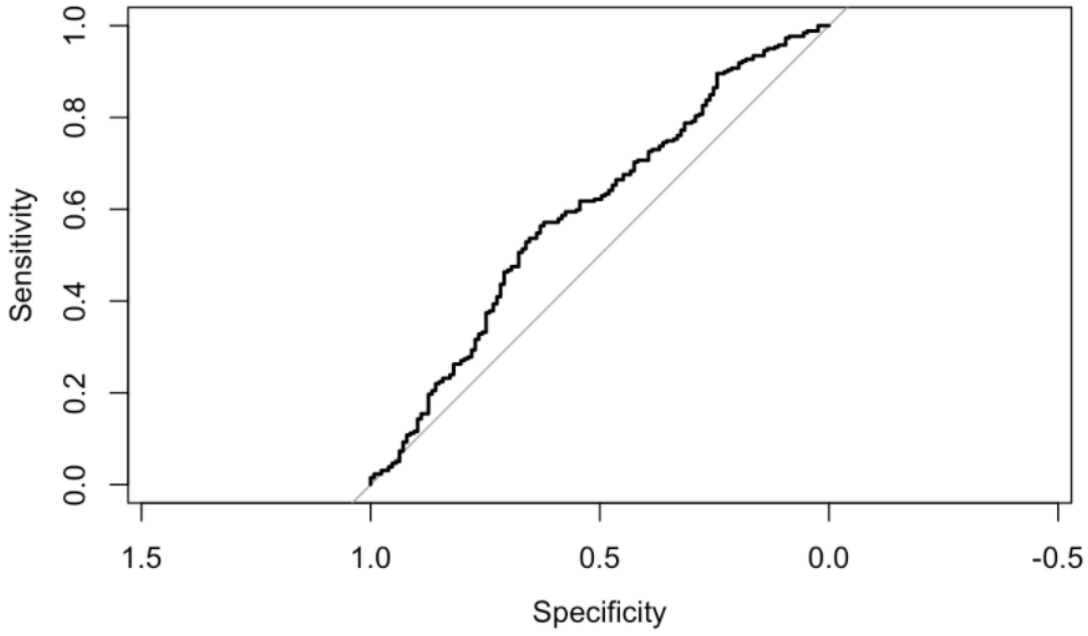

Figure 32

Phonological Awareness AUC for Winter

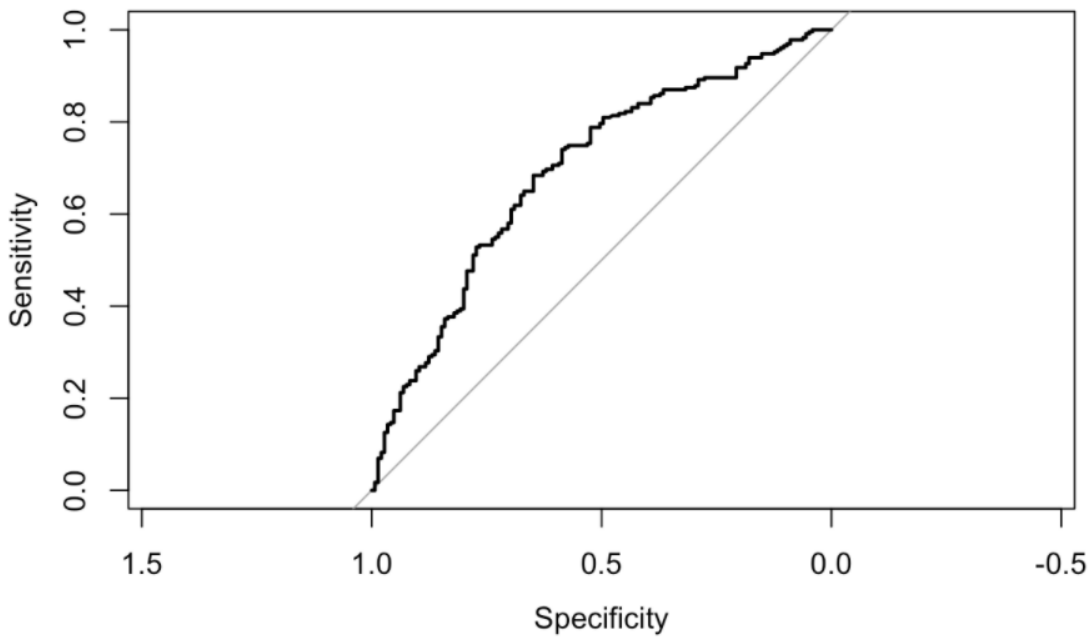

Figure 33

Phonological Awareness AUC for Spring