Rapid Online Assessment of Reading (ROAR)

Technical Manual

Jason D. Yeatman Carrie Townley-Flores Kelly Wentzlof
Wanjing Anya Ma Julian M. Siebert Mia Fuentes-Jimenez
Ana Saavedra Tonya S. Murray

2024-08-24

Contents

A BRIDGE BETWEEN THE LAB AND THE CLASSROOM

Assessments are typically time-consuming and resource-intensive to administer: Individually administering assessments to each student in a classroom means a substantial amount of lost instruction time and requires extensive training for teachers to accurately administer and score measures that are used for high-stakes decisions (e.g., access to intervention). Researchers face these same challenges creating a bottle-neck to research at scale. While education technology companies have built products that lower the demands on teachers, many of these products are expensive, grounded in opaque, proprietary technology, and lack a strong research backing. Hence, these products rarely get used in research, creating a disconnect between educational research and practice.

We launched ROAR envisioning a new model: an open-source, open-access assessment platform, grounded in ongoing academic research, and co-developed in collaboration with school-district stakeholders. Rather than a one-way street from the lab to society (often with a commercial intermediary), ROAR's goal is to inculcate a virtuous cycle between research and practice. We aim to build a suite of completely automated, lightly gamified, online assessments that are grounded in ongoing cognitive neuroscience research and validated against the current "gold standard" of standardized, individually-administered assessments. Our approach is to partner with school districts and community based organizations at each stage of research and development to ensure that our research is grounded in real-world problems and inspired by the deep knowledge of educators who work with children and youth across a diversity of contexts. Through this "Research Practice Partnership" model, we endeavor towards a new assessment methodology that is more valid, precise, efficient, and informative. We aim to design this platform around the diversity of learners in the United States (and abroad). We prioritize transparency at every stage: whenever feasible, materials and technology are made public and each measure within ROAR is published in open-access, peer-reviewed journals with the goal of building more systemic connections between the lab, classroom, and society.

Acknowledgements

ROAR reflects the collective vision and dedication of an incredible team of collaborators. This work would not have been possible without the hundreds of teachers, reading specialists, school administrators, parents, and community based organizations that believed in the ROAR mission (Chapter 1) and collaborated at each stage of ROAR research and development. Additionally, dozens of academics have made important contributions along the way including Ben Domingue, Rebecca Silverman, Joshua Lawrence, Clementine Chou, Amy Burkhardt, Jasmine Tran, Maya Brunton, Aryaman Taore, Kenny Tang, Alby Ungashe, Klint Kanopka, and others.

ROAR would not have been possible without generous funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development[1] (R01HD095861), Advanced Educational Research and Development Fund (AERDF)[2], Stanford-Sequoia K-12 Research Collaborative, Microsoft, Stanford Impact Labs[3], Neuroscience:Translate[4], Klingenstein Foundation[5], Tools Competition[6]. Rapid Visual Processing measures (Section 9.2.2) were developed in collaboration with the UCSF Dyslexia Center[7] with funding from the State of California.

© Jason D. Yeatman, Stanford University

[1] https://www.nichd.nih.gov/
[2] https://aerdf.org/
[3] https://impact.stanford.edu/
[4] https://neuroscience.stanford.edu/programs/research-grants/neurosciencetranslate
[5] https://klingenstein.org/
[6] https://tools-competition.org/winner/rapid-online-assessment-of-reading-roar/
[7] https://dyslexia.ucsf.edu/

Part I

Introduction

1 ROAR Vision and Mission

ROAR emerged out of more than a decade of research in the Brain Development & Education Lab on the neurobiological foundations of literacy. Our goal was to leverage the extensive literature on the cognitive neuroscience of reading development to develop a completely automated, lightly gamified online assessment platform that could replace the resource-intensive and time-consuming conventional approach of individually administering assessments that are scored based on verbal responses. In other words, we endeavored to create a platform that could assess an entire school district in the time typically required to administer an assessment to a single student. We envisioned a new approach to research and development, grounded in the principles of open-science, where each ROAR measure would be grounded in the extensive interdisciplinary literature on reading development, be validated adhering to the highest standards of rigor in each discipline, and be published in open-access journals to support scientific transparency.

1.1 Open-source ideology in educational assessment

The last decade has seen a revolution in scientific transparency. The open-science[1] movement began as a grass roots movement to make science more transparent, accessible and reproducible through open the sharing of code and data to accompany publications in open-access journals. The success of the open-science movement can be appreciated in new public mandates for data sharing by many of the major scientific funders in the United States and Europe, as well as proliferation of organizations like the Center for Open Science[2], and preprint servers like bioRxiv[3], that all make it easier to document, share and reproduce scientific research. In fields like cognitive neuroscience, it is now standard practice for software and algorithms to be open-source, and many journals even require various open-science practices. However, in education, most widely used assessments are grounded in proprietary products, with many of the technical details guarded by pay-walls or made purposefully opaque to maintain a competitive edge in the market. There are, of course, counter examples like DIBELS[4] that have always maintained open-access printed materials, and with projects like the Item Response Warehouse[5] which provides open-access to many educational datasets, their is a clear desire among many educational researchers for a move toward open-science.

[1] https://en.wikipedia.org/wiki/Open_science
[2] https://www.cos.io/
[3] https://www.biorxiv.org/
[4] https://dibels.uoregon.edu/
[5] https://osf.io/preprints/psyarxiv/7bd54

We launched ROAR with the mission to bring the open-source ideology to educational assessment. Our lab has a long track record of developing and supporting open-source software for analysis and sharing of brain imaging data, and for modeling the interplay between brain development and learning. ROAR represents the next phase of this open-science mission: to build tools that fill the needs of educators to assess reading development while, simultaneously, opening the door to research at an unprecedented scale. Not every aspect of ROAR is completely open, but we consciously prioritize open-science at every stage of development including this technical manual which is written as an open-source quarto book[6].

1.2 Approach to validation

Each ROAR measure is rigorously validated both in an academic research setting (i.e., "in the lab") as well as in a typical school setting (i.e., "in the classroom"). We take both these approaches to validation to ensure that ROAR meets the highest standards of rigor across applications in research and practice. Lab validation studies involve recruiting research participants through the typical recruitment avenues of the Brain Development & Education Lab and involve validating new ROAR measures against "gold standard" individually administered diagnostic assessments that are widely accepted by reading and dyslexia researchers. School validation studies are conducted through a Research Practice Partnership model in collaboration with school districts to ensure that ROAR is valid for the desired use cases in the school. Since the question for a school is often "how does ROAR relate to our standard of practice", we report both a) validation of ROAR measures against the current assessments that are used in standard practice in our collaborating schools and b) validation of ROAR measures against validation measures administered by the ROAR research team to students in the district. Together these two approaches to validation have allowed us to extensively examine the accuracy and precision of ROAR relative to a) the constructs it was designed to measure and b) other related measures that are widely used across the United States.

1.3 The ROAR Assessment Suite

ROAR consists of a collection of measures, each designed to tap into a critical aspect of reading. Each individual measure can be run independently and returns raw scores, standard scores, and percentiles relative to national norms. Additionally, measures are also grouped into measurement suites that comprehensively evaluate different constructs in reading development, and produce composite scores and risk indices.

[6] https://github.com/yeatmanlab/roar-tech-manual-public

2 ROAR Assessment Suites

2.1 Foundational Reading Skills

Mastering the code of written language such that text can be fluently decoded into sound and meaning is the foundation of literacy. Foundational reading skills are also the bottleneck for children with dyslexia who struggle to learn letter-sound correspondences and decoding skills early in elementary school and typically have continued struggles with reading fluency. The ROAR Foundational Reading Skills Suite assesses the collection of skills that are at the foundation of reading development and also represent the major challenges for students with dyslexia.

Reading skills develop sequentially with each skill building upon the foundation that has already been established. In an alphabetic script like English where letters represent sounds that are blended together to form words, a critical foundation is Phonological Awareness. Phonological Awareness refers to the ability to identify and manipulate the sounds that make up spoken words. Phonological Awareness typically develops hand in hand with Letter Sound Knowledge. With the development of Phonological Awareness and Letter Sound Knowledge, children begin learning to crack the code of written language, decode text to sound, and read words. Decoding skills and Single Word Reading measure the complexity of words that a student can read going from simple consonant-vowel-consonant words like "cat", to complex, multi-syllabic words like "heterogeneity" (Yeatman et al. 2021). But reading connected text is much more than decoding a sequence of words in isolation; reading sentences efficiently and fluently is critical for comprehension of more complicated texts. ROAR Foundational Reading Skills is a composite of ROAR-Phoneme (Gijbels et al. 2024), ROAR-Letter, ROAR-Word (Yeatman et al. 2021), and ROAR-Sentence (Tran et al. 2023).

2.2 Dyslexia Screening and Subtyping

Developmental dyslexia is an impairment in foundational reading skills. Whereas most children who are provided systematic and structured reading instruction are able to master phonological awareness, letter-sound correspondences, and decoding early in elementary school, children with dyslexia struggle to develop these skills and require substantially more support. For people with dyslexia reading efficiency can remain a challenge throughout life. Thus, the ROAR Foundational Reading Skills Suite (Phoneme, Letter, Word, and Sentence) serves as a reliable and accurate index of the reading challenges associated with dyslexia. Additionally, dyslexia is associated with other challenges including Rapid Automatized Naming (RAN) and various aspects of visual processing. Thus, in addition to the Foundational Reading Skills Suite,

ROAR contains additional measures that have been designed to characterize other challenges that are common in people with dyslexia. These additional measures are useful for both predicting children's struggles with reading development as well as characterizing differences that might contribute to or exacerbate reading challenges.

References

Gijbels, Liesbeth, Amy Burkhardt, Wanjing Anya Ma, and Jason D Yeatman. 2024. "Rapid Online Assessment of Reading and Phonological Awareness (ROAR-PA)." *Sci. Rep.* 14 (1): 1–16.

Tran, Jasmine E, Jason D Yeatman, Amy Burkhardt, Wanjing A Ma, Jamie Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. 2023. "Development and Validation of a Rapid Online Sentence Reading Efficiency Assessment."

Yeatman, Jason D, Kenny An Tang, Patrick M Donnelly, Maya Yablonski, Mahalakshmi Ramamurthy, Iliana I Karipidis, Sendy Caffarra, et al. 2021. "Rapid Online Assessment of Reading Ability." *Sci. Rep.* 11 (1): 6396.

3 ROAR Scores and Norms

ROAR has been developed through a collaborative, co-design process with schools around the United States. The development, validation, score-reporting, and norming samples reflect the contributions of hundreds of schools that have collaborated with the ROAR team through a Research Practice Partnership (RPP) model(Laura Wentworth et al. 2023; L. Wentworth et al. 2021). The goal of this model is to make sure that the diverse interests of stakeholders (teachers, students, parents, school administrators, etc.) representing the incredible diversity in the U.S. education system have a voice in guiding the research and development process of the tools used in their schools. **?@fig-roar-state-map** shows the distribution of ROAR partner schools around the United States.

3.1 ROAR Scores

ROAR assessment report different types of scores that each have intended use cases. The ROAR Families and Teachers Guide[1] provides detailed descriptions of how to interpret scores and use them to guide instruction and/or intervention.

One important consideration for interpreting scores on any assessment is participant effort, concentration, and engagement. A score is only an accurate representative of the participant's ability level if the participant is engaged and tries their hardest even as the assessment gets difficult. For assessments that are individually administered (e.g., by a teacher), the administrator might get a qualitative impression of the participant's effort and focus. If the same person is administering the assessment and interpreting the scores (e.g., classroom teacher or reading specialist), this qualitative impression can be helpful. However it can also be a source of bias and it is hard to standardize the criteria for judging engagement. For automated, online assessments like ROAR, participant disengagement might be a particular concern and needs to be considered when interpreting scores. Each ROAR measure has a defined criteria, grounded in research, for identifying disengaged participants and flagging unreliable scores (for example see Section 15.2 and Section 16.2). This criteria is defined in an algorithm that takes into account a) the response time distribution and b) pattern of responses on the assessment. Scores for any participant that are flagged for disengagement or other issues that might affect the interpretation of the score are flagged in the ROAR Score Report.

Types of Scores

[1] documents/ROAR-Family-Guide-20240814.pdf

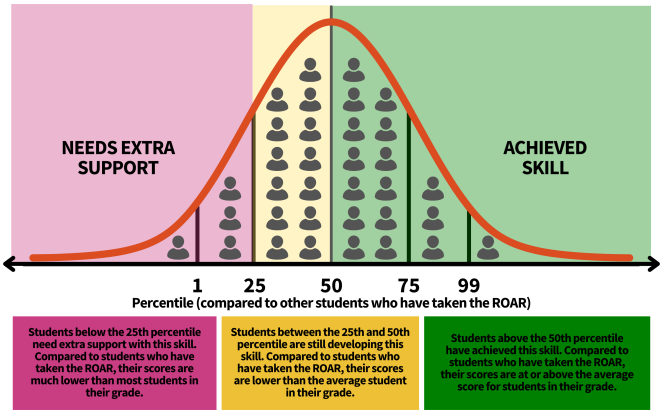


Figure 3.1: Understanding ROAR scores using the normal curve (Gaussian distribution).

- **Raw Scores**: A raw score is the basic measure of a student's performance on the test. Each assessment reports a raw score and the scoring rules for that raw score are detailed in the introduction to that assessment (for example, see Section 5.2 for the scoring rules for ROAR-Word). Most of the assessment use item response theory (IRT; see Section 4.2) and computer adaptive testing (CAT; see Chapter 4) for scoring though some timed assessments like ROAR-Sentence (see Section 12.3) use other types of scoring models. The raw score is comparable across grade levels and over time. **Raw Scores** are useful for tracking growth in reading skills over time.
- **Percentile Scores**: The percentile refers to a student's rank within their grade level on the given skill. The percentile is the number of students out of 100 who have lower scores. Percentile scores are computed by comparing raw scores to a *norming table*. A *norming table* captures the distribution of scores for each age bin in a lookup table providing the percentiles associated with each raw score. Percentile Scores are useful for identifying students who are struggling relative to their peers (or relative to national norms). The norming table for percentile scores are computed in 2 ways:

    1. Based on ROAR Norms. Section 3.2 shows shows the participating ROAR schools, Section 3.3 presents school characteristics and Section 3.4 show student characteristics.
    2. Based on linking ROAR scores to criterion measures like the Woodcock Johnson Basic Reading Skills (WJ BRS) Standard Scores. This linking allows ROAR-Word scores to be interpreted with direct reference to the criterion measure that is often used to define dyslexia risk (for example see Section 22.1).

- **Standard Scores**: A standard score is a way of showing how performance compares to other kids of the same age or grade. The standard score is comparable within a grade level, but not across grade levels or over time. Age standardized scores for ROAR-Word put scores for each age bin on a standard scale (normal distribution, $\mu = 100$, $\sigma = 15$, see Figure 3.1) and are computed in 2 ways:

    1. Based on ROAR Norms

2. Based on linking ROAR scores to WJ BRS Standard Scores.

– **Support Categories**: For each measure, ROAR recommends students who are in need of extra support. Support categories can also be interpreted as indicating risk of reading difficulties such as dyslexia (for more information see Chapter 27 and (?)). Dyslexia refers to the lower end of a continuum of reading skills and there is no agreed upon cutoff. The 25th percentile based on national norms is a common cutpoint that is used to indicate students who are in need of additional support and that is the cut point that is implemented in ROAR Support Categories. Students below the 25th percentile are recommended for additional support. Students between the 25th and 50th percentile are indicated that the skill is still developing. Students above the 50th percentile are indicated as being at grade level (achieved skill). For example, Figure 3.2 shows the distribution of support categories from a hypothetical school district in the ROAR Score Report.

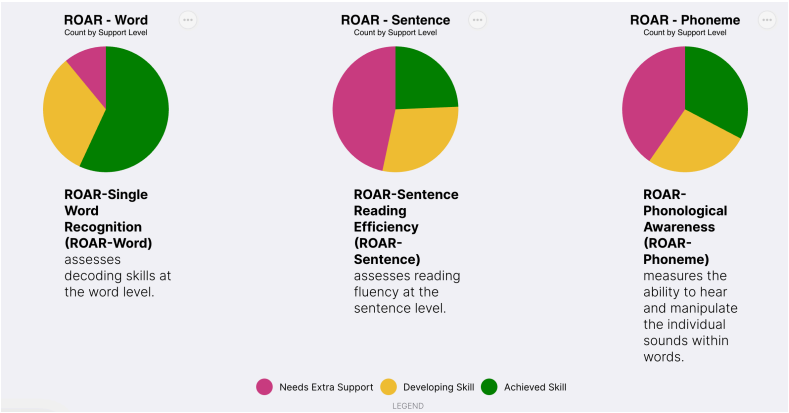


Figure 3.2: Figure from ROAR Score Reports showing the distribution of students for a hypothetical school district that Need Extra Support (red; below 25th percentile), Developing Skill (yellow; 25th - 50th percentile), Achieved Skill at grade level (green; above 50th percentile)

3.2 Map of ROAR schools

```
#> ![Leaflet Map](leaflet_map.png)
```

3.3 Table of school characteristics

	N
Median Students	411
Median Free or Reduced Lunch	269
Race/Ethnicity	
Median Hispanic Ethnicity	155
Median White	31

Median Black or African American	12
Median Asian	14
Median American Indian or Alaska Native	0
Median Hawaiian or Other Pacific Islander	0
Median Multiracial	2402
Organization Type	
Public School	58
Charter School	40
Private School	11
Summer School/Tutor Program/Other	10

3.4 Table of student demographics

	N	%	% Missing
Female	4617	42.91	11.92
Free or Reduced Lunch	1143	10.62	48.87
Race/Ethnicity			
Hispanic Ethnicity	3194	29.69	0.00
White	3713	34.51	0.00
Black or African American	693	6.44	0.00
Asian	1593	14.81	0.00
American Indian or Alaska Native	117	1.09	0.00
Hawaiian or Other Pacific Islander	51	0.47	0.00
Multiracial	976	9.07	0.00
English Language Learner	1792	16.66	40.86
Special Education	381	3.54	43.94
Total	10759		

References

Wentworth, Laura, Paula Arce-Trigatti, Carrie Conaway, and Samantha Shewchuk. 2023. *Brokering in Education Research-Practice Partnerships: A Guide for Education Professionals and Researchers*. Taylor & Francis.

Wentworth, L, R Khanna, M Nayfack, and D Schwartz. 2021. "Closing the Research-Practice Gap in Education." *Stanford Social Innovation Review* 19 (2): 57–58.

4 Computer Adaptive Testing (CAT)

Computer Adaptive Testing (CAT) is a method of administering assessments that adapts to the participant's ability level. As a dynamic approach to assessment, CAT uses algorithms to select items based on the participant's previous answers with the goal of delivering items that are best suited to a participant's ability level. For example, after a correct answer, the next item will be slightly more challenging; if the answer is incorrect, the following item will be easier. This process continues throughout the test, allowing the CAT algorithm to pinpoint the participant's ability level with greater precision and efficiency than traditional fixed tests. CAT offers several advantages, including shorter testing times, reduced test anxiety due to fewer irrelevant questions, and enhanced test security, as each test is unique to the individual. One of the requirements of a computer adaptive test is that responses can be scored in real-time so that the next item can be selected. This is a strength of ROAR as all items are scored immediately. Most of the ROAR assessments including ROAR-Phoneme (see Chapter 7), ROAR-Letter (see Chapter 8), and ROAR-Word (see Chapter 5), are implemented as CATs. Timed assessment including ROAR-Sentence (see Chapter 6) and ROAR-RAN (see Section 9.2.1) are not computer adaptive because for these "fluency" measures response time is fundamental to the measurement so it is critical that each participant sees the same items (or equated items). All computer adaptive ROAR measures use jsCAT[1], an open-source, Javscript CAT package developed by the ROAR team.

The implementation of CAT within each ROAR measure allows for three unique properties of ROAR:

1. Since each measure adapts to the participants ability level, students across a very large age range can be compared on the same "vertical" measurement scale. A vertical scale in educational assessment is a single, continuous scale used to measure student achievement or ability across multiple grade levels or age groups. This type of scaling allows for the comparison of test scores over time, providing a coherent framework to track academic growth and development. For example, a 1st grader and an 8th grader can both take ROAR-Word and the CAT algorithm will ensure that the 1st grader is presented easier items than the 8th grader. The score that is returned by ROAR-Word will put them on the same measurement scale so that an individuals growth can be tracked across the grades, over the course of an intervention, or can be compared to scores in different grades.

[1] https://github.com/yeatmanlab/jscat

2. Individual ROAR measures are highly efficient. Since each individual subtest is controlled by a CAT algorithm, they produce very reliable scores with fewer items than a traditional approach.
3. Our CAT implementation allows ROAR to simultaneously operate as an efficient screener that returns risk metrics based on composite scores while also providing precise measures of specific, actionable sub-skills.

4.1 CAT parameters and item selection

Unless otherwise specified, ROAR assessments use a Rasch model and items are selected based on Fisher information (Linden 2000; Ma et al. 2023).

4.2 IRT models and item validation

Unless otherwise specified, ROAR uses a Rasch model (one parameter logistic), to put items on a vertical scale. Infit and outfit statistics are used to ensure that each item fits the measurement scale well. Any items with with infit/outfit statistics outside the range of 0.7 - 1.3[2] are removed (Wu and Adams 2013). Ensuring that each item fits the measurement scale validates that the item taps into the same latent construct as the other items in the assessment. Finally, to ensure that the measurement scale does not have bias against any demographic group we take two approaches:

1. We run validation studies to ensure that the reliability and criterion validity are equivalent across race, ethnicity, and socio-economic status, school district, and level of English proficiency. We take particular care to validate ROAR for English language learners (ELLs).
2. We run studies of parameter invariance (see (Ma et al. 2023)) to ensure that the difficulty of each item is consistent across samples spanning different school districsts with different demographics. Parameter invariance ensures that the assessments function equivalently across diverse groups of participants.

References

Linden, Wim J van der. 2000. *Computerized Adaptive Testing: Theory and Practice*. Edited by Cees A W Glas. Dordrecht, Netherlands: Kluwer Academic.

Ma, Wanjing A, Adam Richie-Halford, Amy Burkhardt, Klint Kanopka, Clementine Chou, Benjamin Domingue, and Jason D Yeatman. 2023. "ROAR-CAT: Rapid Online Assessment of Reading Ability with Computerized Adaptive Testing."

Wu, Margaret, and Richard J Adams. 2013. "Properties of Rasch Residual Fit Statistics." *J. Appl. Meas.* 14 (4): 339–55.

[2] https://www.rasch.org/rmt/rmt162f.htm

5 Single Word Reading (ROAR-Word)

ROAR-Word uses a lexical decision task to measure single word reading ability. Rapid and automatic word recognition is a foundation of most theories of reading ability (Ehri 2005; Perfetti 1985), and a two alternative forced choice (2AFC) lexical decision task (LDT) has a rich history in the cognitive science literature as a means to probe the processes underlying word recognition. Decades of theory and data have demonstrated that lexical decisions and reading out loud tap into many of the same underlying cognitive processes (Seidenberg and McClelland 1989; Balota, Yap, and Cortese 2006; Katz et al. 2012; Keuleers et al. 2012). While reading out loud is often considered the "gold standard" methodology for assessing single word reading ability, there is no theoretical reason to prefer a verbal versus a silent response. The goal of a single word reading assessment is to measure the complexity of words that can be accurately read and silent reading, not reading out loud, is the ultimate goal. Reading out loud has an easily observable response so is the focus of many assessments of single word reading. But reading out loud also has the drawback of requiring knowledge to be accurately mapped to a very specific motor output which can be a barrier for children with speech impediments as well as for those who speak with accents or language variants that are not reflected in the scoring criteria of the assessment (Brown et al. 2015; Washington and Seidenberg 2021). These issues with language variation also allow potential bias of the administrator to affect the scoring of the assessment: many standardized assessments consider language variants as errors (Washington and Seidenberg 2021). Moreover, reading out loud has the practical drawback of not being amenable to group administration, requiring manual scoring, and often having subjective scoring criteria. For all these reason, we designed ROAR-Word as a silent, lexical decision task to avoid the bias and barriers to scale that can be inherent in measures that are scored based on reading out loud. The item bank was designed based on the extensive cognitive science literature on lexical decisions such that items systematically sample different theoretically important dimensions of lexical and orthographic properties (see Yeatman et al. 2021; Ma et al. 2023 for more a more detailed description of the item bank).

In addition to the theoretical reasons why a properly designed lexical decision task can serve as a valid index of single word reading ability, there is also broad empirical support. For example (Van Bon, Hoevenaars, and Jongeneelen 2004) compared a pencil and paper lexical decision task to a standardized, single word reading aloud measure and found strong correlations between the silent and oral reading measures. (Bon, Tooren, and Eekelen 2000) even found that lexical decisions predicted reading out loud at the item level. Commercially available assessments, such as the CAPTI from ETS, also rely on a modified LDT as a measure of "word recognition and decoding" (Sabatini et al. 2019; Wang et al. 2019). Sabatini et al. (2019) present a strong case for face validity, and report that this measure is distinct from vocabulary and comprehension. In creating ROAR-Word, we a) designed an item bank grounded in

the cognitive science of reading development, b) used item response theory (IRT) to calibrate an efficient and highly reliable unidimensional Rasch measurement scale, and c) undertook a series of studies to demonstrate excellent convergent validity with standardized, individually administered measures of single word reading (Yeatman et al. 2021; Ma et al. 2023; Barrington et al. 2023).

5.1 Structure of the task

Figure 5.1 shows the structure of the ROAR-Word task. ROAR-Word is a computer adative test (CAT) with 84 items that are sampled from a large item bank (>600 items). ROAR-Word begins with instructions that are narrated by characters that introduce the task as a fun adventure. Instructions are followed by practice trials with feedback to ensure that the participant understands the goal of the lexical decision task and how to respond to real and nonsense words. Finally, after practice trials, the participant is presented with 84 CAT items divided across three blocks with breaks in between each block. Throughout the task, participants are rewarded by gold coins and new charachters along their journey which helps with focus and engagement.

For each task trial a real or pseudo word is presented for 350ms and the participant is cued to indicate with a keypress, touchscreen or swipe whether the word was real or made up. The participant can take as long as needed to respond.

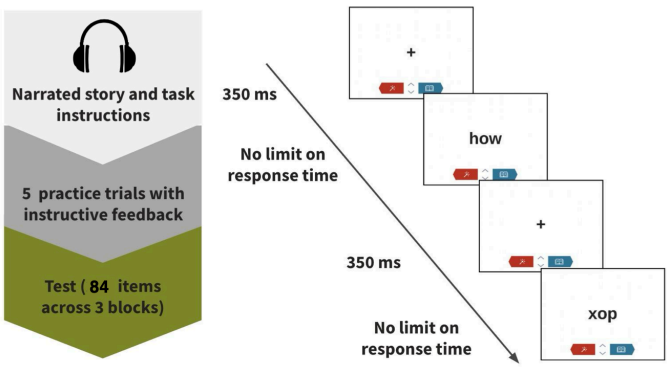


Figure 5.1: ROAR-Word task.

5.2 Scoring

ROAR-Word is a two alternative forced choice (2AFC) task and items are scored as correct or incorrect (dichotomous scoring) by comparing the participant's response (keypress/touchscreen/swipe indicating real word vs. pseudo word), to the true classification (real/pseudo) of the item. Each response is scored in real time. The participant's ability or *raw score* (θ) on ROAR-Word is computed based on an item response theory (IRT) model. IRT puts ability (θ) on an interval scale meaning that scores can be compared over time and across grades.

Types of Scores

- **Raw Scores**: ROAR-Word raw scores are computed based on an IRT model which puts ability (θ) on an interval scale. Raw scores are then put on a scale spanning 100-900 by applying a linear transform to θ estimates. RAOR Score = round $\left(\left(\frac{\theta+6}{3} \times 200\right) + 100\right)$
- **Percentile Scores**: Percentile scores are computed in 2 ways (see Chapter 3):

    1. Based on ROAR Norms
    2. Based on linking ROAR scores to Woodcock Johnson Basic Reading Skills (WJ BRS) Standard Scores. This linking allows ROAR-Word scores to be interpreted with direct reference to the criterion measure that is often used to define dyslexia risk.

- **Standard Scores**: Age standardized scores for ROAR-Word put scores for each age bin on a standard scale (normal distribution, $\mu = 100$, $\sigma = 15$) and are computed in 2 ways:

    1. Based on ROAR Norms
    2. Based on linking ROAR scores to WJ BRS Standard Scores.

5.3 Design and validation of items

The ROAR-Word measurement scale has already been validated in previous publications (Yeatman et al. 2021; Ma et al. 2023; Barrington et al. 2023). Each individual item is validated based on its fit to the IRT measurement scale. ROAR-Word uses a Rasch (one parameter logistic) model (with a guess rate of 0.5) and item fit is assessed based on infit and outfit statistics using standard criteria (Wu and Adams 2013). Items are included in the assessment if infit and outfit are betwen 0.7 and 1.3.

Additionally, item parameters are validated to ensure that items function similarly across diverse demographic groups. Ma et al. (2023) report an analysis of parameter invariance across 4 groups of school districts that differ dramatically in terms of:

- Race/Ethnicity
- Percentage of English language learners
- Median income
- Percentage of students who qualify for free and reduced price lunch

Ma et al. (2023) demonstrated that after removing a small number of items, item difficulty was consistent across all the school districts included in the sample.

References

Balota, David A, Melvin J Yap, and Michael J Cortese. 2006. "Visual Word Recognition." In *Handbook of Psycholinguistics*, 285–375. Elsevier.

Barrington, Elizabeth, Sadie Mae Sarkisian, Heidi M Feldman, and Jason D Yeatman. 2023. "Rapid Online Assessment of Reading (ROAR): Evaluation of an Online Tool for Screening Reading Skills in a Developmental-Behavioral Pediatrics Clinic." *J. Dev. Behav. Pediatr.* 44 (9): e604–10.

Bon, Wim H J van, Paula H Tooren, and Kees W J M van Eekelen. 2000. "Lexical Decision and Oral Reading by

Poor and Normal Readers." *Eur. J. Psychol. Educ.* 15 (3): 259–70.
Brown, Megan C, Daragh E Sibley, Julie A Washington, Timothy T Rogers, Jan R Edwards, Maryellen C MacDonald, and Mark S Seidenberg. 2015. "Impact of Dialect Use on a Basic Component of Learning to Read." *Front. Psychol.* 6 (March): 196.
Ehri, Linnea C. 2005. "Learning to Read Words: Theory, Findings, and Issues." *Sci. Stud. Read.* 9 (2): 167–88.
Katz, Leonard, Larry Brancazio, Julia Irwin, Stephen Katz, James Magnuson, and D H Whalen. 2012. "What lexical decision and naming tell us about reading." *Read. Writ.* 25 (6): 1259–82.
Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. "The British Lexicon Project: Lexical Decision Data for 28,730 Monosyllabic and Disyllabic English Words." *Behav. Res. Methods* 44 (1): 287–304.
Ma, Wanjing A, Adam Richie-Halford, Amy Burkhardt, Klint Kanopka, Clementine Chou, Benjamin Domingue, and Jason D Yeatman. 2023. "ROAR-CAT: Rapid Online Assessment of Reading Ability with Computerized Adaptive Testing."
Perfetti, Charles A. 1985. "Reading Ability" 282.
Sabatini, John, Jonathan Weeks, Tenaha O'Reilly, Kelly Bruce, Jonathan Steinberg, and Szu-Fu Chao. 2019. "SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 3rd Edition." *ETS Res. Rep. Ser.* 2019 (1): 1–30.
Seidenberg, M S, and J L McClelland. 1989. "A distributed, developmental model of word recognition and naming." *Psychol. Rev.* 96 (4): 523–68.
Van Bon, Wim H J, Lotje T M Hoevenaars, and Joyce J Jongeneelen. 2004. "Using Pencil□and□paper Lexical□decision Tests to Assess Word Decoding Skill: Aspects of Validity and Reliability." *J. Res. Read.* 27 (1): 58–68.
Wang, Zuowei, John Sabatini, Tenaha O'Reilly, and Jonathan Weeks. 2019. "Decoding and Reading Comprehension: A Test of the Decoding Threshold Hypothesis." *J. Educ. Psychol.* 111 (3): 387–401.
Washington, J A, and M S Seidenberg. 2021. "Teaching Reading to African American Children: When Home and School Language Differ." *American Educator*.
Wu, Margaret, and Richard J Adams. 2013. "Properties of Rasch Residual Fit Statistics." *J. Appl. Meas.* 14 (4): 339–55.
Yeatman, Jason D, Kenny An Tang, Patrick M Donnelly, Maya Yablonski, Mahalakshmi Ramamurthy, Iliana I Karipidis, Sendy Caffarra, et al. 2021. "Rapid Online Assessment of Reading Ability." *Sci. Rep.* 11 (1): 6396.

6 Sentence Reading Efficiency (ROAR-Sentence)

ROAR-Sentence measures the speed or efficiency with which participants can read sentences for understanding. Being able to efficiently read connected text for understanding is at the foundation of reading development and is a major bottle-neck for children with dyslexia. ROAR-Sentence is specifically designed to tap into reading efficiency and sentences are syntactically and semantically simple, avoid low frequency words and complex sentence structures, and have unambiguous answers that require little, if any, background knowledge.

As reading skills develop, the speed with which children can read connected text becomes particularly important (Silverman et al. 2013). "Efficient word recognition" was highlighted in the original conceptualization of the "simple view of reading" (Hoover and Gough 1990), and fluent reading has been implicated as a bridge between decoding skills and reading comprehension (Pikulski and Chard 2005; Silverman et al. 2013). Children with dyslexia and other word reading difficulties often struggle to achieve fluency, and struggles with word reading speed and fluency have always been core to the definition of dyslexia (Catts et al. 2024; Lyon, Shaywitz, and Shaywitz 2003). Tran et al. (2023) provide a detailed description of the development of a silent sentence reading efficiency (SRE) measure that was designed to be fast, reliable, efficient at scale, and targeted to the issues with speed/fluency that present a bottleneck for so many struggling readers. We provide key details here in the technical manual (using much of the same text) and refer readers to the peer-reviewed publication for more details of the research and development process (Tran et al. 2023).

6.1 Defining the construct of Sentence Reading Efficiency

Timed reading measures go under a variety of names (e.g., reading fluency, reading efficiency, etc) and involve different levels of demands on comprehension and articulation, making it hard to interpret the extent to which scores reflect differences in reading efficiency versus separate constructs. We designed ROAR-Sentence to isolate reading efficiency by minimizing comprehension demands while maintaining checks for understanding, and we used a silent reading task to a) avoid the confounds of articulation that are inherent to oral reading tasks and b), measure the most ecologically valid form of reading (silent reading). This stands in contrast to other reading fluency measures that confound articulation, comprehension and efficiency leading to a less interpretable score.

6.2 Other measures of oral and silent reading efficiency and fluency

Traditional measures that are most similar to ROAR-SRE are sometimes referred to as sentence reading fluency tasks, and while they are not administered online, they do elicit silent responses from students. For example, the Woodcock Johnson (WJ) Tests of Achievement "Sentence Reading Fluency" subtest (Schrank et al. 2014), and Test Of Silent Reading Efficiency and Comprehension (TOSREC) (Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. 2010), rely on an established design: A student reads a set of sentences and endorses whether each sentence is true or false. For example, the sentence, Fire is hot, would be endorsed as True. A student endorses as many sentences as they can within a fixed time limit (usually three minutes). The final score is the total number of correctly endorsed sentences minus the total number of incorrectly endorsed sentences. Both the WJ and TOSREC are standardized to be administered in a one-on-one setting (though TOSREC can also be group administered) and the stimuli consist of printed lists of sentences which students read silently and mark True/False with a pencil. Even though the criteria for item development on these assessments is not specified in detail, there is a growing literature showing the utility of this general approach. First of all, this quick, 3 minute assessment is straightforward to administer and score and has exceptional reliability, generally between 0.85 and 0.90 for alternate form reliability (Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. 2010; Johnson, Pool, and Carter 2011; Wagner 2011). Moreover, this measure has been shown to be useful for predicting performance on state reading assessments: For example, Johnson and colleagues demonstrated that TOSREC scores could accurately predict students who did not achieve grade-level performance benchmarks on end-of-the-year state testing of reading proficiency (Johnson, Pool, and Carter 2011).

Further evidence for validity comes from the strong correspondence between silent sentence reading measures such as the TOSREC and Oral Reading Fluency (ORF) measures (Denton et al. 2011; Wagner 2011; Johnson, Pool, and Carter 2011; Kang and Shin 2019; Y.-S. Kim, Wagner, and Lopez 2012; Y.-S. G. Kim, Park, and Wagner 2014; Price et al. 2016). ORF is one of the most widely used measures of reading development in research and practice, and some have even argued for ORF as an indicator of overall reading competence (Fuchs et al. 2001). ORF is widely used to chart reading progress in the classroom, providing scores with units of words per minute that can be examined longitudinally (e.g., for progress monitoring (Cummings, Park, and Bauer Schaper 2013; Good, Gruba, and Kaminski 2002; Hoffman, Jenkins, and Dunlap 2009)), compared across classrooms and districts, and can inform policy decisions such as how to confront learning loss from the Covid-19 pandemic Domingue et al. (2022). Even though silent reading and ORF are highly correlated, the measures also have unique variance (Hudson et al. 2008; ?; ?) and, theoretically, have different strengths and weaknesses. For example, even though there are strong empirical connections between ORF and reading comprehension (Y.-S. G. Kim, Park, and Wagner 2014), ORF does not require any understanding of the text and has been labeled by some as "barking at print" (Samuels 2007). Silent reading, on the other hand, is the most common form of reading, particularly as children advance in reading instruction. In line with this theoretical perspective, Kim and colleagues found that silent sentence reading fluency was a better predictor of reading comprehension than ORF starting in second grade (Y.-S. Kim, Wagner, and Lopez 2012). Thus, given the practical benefits of silent read-

ing measures (easy to administer and score at scale), along with the strong empirical evidence of reliability, concurrent, and predictive validity, and face validity of the measure, an online measure of silent sentence reading efficiency would be useful for both research and practice.

6.3 Structure of the task and design of the items

ROAR-Sentence uses a similar task design as the Test of Silent Reading Efficiency and Comprehension (TOSREC) and WJ Sentence Reading Fluency sub-test with the major differences being: a) ROAR-Sentence is a gamified online task rather than pencil and paper and b) the ROAR-Sentence items are designed specifically to tap into silent sentence reading efficiency. Figure 6.1 shows the ROAR-Sentence task.

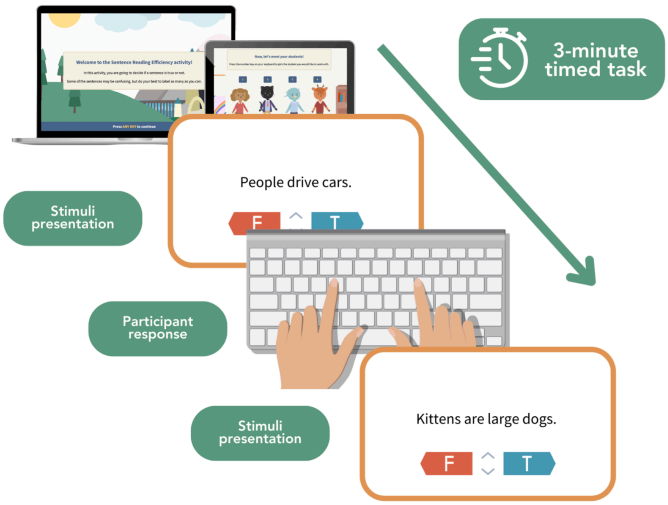


Figure 6.1: ROAR-Sentence task. In the ROAR-Sentence task, participants first choose the character they want to read with and then they are instructed to read the sentences as quickly as possible and indicate with a button-press whether each sentence is true or false. Their score is computed as the number of correct responses minus the number of incorrect responses in either 180 second or 90 seconds depending on the version of the task

The strength of silent reading fluency/efficiency tasks is also their weakness: On the one hand, these tasks include comprehension, which bolsters the argument for the face validity of silent reading measures. On the other hand, what is meant by comprehension in these sentence reading tasks is often ill-defined and, thus, a low score lacks clarity on whether the student is struggling due to difficulties with "comprehension" or "efficiency". As a concrete example, sentences in the TOSREC incorporate low frequency vocabulary words (e.g., porpoise, bagpipes, locomotive, greyhounds, buzzards) meaning that vocabulary knowledge as well as specific content knowledge (e.g., knowledge about porpoises, bagpipes and locomotives) will affect scores. While this design decision might be a strength in some scenarios (e.g., generalizability to more complex reading measures such as state testing), it presents a challenge for interpretability. An interpretable construct is critical if scores are used to individualize instruction. For example, does a fourth grade student with a low TOSREC score need targeted instruction and practice focused on a) building greater automaticity and efficiency in reading

or b) vocabulary, syntax and background knowledge. Our goal in designing a new silent sentence reading efficiency measure was to more directly target reading efficiency by designing simple sentences that are unambiguously true or false and have minimal requirements in terms of vocabulary, syntax and background knowledge. Ideally, this measure could be used to track reading rate in units of words per minute, akin to a silent reading version of the ORF task, but with a check to ensure reading for understanding.

To consider the ideal characteristics of these sentences, it may be helpful to begin by considering the ORF task which is used to compute an oral reading rate (words per minute) for connected text. In an ORF task, the test administrator can simply count the number of words read correctly to assess each student's reading rate. Translating this task to a silent task that can be administered at scale online poses an issue because an administrator is unable to monitor the number of sentences read by the student. A student could be instructed to press a button on the keyboard after the completion of a sentence in order to proceed to the next one. However, the validity of this method depends on the student's ability to exhibit restraint and wait until the completion of each sentence before proceeding to the next sentence.

In the interest of preserving the validity of the interpretations of the scores, we retain the True/False endorsement of the TOSREC and WJ, but reframe its use. That is, for the ROAR-Sentence task, the endorsement of True/False should be interpreted as an indication that the student has read the sentence, rather than as an evaluation of comprehension per se. In this context, if the student has difficulty comprehending a sentence, or if the student takes a long time to consider the correct answer because the sentence is confusing, syntactically complex, or depends on background knowledge and high-level reasoning, we lose confidence in the inferences that we can make about a student's reading efficiency. As such, it is important that sentences designed for this task are simple assertions that are unambiguously true or false. However, creating sentences to adhere to these basic standards may not always be straightforward. For example, the statement "the sky is blue" may be true for a student in the high-plain desert in Colorado but may be a controversial statement for a student in Seattle. Thus, careful consideration must be given to crafting sentences that do not depend on specific background knowledge and are aligned with the goal of measuring reading efficiency. (Tran et al. 2023) provides a detailed description of the iterative research and design process that went into defining and validating this construct and the reader is referred to that publication for more details on the item bank.

6.4 Scoring

ROAR-Sentence is a two alternative forced choice (2AFC) task and is scored as the total number of correct responses minus the total number of incorrect responses in the alloted (3 minute) time window. This scoring method controls for guessing by controlling for the number of incorrect responses in the calculation of the scores.

Types of Scores

- **Raw Scores**: The student's *raw score* will range between 0-130.

- **Percentile Scores**: Percentile scores are computed in 2 ways (see Chapter 3):
  1. Based on ROAR Norms
  2. Based on linking ROAR scores to Test of Silent Reading Efficiency and Comprehension (TOSREC) and Woodcock Johnson Sentence Reading Fluency Standard Scores. This linking allows ROAR-Sentence scores to be interpreted with direct reference to the criterion measure that is often used to define dyslexia risk.
- **Standard Scores**: Age standardized scores for ROAR-Sentence put scores for each age bin on a standard scale (normal distribution, $\mu = 100$, $\sigma = 15$) and are computed in 2 ways:
  1. Based on ROAR Norms
  2. Based on linking ROAR scores to TOSREC and WJ Standard Scores.

References

Catts, Hugh W, Nicole Patton Terry, Christopher J Lonigan, Donald L Compton, Richard K Wagner, Laura M Steacy, Kelly Farquharson, and Yaacov Petscher. 2024. "Revisiting the Definition of Dyslexia." *Ann. Dyslexia*, January.

Cummings, Kelli D, Yonghan Park, and Holle A Bauer Schaper. 2013. "Form Effects on DIBELS Next Oral Reading Fluency Progress-Monitoring Passages." *Assessment for Effective Intervention* 38 (2): 91–104.

Denton, Carolyn A, Amy E Barth, Jack M Fletcher, Jade Wexler, Sharon Vaughn, Paul T Cirino, Melissa Romain, and David J Francis. 2011. "The Relations Among Oral and Silent Reading Fluency and Comprehension in Middle School: Implications for Identification and Instruction of Students with Reading Difficulties." *Sci. Stud. Read.* 15 (2): 109–35.

Domingue, Benjamin W, Madison Dell, David Lang, Rebecca Silverman, Jason Yeatman, and Heather Hough. 2022. "The Effect of COVID on Oral Reading Fluency During the 2020–2021 Academic Year." *AERA Open* 8: 23328584221120254.

Domingue, Benjamin W, Heather J Hough, David Lang, and Jason Yeatman. 2021. "Changing Patterns of Growth in Oral Reading Fluency During the COVID-19 Pandemic. Working Paper." *Policy Analysis for California Education, PACE*.

Fuchs, Lynn S, Douglas Fuchs, Michelle K Hosp, and Joseph R Jenkins. 2001. "Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis." *Sci. Stud. Read.* 5 (3): 239–56.

Good, Roland H, Jerry Gruba, and Ruth A Kaminski. 2002. "Best Practices in Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an Outcomes-Driven Model."

Hoffman, Amy R, Jeanne E Jenkins, and S Kay Dunlap. 2009. "Using DIBELS: A Survey of Purposes and Practices." *Reading Psychology* 30 (1): 1–16.

Hoover, Wesley A, and Philip B Gough. 1990. "The Simple View of Reading." *Read. Writ.* 2 (2): 127–60.

Hudson, Roxanne F, Paige C Pullen, Holly B Lane, and Joseph K Torgesen. 2008. "The Complex Nature of Reading Fluency: A Multidimensional View." *Reading & Writing Quarterly* 25 (1): 4–32.

Johnson, Evelyn S, Juli L Pool, and Deborah R Carter. 2011. "Validity Evidence for the Test of Silent Reading Efficiency and Comprehension (TOSREC)." *Assess. Eff. Interv.* 37 (1): 50–57.

Kang, Eun Young, and Mikyung Shin. 2019. "The Contributions of Reading Fluency and Decoding to Reading Comprehension for Struggling Readers in Fourth Grade." *Read. Writ. Q.* 35 (3): 179–92.

Kim, Young-Suk Grace, Chea Hyeong Park, and Richard K Wagner. 2014. "Is Oral/Text Reading Fluency a 'Bridge' to Reading Comprehension?" *Read. Writ.* 27 (1): 79–99.

Kim, Young-Suk, Richard K Wagner, and Danielle Lopez. 2012. "Developmental Relations Between Reading Fluency and Reading Comprehension: A Longitudinal Study from Grade 1 to Grade 2." *J. Exp. Child Psychol.*

113 (1): 93–111.
Lyon, G Reid, Sally E Shaywitz, and Bennett A Shaywitz. 2003. "A Definition of Dyslexia." *Ann. Dyslexia* 53 (1): 1–14.
Pikulski, J J, and D J Chard. 2005. "Fluency: Bridge Between Decoding and Reading Comprehension." *Read. Teach.*
Price, Katherine W, Elizabeth B Meisinger, Max M Louwerse, and Sidney D'Mello. 2016. "The Contributions of Oral and Silent Reading Fluency to Reading Comprehension." *Read. Psychol.* 37 (2): 167–201.
Samuels, S Jay. 2007. "The DIBELS Tests: Is Speed of Barking at Print What We Mean by Reading Fluency?"
Schrank, F A, K S McGrew, N Mather, B J Wendling, and E M LaForte. 2014. "Woodcock-Johnson IV Tests of Achievement." Riverside Publishing Company.
Silverman, Rebecca D, Deborah L Speece, Jeffrey R Harring, and Kristen D Ritchey. 2013. "Fluency Has a Role in the Simple View of Reading." *Sci. Stud. Read.* 17 (2): 108–33.
Tran, Jasmine E, Jason D Yeatman, Amy Burkhardt, Wanjing A Ma, Jamie Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. 2023. "Development and Validation of a Rapid Online Sentence Reading Efficiency Assessment."
Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. 2010. *Test of Silent Reading Efficiency and Comprehension*. Pro Ed.
Wagner, Richard K. 2011. "Relations Among Oral Reading Fluency, Silent Reading Fluency, and Reading Comprehension: A Latent Variable Study of First-Grade Readers." *Sci. Stud. Read.* 15 (4): 338–62.

7 Phonological Awareness (ROAR-Phoneme)

Phonological awareness (PA), or phonological processing more broadly, refers to a metalinguistic skill that enables an individual to reflect upon and manipulate the sound structure of spoken language (Tunmer, Herriman, and Nesdale 1988), at the level of (1) words and syllables, (2) onsets and rimes, or (3) phonemes independent of their meaning (Stanovich 2017; Treiman and Zukowski 1991). PA skills emerge early, develop rapidly throughout childhood (Nittrouer, Studdert-Kennedy, and McGowan 1989), and have important implications for literacy achievement (Moyle, Heilmann, and Berman 2013). To learn how to read, a child needs to be aware of the arbitrary and conventional correspondence between the sound structure of language and the rules for how it is written (Anthony and Francis 2005). Thus, measures of PA are an important indicator of early reading development and are one of the most established measures in dyslexia screening. ROAR-Phoneme was developed and validated as an efficient, precise and automated measure of PA that does not depend on verbal responses (Gijbels et al. 2024). Gijbels et al. (2024) provide a detailed account of the research, development and validation process, and we lay out the pertinent technical details here.

7.1 Structure of the task

ROAR-Phoneme has 3 sub-tests that measure different dimensions of phonological awareness:

- First-Sound Matching (FSM)
- Last-Sound Matching (LSM)
- Deletion (DEL)

And two optional sub-test that were deemed unnecessary for obtaining a reliable and valid measure of PA but still have utility in specific use cases (for more information see (Gijbels et al. 2024)):

- Blending (BLE)
- Rhyming (RHY)

ROAR-Phoneme employs a one-interval, three-alternative forced choice task. Instructions are narrated by a character, and the participant selects the appropriate response with a touchscreen or mouse click. ROAR-Phoneme consists of 3 subtests (19 items per subtest), with each subtest consisting of 2 or 3 blocks (divided by difficulty level). Each subtest starts off with 3 training items with feedback. Training items have to be completed correctly before the task will continue to the test items. To engage children from pre-k through 5th grade, the task is

embedded in a story where the child has to help a monkey and his friends (rabbit, bear, and otter) collect their favorite foods. At the end of every trial, images of food are displayed. At the end of every block, a visualization of the collected food is presented, and the character provides encouragement (e.g., "Great job! So many bananas! Let's get a few more!"). Every character provides the instructions of their own subtest, guides throughout the task, motivates the participants to take short breaks between blocks, and introduces their next friend.

Participants are instructed to work on a computer or tablet sitting at a desk. Sound has to be turned on and set to a comfortable level, based on the instructions of one of the characters. All game instructions are provided in both text and audio. For all trials of all subtests, one image, accompanied by an audio fragment, recorded by a native English-speaker, provides the specific instruction of that trial (e.g., subtest FSM: "Which picture starts with the same sound as dog?"). This screen shown in Figure 7.1 is followed by the instruction image (top) plus three answer options (left, middle, right). All images are verbalized (e.g., "dam" (target), "goat" (foil 1), "mop" (foil 2)) and the position of the images is randomized for each trial. For the DEL subtest, the images are not verbalized as this could give away the correct answer. In contrast, participants are allowed to listen to the instruction phrase two times for this subtest. We did not implement the ability to listen to the instructions more than once throughout the entire task to stay as consistent as possible with standardized PA tasks like the CTOPP-2 in which instructions are not repeated. Participants pick a response by clicking the image, which is followed by a visualization of the response with a random number of food images as motivation (Figure 7.1).

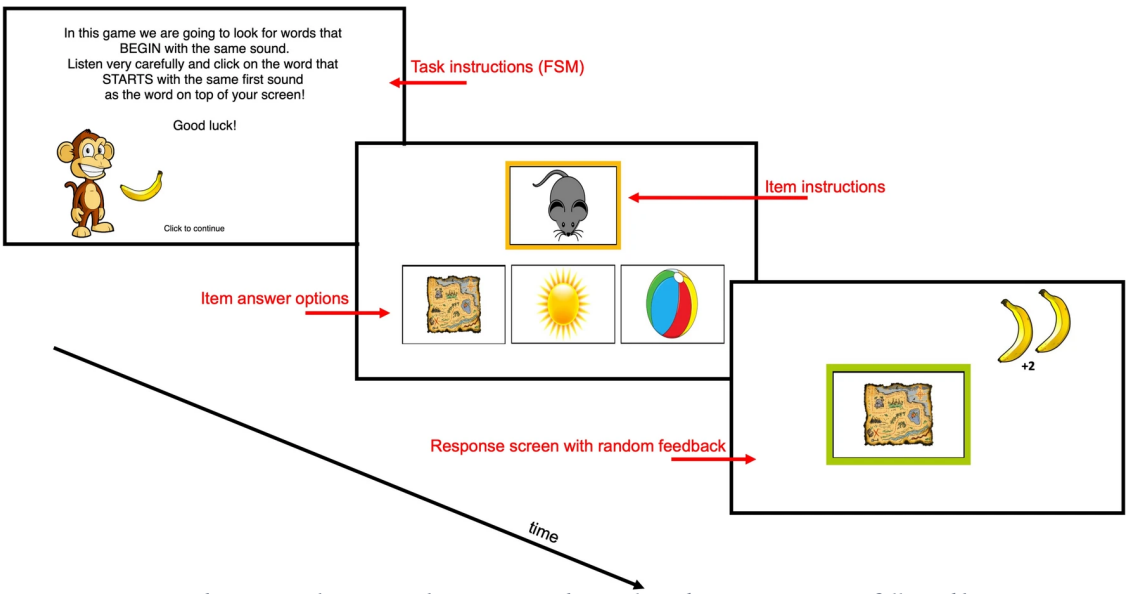


Figure 7.1: ROAR-Phoneme task. Every subtest starts with visual+auditory instructions, followed by some practice items with feedback. After the practice items, 19 items are presented in a one-interval three-alternative forced choice (1I-3AFC) task with random feedback (independent of answer, as motivation). The items are presented in semi-random order (within every block) and there are 2–3 sections per subtest.

References

Anthony, Jason L, and David J Francis. 2005. "Development of Phonological Awareness." *Curr. Dir. Psychol. Sci.* 14 (5): 255–59.

Gijbels, Liesbeth, Amy Burkhardt, Wanjing Anya Ma, and Jason D Yeatman. 2024. "Rapid Online Assessment of Reading and Phonological Awareness (ROAR-PA)." *Sci. Rep.* 14 (1): 1–16.

Moyle, Maura Jones, John Heilmann, and S Sue Berman. 2013. "Assessment of Early Developing Phonological Awareness Skills: A Comparison of the Preschool Individual Growth and Development Indicators and the Phonological Awareness and Literacy Screening–PreK." *Early Educ. Dev.* 24 (5): 668–86.

Nittrouer, Susan, Michael Studdert-Kennedy, and Richard S McGowan. 1989. "The Emergence of Phonetic Segments: Evidence from the Spectral Structure of Fricative-Vowel Syllables Spoken by Children and Adults." *J. Speech Lang. Hear. Res.* 32 (1): 120–32.

Stanovich, Keith E. 2017. "Speculations on the Causes and Consequences of Individual Differences in Early Reading Acquisition." In *Reading Acquisition*, 1st Edition, 307–42. Routledge.

Treiman, Rebecca, and Andrea Zukowski. 1991. "Levels of Phonological Awareness." *Phonological Processes in Literacy: A Tribute to Isabelle Y. Liberman.*

Tunmer, William E, Michael L Herriman, and Andrew R Nesdale. 1988. "Metalinguistic Abilities and Beginning Reading." *Read. Res. Q.* 23 (2): 134.

8 Letter Sound Knowledge (ROAR-Letter)

ROAR-Letter measures knowledge of upper-case and lower-case letter names and sounds. ROAR-Letter is designed to run as a computer adaptive test (CAT) to assess where a student is on the continuum of letter sound knowledge. ROAR-Letter can also be used as a diagnostic measure to guide teaching as it returns specific information about the letter names and letter-sound correspondences that the student does and does not know.

8.1 Structure of the task

ROAR-Letter is a four alternative fourced choice (4AFC) task divided into 3 blocks: - Upper-case letter names (26 items) - Lower-case letter names (26 items) - Letter-sound correspondences (36 items) Like all ROAR measures, ROAR-Letter is lightly gamified. The task begins with instructions and practice trials with feedback until the student understands the game. Then, in each block, the student is presented with the name or sound of a letter and asked to select the correct letter from the four choices (see Figure 8.1).

8.2 Design and implementation of computer-adaptive letter-sound assessment

To determine the optimal ROAR-Letter CAT we first collected data in 4,041 students in kindergarten and first grade. Table 8.1 shows the demographics of the participants and Table 8.2 shows the characteristics of the schools that participated. 2,840 of these students were administered all 88 items (26 lower case letter names; 26 upper case letter names; 36 letter sound correspondences). Based on these data, we ran a CAT simulation to a) determine how these students would have responded under different item-selection criteria and b) choose the number of items and optimal CAT parameters to achieve reliable estimates of letter-sound knowledge in the fewest number of trials. Figure 18.1 shows the upper and lower bounds on reliability as a function of the number of items that a participant completes. Based on these data we were able to develop an extremely efficient and precise computer adaptive test of letter sound knowledge (see Chapter 18).

	N	%	% Missing
Female	673	23.43	51.18
Free or Reduced Lunch	547	19.05	56.86
Race/Ethnicity			

Hispanic Ethnicity	700	24.37	48.33
White	362	12.60	48.33
Black or African American	78	2.72	48.33
Asian	96	3.34	48.33
American Indian or Alaska Native	3	0.10	48.33
Hawaiian or Other Pacific Islander	12	0.42	48.33
Multiracial	26	0.91	48.33
Total	2872		

Table 8.1: Demographics of ROAR-Letter calibration sample.

	N
Median Students	441
Median Free or Reduced Lunch	247
Race/Ethnicity	
Median Hispanic Ethnicity	69
Median White	129
Median Black or African American	3
Median Asian	25
Median American Indian or Alaska Native	0
Median Hawaiian or Other Pacific Islander	1
Median Multiracial	794
Organization Type	
Public School	21
Charter School	2
Private School	0
Summer School/Tutor Program/Other	2

Table 8.2: Characteristics of participating schools in the calibration sample.

8.3 Scoring

ROAR-Letter is a four alternative forced choice (4AFC) task and items are scored as correct or incorrect (dichotomous scoring) by comparing the participant's response (mouse click/touchscreen indicating the chosen letter), to the correct answer. Each response is scored in real time. The participant's ability or *raw score* (θ) on ROAR-Letter is computed based on an item response theory (IRT) model. IRT puts ability (θ) on an interval scale meaning that scores can be compared over time and across grades. A CAT algorithm is used to optimize the precision and efficiency of ROAR-Letter (see Chapter 18)

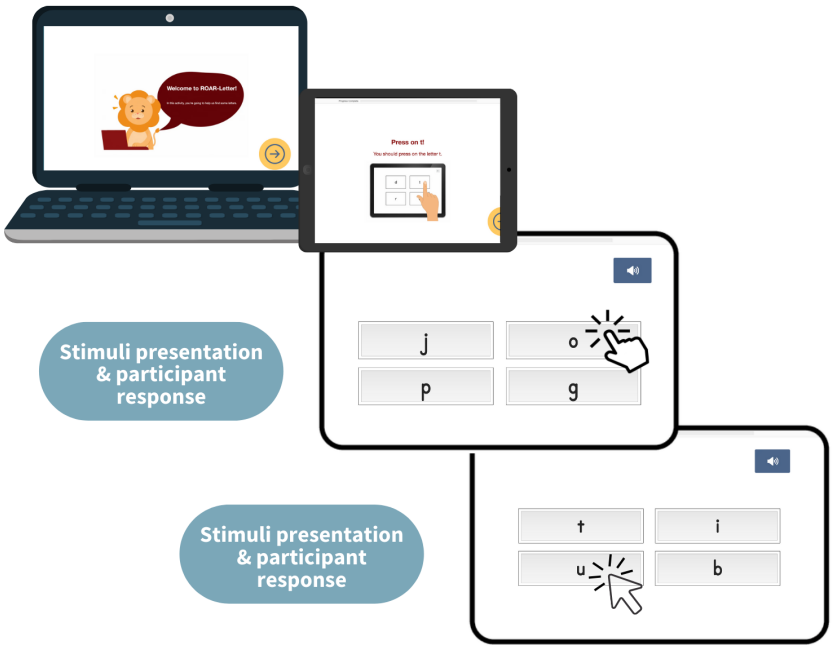


Figure 8.1: ROAR-Letter task. In the ROAR-Letter task, participants are first directed to choose the spoken letter name and then the letter sound in separate blocks. The task is divided into several blocks with encouragement throughout. Students have the opportunity to replay any letter name or sound audio they miss. There is no time limit for responses.

9 ROAR Dyslexia Screening and Subtyping

Beyond measures of reading skills, many state dyslexia screening initiatives require additional *predictors* of dyslexia. Dyslexia is defined in terms of reading skills (Maggie Snowling and Hulme 2024; O'Brien and Yeatman 2021; Catts et al. 2024; Lyon, Shaywitz, and Shaywitz 2003; Fletcher et al. 2006; Elliott and Grigorenko 2024), but there are two reasons why additional *predictors* are useful:

1. **Improved prediction accuracy**: Even though dyslexia is defined in terms of reading skills, that does not mean dyslexia begins in school. There are a collection of (potentially but not necessarily causal) factors that can be measured prior to formal reading instruction that predict future risk for dyslexia. There are also skills that can be measured at the earliest stages of reading instruction that are useful for prediction and early risk assessment. Thus, from a practical standpoint, including early predictors of dyslexia in a screener improves sensitivity and specificity (irrespective of the causal relationship between the predictors and reading development).
2. **Dyslexia subtyping and individualized intervention**: Dyslexia research has embraced multifactorial models where dyslexia is considered a probabilistic outcome that emerges from the combined influence of a collection of *risk factors* and *protective factors* (O'Brien and Yeatman 2021; Catts et al. 2024, 2024; B. F. Pennington 2006; Compton 2021; Zuk et al. 2021; Bergen, Leij, and Jong 2014; Wolf and Bowers 1999). Even though phonological awareness (PA) is still considered one of the most important mechanisms underlying dyslexia (Margaret Snowling 1998; M. J. Snowling, Hulme, and Nation 2020; Wagner and Torgesen 1987; Stanovich 1998), the field of dyslexia research has broadly reached consensus that a collection of mechanisms beyond difficulties with PA confer risk for dyslexia. This new understanding of dyslexia embraces heterogeneity: not every student's struggles are the same and knowing the root of a student's struggles could help plan the most efficacious intervention. However, currently, the notion of personalized intervention for particular dyslexia subtypes is more of an aspiration than a reality. One of the challenges is that the measures that are useful for phenotyping are not necessarily the most effective intervention targets. For example measures of Rapid Automatized Naming (RAN) and Rapid Visual Processing are prime examples: even though a wealth of research has established that these measures are useful for prediction, and can identify different profiles of struggling readers, it has not been established how intervention should differ based on these profiles.

We organize the ROAR Dyslexia Screening and Subtyping battery into screening measures that target:

1. **Foundational Reading Skills** that indicate dyslexia Section 9.1 and are established intervention targets
2. **Dyslexia Prediction and Subtyping** which includes measures of various mechanisms that are hypothesized to be causally related to dyslexia, improve screening sensitivity and specificity, but aren't established intervention targets Section 9.2.

ROAR has been validated as a dyslexia screener beginning in the Spring of kindergarten. Many of the individual measures are valid earlier (as young as 4 years of age (Gijbels et al. 2024)), but validation of ROAR dyslexia screening measures in the Fall of kindergarten is ongoing.

9.1 Dyslexia screening based on foundational reading skills

People with dyslexia struggle learning how to read. The most direct way to screen (or diagnose) dyslexia is based on measures of foundational reading skills. Children with dyslexia are delayed relative to their peers in the development of: (1) Phonological Awareness (see Chapter 7), (2) Letter Sound Knowledge (see Chapter 8), (3) real word and pseudo word reading (see Chapter 5), and (4) reading speed, efficiency or fluency (see Chapter 6). Difficulties with all these skills can persist through adulthood without proper identification and intervention. The goal of screening based on foundational reading skills is to a) identify challenges early in in elementary school and b) intervene while the developing brain is optimally plastic and interventions are most effective (Gaab and Petscher 2022; Blachman 2013; Torgesen 2004, 1998; Lovett et al. 2017). For example, Lovett et al. (2017) has shown that intervening early is more efficient (larger effects per hour of intervention) compared to waiting until later in elementary school. To quote Torgesen (1998), the goal of early screening is to "catch them before they fall".

9.2 Dyslexia prediction and subtyping

Wolf and Bowers (1999) and colleagues first introduced the "double deficit hypothesis" as a response to the "core phonological deficit hypothesis" and demonstrated that a) some children with typical PA skills still struggle learning to read and b) many of these struggling readers have early challenges with Rapid Automatized Naming (RAN) (Denckla and Cutting 1999; Denckla and Rudel 1976; Wolf and Bowers 1999; Compton, DeFries, and Olson 2001; Wolf et al. 2002). Additionally, children who struggle with PA and RAN tend to have even larger challenges with reading than those who only struggle on one skill. RAN is now required by most dyslexia screening legislation. However, RAN is not a useful intervention target per se: whereas a child who struggles with PA will benefit from training targeting PA (Bradley L and Bryant P E 1983), a child struggling with RAN does not simply need to practice RAN. Each of the measures in the **ROAR Foundational Reading Skills** battery is a core component of reading development and a useful intervention target. Measures in the **Dyslexia Prediction and Subtyping** battery are predictive of reading development and often help to understand the mechanisms underlying a student's struggles, but are not proven intervention targets. For example, **RAN** (see Section 9.2.1) is highly predictive of future reading development, and also indicates

a different type of struggle than PA (i.e., automatization or connectivity between visual and verbal processing), but is not a skill that needs direct instruction. **Rapid Visual Processing** (see Section 9.2.2)is another measure that has mounting evidence of a causal relationship to reading development and has utility as a screener but is not a skill that should be directly taught (Ramamurthy, White, and Yeatman 2023a; Ramamurthy et al. 2024; Lobier and Valdois 2015; Marie Line Bosse, Tainturier, and Valdois 2007).

9.2.1 Rapid Automatized Naming (ROAR-RAN)

9.2.1.1 Structure of the task, administration and scoring

ROAR-RAN is the only ROAR measure that requires verbal responses. Thus, ROAR-RAN has unique considerations for administration and scoring. Whereas all other ROAR measures are specifically designed and validated to produce accurate and reliable results in a large group setting (e.g., a classroom) where dozens (or hundreds) of students are silently completing ROAR assessments at the same time, ROAR-RAN requires students to be assessed in a quiet and private space. RAN is specifically designed to measure naming speed. Since speed is the fundamental unit of the measure, and since naming must be done out loud, it is important that participants are in a space where they can speak rapidly and without distraction.

Similar to other ROAR measures, RAN is scored automatically and instructions are narrated by characters in a gamelike setting. RAN does not require a test administrator to administer or score the assessment though young students might need some monitoring. Students should be in a private space where they can speak out loud without distractions and, akin to other ROAR measures, they log in to the dashboard and click to launch the ROAR-RAN assessment. Upon launch, a character will narrate the instructions. The participant will first be cued to name each individual item - letters, numbers or colors. This both serves as a check to ensure that the participant knows the name of each item and also to calibrate the automated scoring algorithm. After the quick calibration phase, the participant is the instructed to name all the items in sequence as quickly as possible. A countdown indicates the beginning of the measure, the stimuli are presented, and the participant responses are recorded through the webcam microphone. The webcam is also used to track the participants gaze so that the speech data can be co-registered to the item the participant is fixating, allowing for more precise scoring of individual items.

Scoring is performed automatically by the ROAR-RAN automated scoring algorithm. The algorithm processes the speech data, records the timestamp and duration for each spoken item, measuring both the speed and accuracy of the participant's responses. If a participant incorrectly identifies more than three items, the test is invalidated. For valid tests, the system calculates the total time taken to complete the task by recording the timestamp of the participant's response to the first symbol and the timestamp of their response to the last symbol. This total duration, measured in seconds, is then reported as the participant's score on the ROAR-RAN assessment (see Section 19.1 for validation of the scoring algorithm).

> **! A Webcam is Required for ROAR-RAN**
>
> Since RAN is fundamentally about naming, it is the **only ROAR assessment** that requires a webcam. Responses are recorded, securely stored, scored with an algorithm, and scores are displayed in the ROAR score report.

9.2.1.2 RAN-Letters

RAN is intended to measure naming that is "automatized". To select the ideal upper case letters for RAN-Letters we assessed knowledge of upper case letter names in 4,022 kindergarten and first grade students and calculated the proportion of students that knew the name of each upper case letter name. For the RAN-Letter stimuli, we chose the 6 monosyllabic letter names with a) the highest accuracy and b) phonetically distinct: **X**, **A**, **O**, **Z**, **F**, **M**

9.2.1.3 RAN-Colors

For young children who have not yet established sufficient letter knowledge for RAN-Letters, RAN-Colors is more appropriate. In RAN-Colors, rather than upper case letters, the stimuli are an array of color patches. The task is the same: the participant sequentially names all the colors as quickly and accurately as possible. RAN-Colors uses the following colors which have distinct and unambiguous names: Red, Blue, Pink, Black, Yellow. RAN-Colors does not use Green due to the high occurance of Red/Green color blindness.

9.2.1.4 RAN-Numbers

Another option for young children who have not yet established knowledge of letter names is RAN-Numbers. The following numbers are used as stimuli in RAN-Numbers: 2, 3, 4, 5, 7, 8.

9.2.2 Rapid Visual Processing

9.2.2.1 Theoretical background

There is a long-standing controversy about the visual factors associated with dyslexia. While phonological difficulties have been widely recognized as a core feature of dyslexia, debate has persisted over visual processing theories, suggesting a more complex, multifaceted etiology (R. Pennington and Pennington 2011; Vellutino et al. 2004; O'Brien and Yeatman 2021). The field has yet to reach consensus because existing theory on visual processing differences in dyslexia is often built upon small samples, using experimental measures without established psychometric properties, that are deployed across different age ranges often without prior work demonstrating the validity of the measure in each developmental window. Validated measures of visual

processing that predict future reading development could contribute in overcoming the challenges faced by other screening measures since visual development is language-agnostic and not directly taught in preschool.

Rapid Visual Processing (RVP) refers to the ability to rapidly encode and recall multiple visual elements simultaneously in a brief glimpse (Sperling 1960, 1983). In this task, a string of letters or symbols is briefly flashed at the center of the screen and then the participant is cued to report the identity of a single randomly chosen element. Of all the measures of sensory processing that have been studied in relation to word reading difficulties, the rapid visual processing task has the strongest evidence for identifying a subgroup of struggling readers who are not captured by conventional measures of phonological awareness. Previous studies have demonstrated that this task:

1. Consistently correlates with reading ability (Marie-Line Bosse and Valdois 2009; Ramamurthy, White, and Yeatman 2023b).
2. Differs in children with dyslexia (Lobier, Zoubrinetzky, and Valdois 2012; Ramamurthy, White, and Yeatman 2023b).
3. Cannot be explained as a consequence of dyslexia (Lobier and Valdois 2015).
4. Might be a useful intervention target (Valdois et al. 2014; Zhao et al. 2019; Zoubrinetzky et al. 2019).
5. Identifies a subset of poor readers that have high phonological awareness despite their reading difficulties (Saksida et al. 2016; Valdois et al. 2020).
6. Recent evidence suggests that the difficulties with Rapid Visual Processing in children with dyslexia are consistent across languages including French, English, Dutch and Chinese (Huang, Liu, and Zhao 2021; Lobier and Valdois 2015) making this an ecologically relevant measure related to reading across different languages.

By including both **letter** and **symbol** stimuli, the task allows for the assessment of rapid visual processing skills both within and independent of language experience. The measure's language-agnostic nature makes it particularly valuable for early screening, as visual processing issues precede reading difficulties. This approach, of including visual measures that are linked to reading, aligns with emerging multifactorial models of dyslexia, contributing to a more comprehensive understanding of the various cognitive factors influencing reading development (O'Brien and Yeatman 2021; Catts et al. 2024; B. F. Pennington 2006; Compton 2021; Zuk et al. 2021; Bergen, Leij, and Jong 2014; Wolf and Bowers 1999).

9.2.2.2 Structure of the task

The **Rapid Visual Processing Task** was developed in close collaboration with the UCSF Multitudes project[1]. Task design, data colletion and analysis was shared across the two projects.

ROAR-RVP has 2 versions that measure the same construct of rapid visual processing: Rapid Visual Processing with Letters (RVPL) and Rapid Visual Processing with Symbols (RVPS).

[1] https://multitudesinfo.ucsf.edu/

RVPL measures the ability to rapidly locate and identify letters in 2-, 4-, and 6-letter strings. RVPS assesses the ability to rapidly locate and identify non-namable visual symbols, making it language-agnostic. These tasks are considered promising tools for early identifying of struggling readers not captured by conventional phonological awareness measures.

- Letters (RVPL)
- Symbols (RVPS)

ROAR-RVP employs a six-alternative forced choice task presented as an engaging underwater adventure game. There are two versions, each with two difficulty blocks (2- and 4- element strings. 6-element string can be added for older students). Participants help a lost dolphin (RVPL) or whale (RVPS) find friends and treasures. The task sequence involves fixating on a central point where 2-4 elements (letters or pseudo-letters) briefly appear (240ms). A post-cue then indicates the target position, and participants select the correct element from six choices. Each version begins with 6 longer-duration practice items. Narrated instructions, encouragement animations, and feedback sounds guide participants throughout. The game is designed for K-2 children but can be extended to older students and adults.

Figure 9.1 shows the rapid visual processing task with letters (RVPL) a string of letters is briefly flashed at the center of the screen (240ms) and then the participant is cued to report the identity of a single randomly chosen letter. The participant's task is to report the identity of the target element that was at the prompted single location, by tapping on one letter from a set of 6 letter choices provided. This task was optimized to provide highly reliable metrics of rapid visual processing in children as young as five years of age (see Section 19.2)

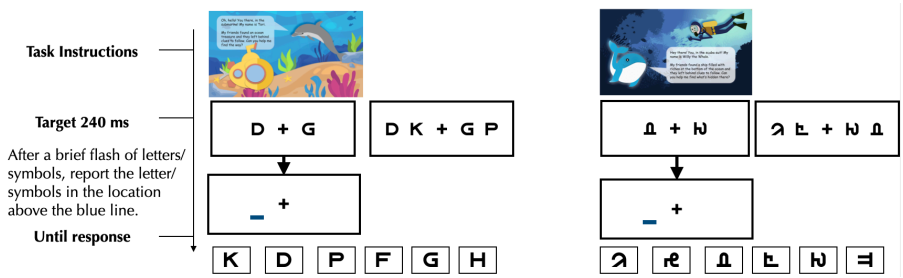


Figure 9.1: Schematic of the ROAR Rapid Visual Processing (ROAR-RVP) Task

9.2.2.3 *Are visual measures biased to social factors like one's socio-economic status and primary language?*

While different cognitive and language processes have been increasingly reported to be influenced by socioeconomic status and childhood experience Schwab and Lew-Williams (2016), in theory, visual processing measures should not be influenced by language, socioeconomic status or early childhood experience. This is because most sensory processes have a developmental trajectory that precedes Stein (2022) formal reading instructions, making them valuable screening measures for identifying potential future reading challenges.

We first examined whether the visual measures we developed exhibit bias related to students' primary language or socio-economic status as indexed by eligibility for free and reduced-price meals (see Figure 9.2). This investigation addresses a fundamental challenge in early screening: many measures show bias across various social factors. Consequently, even when measures demonstrate good predictive power for reading challenges, it often remains unclear whether this predictive ability stems from the measure assessing a construct fundamental to reading skill or simply from capturing the same differences in social factors that are known to influence reading outcomes. Our findings show that ROAR-RVP were not affected by English language proficiency or socioeconomic status, which shows promise for addressing a major challenge in the field.

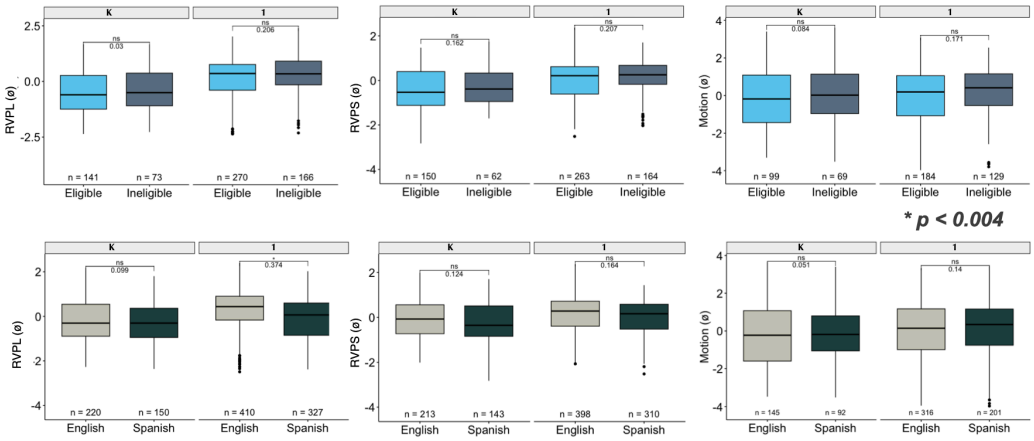


Figure 9.2: ROAR Rapid Visual Processing is not affected by spoken language or socioeconomic status

Part II

Introduction to ROAR-Español

References

Bergen, Elsje van, Aryan van der Leij, and Peter F de Jong. 2014. "The Intergenerational Multiple Deficit Model and the Case of Dyslexia." *Frontiers in Human Neuroscience* 8: 346.

Blachman, Benita A. 2013. *Foundations of Reading Acquisition and Dyslexia: Implications for Early Intervention*. Routledge.

Boets, Bart, Jan Wouters, Astrid van Wieringen, Bert De Smedt, and Pol Ghesquière. 2008. "Modelling Relations Between Sensory Processing, Speech Perception, Orthographic and Phonological Ability, and Literacy Achievement." *Brain Lang.* 106 (1): 29–40.

Bosse, Marie Line, Marie Josèphe Tainturier, and Sylviane Valdois. 2007. "Developmental dyslexia: The visual attention span deficit hypothesis." *Cognition* 104 (2): 198–230.

Bosse, Marie-Line, and Sylviane Valdois. 2009. "Influence of the Visual Attention Span on Child Reading Performance: A Cross-Sectional Study." *J. Res. Read.* 32 (2): 230–53.

Bradley L, and Bryant P E. 1983. "Categorizing sounds and learning to read – a causal connection." *Nature* 301 (3): 419–21.

Bronfenbrenner, U, and P Morris. 2007. "The Bioecological Model of Human Development." *Handbook of Child Psychology*, June, 793–828.

Catts, Hugh W, Nicole Patton Terry, Christopher J Lonigan, Donald L Compton, Richard K Wagner, Laura M Steacy, Kelly Farquharson, and Yaacov Petscher. 2024. "Revisiting the Definition of Dyslexia." *Ann. Dyslexia*, January.

Compton, Donald L. 2021. "Focusing Our View of Dyslexia Through a Multifactorial Lens: A Commentary." *Learning Disability Quarterly* 44 (3): 225–30.

Compton, Donald L, John C DeFries, and Richard K Olson. 2001. "Are RAN-and Phonological Awareness-Deficits Additive in Children with Reading Disabilities?" *Dyslexia* 7 (3): 125–49.

Denckla, Martha Bridge, and Laurie E Cutting. 1999. "History and Significance of Rapid Automatized Naming." *Annals of Dyslexia* 49: 29–42.

Denckla, Martha Bridge, and Rita G Rudel. 1976. "Rapid 'Automatized'naming (RAN): Dyslexia Differentiated from Other Learning Disabilities." *Neuropsychologia* 14 (4): 471–79.

Elliott, Julian G, and Elena L Grigorenko. 2024. "Dyslexia in the Twenty-First Century: A Commentary on the IDA Definition of Dyslexia." *Ann. Dyslexia*, June.

Fletcher, J M, G R Lyon, L S Fuchs, and M A Barnes. 2006. *Learning disabilities: From identification to intervention*. New York: Guilford PRess.

Gaab, Nadine, and Yaacov Petscher. 2022. "Screening for Early Literacy Milestones and Reading Disabilities: The Why, When, Whom, How, and Where." *Perspectives on Language and Literacy* 48 (1): 11–18.

Gijbels, Liesbeth, Amy Burkhardt, Wanjing Anya Ma, and Jason D Yeatman. 2024. "Rapid Online Assessment of Reading and Phonological Awareness (ROAR-PA)." *Sci. Rep.* 14 (1): 1–16.

Huang, Chen, Ningyu Liu, and Jing Zhao. 2021. "Different Predictive Roles of Phonological Awareness and Visual Attention Span for Early Character Reading Fluency in Chinese." *J. Gen. Psychol.* 148 (1): 45–66.

Lobier, Muriel, and Sylviane Valdois. 2015. "Visual attention deficits in developmental dyslexia cannot be ascribed solely to poor reading experience." *Nat. Rev. Neurosci.* 16 (4): 1.

Lobier, Muriel, Rachel Zoubrinetzky, and Sylviane Valdois. 2012. "The visual attention span deficit in dyslexia is visual and not verbal." *Cortex* 48 (6): 768–73.

Lovett, Maureen W, Jan C Frijters, Maryanne Wolf, Karen A Steinbach, Rose A Sevcik, and Robin D Morris. 2017. "Early Intervention for Children at Risk for Reading Disabilities: The Impact of Grade at Intervention and Individual Differences on Intervention Outcomes." *Journal of Educational Psychology* 109 (7): 889.

Lyon, G Reid, Sally E Shaywitz, and Bennett A Shaywitz. 2003. "A Definition of Dyslexia." *Ann. Dyslexia* 53 (1): 1–14.

O'Brien, Gabrielle, and Jason D Yeatman. 2021. "Bridging Sensory and Language Theories of Dyslexia: Toward a Multifactorial Model." *Dev. Sci.* 24 (3): e13039.

Pennington, Bruce F. 2006. "From Single to Multiple Deficit Models of Developmental Disorders." *Cognition* 101

(2): 385–413.
Pennington, Rob, and Robert E Pennington. 2011. *Find the Upside of the down Times: How to Turn Your Worst Experience*. Resource International.
Ramamurthy, Mahalakshmi, Klint Kanopka, Adam Richie-Halford, Benjamin Domingue, Francesca Pei, Phaedra Bell, Lucy Yan, Andrea Hartsough, Maria L Gorno-Tempini, and Jason D Yeatman. 2024. "Design and Validation of a Rapid Visual Processing Measure for Screening Reading Difficulties in Early Childhood," February.
Ramamurthy, Mahalakshmi, Alex L White, and Jason D Yeatman. 2023a. "Children with Dyslexia Show No Deficit in Exogenous Spatial Attention but Show Differences in Visual Encoding." *Dev. Sci.*, November, e13458.
Ramamurthy, Mahalakshmi, Alex White, and Jason D Yeatman. 2023b. "Children with Dyslexia Show No Deficit in Exogenous Spatial Attention but Show Differences in Visual Encoding."
Saksida, Amanda, Stéphanie Iannuzzi, Caroline Bogliotti, Yves Chaix, Jean François Démonet, Laure Bricout, Catherine Billrd, et al. 2016. "Phonological skills, visual attention span, and visual stress in developmental dyslexia." *Dev. Psychol.* 52 (10): 1503–16.
Schwab, Jessica F, and Casey Lew-Williams. 2016. "Language Learning, Socioeconomic Status, and Child-Directed Speech." *Wiley Interdiscip. Rev. Cogn. Sci.* 7 (4): 264–75.
Snowling, Maggie, and Charles Hulme. 2024. "Do We Really Need a New Definition of Dyslexia? A Commentary." *Ann. Dyslexia*, March.
Snowling, Margaret. 1998. "Dyslexia as a Phonological Deficit: Evidence and Implications." *Child Psychology and Psychiatry Review* 3 (1): 4–11.
Snowling, Margaret J, Charles Hulme, and Kate Nation. 2020. "Defining and Understanding Dyslexia: Past, Present and Future." *Oxford Review of Education* 46 (4): 501–13.
Sperling, George. 1960. "The Information Available in Brief Visual Presentations." *Psychological Monographs: General and Applied* 74 (11): 1.
———. 1983. "Why We Need Iconic Memory." *Behav. Brain Sci.* 6 (1): 37–39.
Stanovich, Keith E. 1998. "Refining the Phonological Core Deficit Model." *Child Psychology and Psychiatry Review* 3 (1): 17–21.
Stein, John. 2022. "The Visual Basis of Reading and Reading Difficulties." *Front. Neurosci.* 16 (November): 1004027.
Taylor, Ellie K, Gavkhar Abdurokhmonova, and Rachel R Romeo. 2023. "Socioeconomic Status and Reading Development: Moving from 'Deficit' to 'Adaptation' in Neurobiological Models of Experience-Dependent Learning." *Mind Brain Educ.* 17 (4): 324–33.
Torgesen, Joseph K. 1998. "Catch Them Before They Fall." *American Educator* 22: 32–41.
———. 2004. "Preventing Early Reading Failure." *American Educator* 28 (3): 6–9.
Valdois, Sylviane, Carole Peyrin, Delphine Lassus-Sangosse, Marie Lallier, Jean-François Démonet, and Sonia Kandel. 2014. "Dyslexia in a French–Spanish Bilingual Girl: Behavioural and Neural Modulations Following a Visual Attention Span Intervention." *Cortex* 53 (April): 120–45.
Valdois, Sylviane, Caroline Reilhac, Emilie Ginestet, and Marie Line Bosse. 2020. "Varieties of Cognitive Profiles in Poor Readers: Evidence for a VAS-Impaired Subtype." *J. Learn. Disabil.*, September, 22219420961332.
Vellutino, Frank R, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. 2004. "Specific Reading Disability (Dyslexia): What Have We Learned in the Past Four Decades?" *J. Child Psychol. Psychiatry* 45 (1): 2–40.
Wagner, Richard K, and Joseph K Torgesen. 1987. "The Nature of Phonological Processing and Its Causal Role in the Acquisition of Reading Skills." *Psychological Bulletin* 101 (2): 192.
Wolf, Maryanne, and Patricia Greig Bowers. 1999. "The Double-Deficit Hypothesis for the Developmental Dyslexias." *Journal of Educational Psychology* 91 (3): 415.
Wolf, Maryanne, Alyssa Goldberg O'rourke, Calvin Gidney, Maureen Lovett, Paul Cirino, and Robin Morris. 2002. "The Second Deficit: An Investigation of the Independence of Phonological and Naming-Speed Deficits in Developmental Dyslexia." *Reading and Writing* 15: 43–72.
Zhao, Jing, Hanlong Liu, Jiaxiao Li, Haixia Sun, Zhanhong Liu, Jing Gao, Yuan Liu, and Chen Huang. 2019. "Improving Sentence Reading Performance in Chinese Children with Developmental Dyslexia by Training

Based on Visual Attention Span." *Sci. Rep.* 9 (1): 18964.
Zoubrinetzky, Rachel, Gregory Collet, Marie-Ange Nguyen-Morel, Sylviane Valdois, and Willy Serniclaes. 2019. "Remediation of Allophonic Perception and Visual Attention Span in Developmental Dyslexia: A Joint Assay." *Front. Psychol.* 10 (July): 1502.
Zuk, Jennifer, Jade Dunstan, Elizabeth Norton, Xi Yu, Ola Ozernov-Palchik, Yingying Wang, Tiffany P Hogan, John DE Gabrieli, and Nadine Gaab. 2021. "Multifactorial Pathways Facilitate Resilience Among Kindergarteners at Risk for Dyslexia: A Longitudinal Behavioral and Neuroimaging Study." *Developmental Science* 24 (1): e12983.

10 Multilingualism

A widely accepted definition of *multilingualism* (often used interchangeably with the term *bilingualism*) describes "anyone who can communicate in more than one language, be it active (through speaking and writing) or passive (through listening and reading)" as multilingual (Wei 2008, 4). Moreover, multilingual individuals are a very heterogeneous group, with differences in, for example, the ages of acquisition, proficiency, contexts of use, and degree of balancedness between their languages. This, in turn, means that multilingualism is a very individual phenomenon (Blumenfeld, Bobb, and Marian 2016; Grosjean 2008).

Most importantly, multilingual individuals should not be viewed as "two monolinguals in one person" (Grosjean 1989, 4). This implies that their performance on a test measuring a given construct must not be interpreted using monolingual norms in either of their languages without appropriate considerations and qualifications. While performing at or above monolingual norms might indicate the absence of reading difficulties Section 10.3, the inverse is not true: If multilingual students do not meet monolingual norms, this must not be interpreted as an indication of reading difficulty, because multilingual students follow distinct developmental trajectories in their languages.

10.1 Prevalence

Based on responses to the mandatory Home Language Survey, about 39.5 % (2,310,311) of students enrolled in the Californian public school system speak a language other than English at home (California Department of Education[1], 2022). About 19.1 % (1.3 million) of Californian students are classified as English Learners (ELs), meaning that they did not meet the criteria, based on the English Language Proficiency Assessment of California (ELPAC), to be (re)classified as English-proficient. While these students speak more than 100 different languages, 81.9 % are Spanish-speakers, highlighting the need for universal screening instruments in Spanish, such as ROAR-Español.

10.2 Choosing the Language(s) of Assessment

Determining the appropriate language(s) of assessment for a multilingual student is challenging. Factors to consider are the languages reported to be spoken at home, languages of previous and current instruction, as well as the outcome of interest. Currently, we offer ROAR-English and ROAR-Español as two standalone suites of tests (Italian, Portuguese, German and French are

[1] https://www.cde.ca.gov/ds/ad/cefelfacts.asp

in development). However, we are continuously exploring ways of combining multilingual students' scores in multiple languages so that we will be able to provide a unified reading risk estimation for multilingual students.

In the meantime, it is best practice to assess students in all their languages. For Spanish-speaking ELs that means administering both ROAR-English and ROAR-Español and qualitatively interpreting a student's results in both languages in conjunction with other available information on, for example, their home language environment and prior languages of instruction. Multilingual students meeting or exceeding standards derived from monolingual populations may generally be considered as meeting those standards. However, in cases where a multilingual student performs below proficiency thresholds, this may be due to a number of reasons (the ongoing acquisition of the language of assessment, different expected developmental trajectories, etc.) and one should not conclude that this result indicates a screening flag. For example, consider a 2nd grade student who grew up in a home that primarily spoke Spanish, began learning to speak English when they entered Kindergarten, and was primarily taught to read in English. Their scores on ROAR-English and ROAR-Spanish would both be useful for gauging their reading development and making planning instruction, but we would not expect their ROAR-Español scores to be at the same level as a monolingual Spanish speaker who is being taught to read in Spanish. ROAR scores provide detailed information about reading skills in each language of assessment but must be interpreted in context alongside other sources of information.

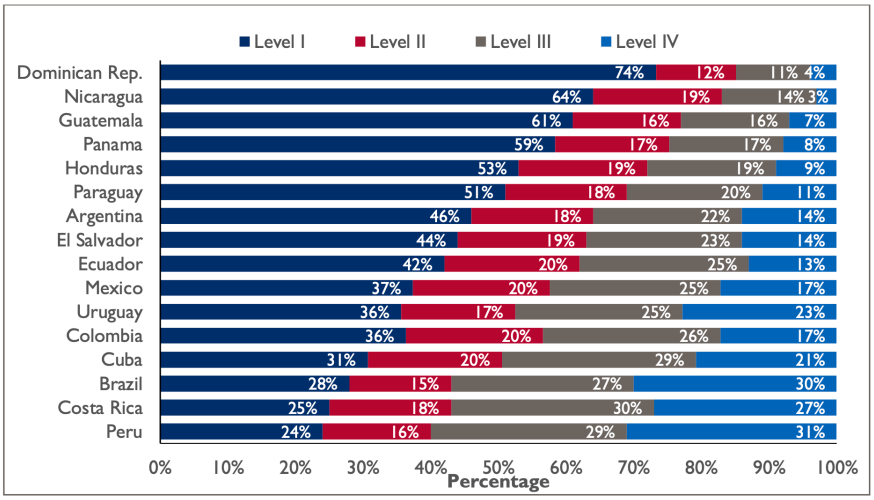
10.3 ROAR-Español Scores

ROAR-Español has been validated in a sample of multilingual learners in California and a sample of (primarily) monolingual Spanish speakers in Colombia (see Bhat et al. (2024) for an initial publication based on ROAR-Español). Care has been taken in the design and validation of each ROAR-Español measure to design the items around the linguistic diversity of Spanish speakers. ROAR-Español returns the same types of scores as English ROAR measures (see Section 3.1 and ROAR Families and Teachers Guide[2] for detailed information) and, additionally, ROAR-Español returns grade level equivalent scores based on the normative data from monolingual Spanish speakers in Colombia. Every country will, of course, have different normative trajectories of reading development. Literacy levels among school aged children in Colombia is similar to Mexico, but below average compared to other countries in Latin America[3]. Figure 10.1 shows literacy achievement data across different countries in Latin America (reproduced from US AID A SUMMARY ANALYSIS OF EDUCATION TRENDS IN LATIN AMERICA AND THE CARIBBEAN 2022 UPDATE[4]). The large, urban, school districts we partnered with in Colombia perform above that national average and could be considered reasonably representative of a typical learning trajectory for a (primarily) monolingual Spanish speaker in an urban area of Latin America. Moreover, the representative data we collected in California schools places Spanish speaking students from California within the typical range

[2] documents/ROAR-Family-Guide-20240814.pdf
[3] documents/USAID-Education-Trends-LA.pdf
[4] https://pdf.usaid.gov/pdf_docs/pa00zk72.pdf

of the Colombian norms. Thus, ROAR-Español Grade Level Equivalent (GLE) Scores can provide useful information to teachers in the United States about a student's Spanish reading proficiency relative to monolingual learners in Latin America. These data can be useful for interpreting scores on an English screener. For example, a new second grade student who has just moved from a monolingual Spanish environment to the US who has low scores on ROAR measures in English but with GLE Scores on ROAR-Español that are in the typical range for a 2nd grader still needs additional reading instruction, but is not at high risk for dyslexia.

GRAPH 11: DISTRIBUTION OF ACHIEVEMENT LEVELS ON ERCE READING TEST (3RD GRADE STUDENTS)



Notes: Level II is the minimum level of performance on 3rd grade reading test. Level III is the minimum level of performance on 6th grade reading test

Source: UNESCO ERCE 2019

Figure 10.1: Literacy achievement levels in Latin America

10.4 Vignettes: Interpretation and Use of Multilingual Students' Scores

10.4.1 Vignette 1

A student who grew up and attended (pre-)school in Mexico arrives in the US with his family in the summer before starting second grade. The family reports speaking Spanish at home and that their child had attended two years of prior instruction in Spanish. In this case, knowing how this student fares in relation to his monolingual Spanish peers is of value, as this is a fair comparison–the student has, hitherto, lived in a linguistic environment and had a scholastic experience that is represented by the largely monolingual Colombian norming sample (Colombia and Mexico also have comparable literacy achievement levels. See Figure 10.1). This student will not have had many opportunities to learn English. In this case, even if the student is deemed to have met the minimal requirements to be assessed in English, their performance on any English assessment is not particularly meaningful, as it will be a function of their lacking exposure to the English language, rather than of any underlying reading difficulty.

Thus, if this students' English score suggest that they are in need of additional support, this result should *not* be interpreted in isolation. Rather, one ought to obtain the student's ROAR-Español scores. If the student's ROAR-Español grade level equivalent score classifies them as performing at grade-level/on-track, then one can conclude that the student is not at risk of developing a reading difficulty/dyslexia, but instead is in need of exposure to and structured support with learning English. They will likely also need support in translating their Spanish reading ability to English since English is a highly irregular writing system (i.e., an opaque orthography). But needing instruction is qualitatively different than being at risk for long-term reading difficulties like dyslexia. If the student's ROAR-Español grade level equivalent score results in a 'needing support'/not-on-track classification, further assessment is recommended.

10.4.2 Vignette 2

An incoming first-grade student grows up in a predominantly Spanish-speaking family in California. While the parents were born in Mexico and had moved to the US in their adult life, the student was born in the US, lived in a linguistically diverse neighborhood where both English and Spanish are spoken, and attended pre-school and kindergarten classes where English was the language of instruction. In this case, though the student only hears and speaks Spanish at home, their instructional experience and formal language instruction was exclusively English.

Here, we cannot reasonably expect the student to perform similarly to monolingual Spanish speakers on the ROAR-Español because their linguistic and educational experiences were spread out across multiple languages. Therefore, their grade level equivalent score is much less meaningful and no "risk" classification should be made on the basis of their Spanish performance, alone. At the same time, the students cannot be fairly compared to monolingual English speakers on English assessments. Ultimately, testing in both languages is strongly recommended and, due to multilingual individuals heterogeneity and distinct developmental trajectories, the student's performance is best judged in conjunction with other qualitative information available. This vignette serves to highlight the challenges in assessment of multilingual learners. Even though the goal of screeners is often to produce a simple and interpretable "risk metric", interpreting assessment data in multilingual learners requires nuance and expertise and should always take into account the learning context.

References

Bhat, Kruttika G., Alexa Mogan, Ana Saavedra, Mia Fuentes-Jimenez, Julian M. Siebert, Wanjing Anya Ma, Carrie Townley-Flores, et al. 2024. "Shared and Unique Influences of Phonological Processing on Reading and Math[a]."

Blumenfeld, Henrike K, S C Bobb, and Viorica Marian. 2016. "The role of language proficiency, cognate status and word frequency in the assessment of Spanish–English bilinguals' verbal fluency." *International Journal of Speech-Language Pathology* 18 (2): 190–201.

Grosjean, François. 1989. "Neurolinguists, beware! The bilingual is not two monolinguals in one person." *Brain and Language* 36 (1): 3–15. https://doi.org/10.1016/0093-934x(89)90048-5.

———. 2008. *Studying Bilinguals*. Oxford University Press. Oxford University Press.
Wei, Li. 2008. "Research perspectives on bilingualism and multilingualism." In, edited by Li Wei and Melissa G Moyer, 1–17. The Blackwell Guide to Research Methods in Bilingualism and Multilingualism. Blackwell Publishing.

---

[a]

11 Spanish Single Word Reading (ROAR-Palabra)

ROAR-Palabra uses a lexical decision task to measure single word reading ability in Spanish. To our knowledge, the use of lexical decision tasks as a proxy of single word reading ability in Spanish has not been investigated before and we are the first to propose such task for the purpose of a reading screener. We developed this task in response to the need for reading screening instruments in languages other than English (also see Chapter 10). While additional languages are to follow, we began developing version of our subtests in Spanish, because it is the most widely spoken language other than English in the United States (and California specifically). ROAR-Palabra is still under active development and we report on the current fixed-length (62 items) version of the task (i.e., a non-computer-adaptive) version. For the rationale behind using lexical decision tasks to approximate word reading in general, refer to Chapter 5. For a description of the structure of the task, see Section 5.1, particularly Figure 5.1.

Unlike English, Spanish is a highly phonologically transparent language, meaning that there is a close correspondence between spelling and pronunciation. This transparency requires careful selection of non-words, as they must adhere to Spanish phonotactic rules to avoid creating stimuli that are easily identifiable as non-words based solely on their orthographic structure. Additionally, Spanish morphology is more complex than English, with a rich system of verb conjugations and gendered noun-adjective agreements. Therefore, attention must be given to the morphological structure of the stimuli to ensure that the task captures genuine lexical decision processes rather than responses driven by morphological cues.

11.1 Task Development

ROAR-Palabra is explicitly not a translation of the ROAR-Word— simple translations usually fail to produce equivalent versions of a test (Solano-Flores, Backhoff, and Contreras-Niño 2009). In contrast to many other non-English measures, we started the development process from a Spanish perspective: We created an initial list of stimuli by prompting ChatGPT to produce a list of frequent Spanish words, known to pre- and middle-schoolers across the Americas and occurring in all the varieties of Spanish spoken there. We then then used the Wuggy algorithm (Keuleers and Brysbaert 2010) to create matching, word-like pseudowords—stimuli conforming to Spanish orthographic rules and matching the real word list in terms of word length, letter-transition frequencies, and orthographic neighbourhood size. Candidate items were then inspected by experts in reading development who speak various regional varieties of Spanish. The goal was to create an assessment that would be equitable across the wide variety of Spanish language varieties that are spoken in the United States. Spanish speakers from Chile, Colombia, Mexico, and Spain independently reviewed both the real and pseudowords. Items

flagged as problematic due to, for example, low frequency of occurrence or inappropriate slang meanings in any one of the Spanishes were removed.

This process resulted in an initial item bank with 378 item pairs (that is 189 real words and 189 matched pseudowords). To keep administration time reasonable, we randomly selected 70 core items (35 real and 35 pseudowords), which form the basis of the current version. Every test-taker responded to these core items, as well as 30 additional items (15 real and 15 pseudowords) randomly selected from the broader item pool. In future versions, these additional items will be calibrated, too, so that the task can be made computer-adaptive.

11.1.1 Californian Calibration Sub-sample

To account for regional variations in the Spanish language, as well as for the different linguistic profiles of monolingual and multilingual speakers of Spanish (see Chapter 10), we developed the ROAR-Palabra using data from both a native Spanish-speaking sample in Chile, as well as a multilingual sample from a school district in California. This allowed us to ensure that the task is appropriate for use in Spanish-speaking populations with different linguistic and cultural backgrounds. Table 11.1 shows the demographic characteristics of our California sub-sample.

Table 11.1: ROAR-Palabra Calibration Sub-sample from California

Characteristic	**Kindergarten** N = 5	**Grade 1** N = 201	**Grade 2** N = 220
Female	3 (60%)	96 (48%)	81 (38%)
Unknown	0	3	4
Free/Reduced-price Lunch Eligibility	3 (60%)	140 (70%)	155 (70%)
Special Educational Needs	1 (20%)	21 (11%)	26 (12%)
Unknown	0	3	3
English Proficiency Status			
EL	2 (40%)	121 (61%)	145 (67%)
EO	2 (40%)	52 (26%)	49 (23%)
IFEP	1 (20%)	14 (7.1%)	12 (5.5%)
RFEP	0 (0%)	11 (5.6%)	10 (4.6%)
TBD	0 (0%)	0 (0%)	1 (0.5%)
Unknown	0	3	3
Primary Language			
English	2 (40%)	74 (41%)	93 (46%)
Spanish	3 (60%)	107 (59%)	111 (54%)
Unknown	0	20	16
Race			
Asian	0 (0%)	3 (1.5%)	0 (0%)
Black or African American	0 (0%)	3 (1.5%)	0 (0%)
Filipino	0 (0%)	1 (0.5%)	0 (0%)

Table 11.1: ROAR-Palabra Calibration Sub-sample from California

Characteristic	Kindergarten N = 5	Grade 1 N = 201	Grade 2 N = 220
Hawaiian or Other Pacific Islander	0 (0%)	1 (0.5%)	0 (0%)
Hispanic	3 (60%)	178 (90%)	195 (100%)
White	2 (40%)	12 (6.1%)	0 (0%)
Unknown	0	3	25

11.1.2 Response Time

Figure 11.1 shows the distribution of students' median response times on the ROARA-Palabra. Following the rationale outlined in Chapter 15), participants with a median response time <450ms were excluded from further analyses. This results in 374 of 6035 students (6.2 %) being excluded from further analyses due to guessing or other unreliable testing behaviours.

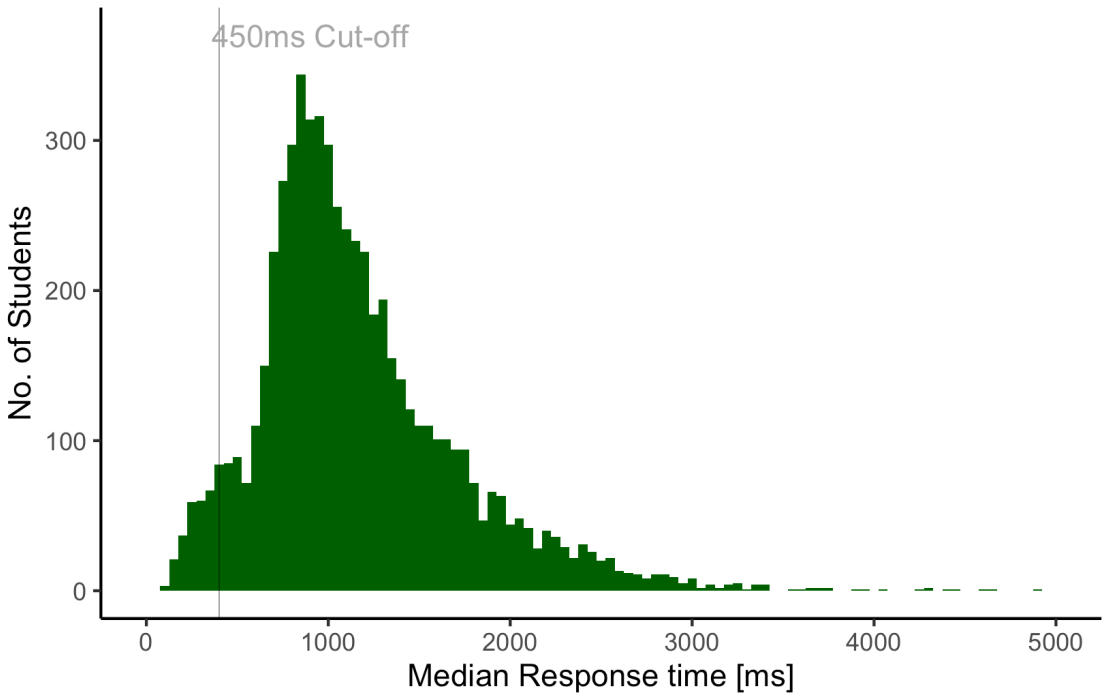


Figure 11.1: Distribution of median response times (both locations) with indication of cut-off (450ms).

11.1.3 Item Properties

Overall, items tended to be easier for students in Colombia; this is expected, as these students are native Spanish speakers instructed in their first language, while the Californian sub-sample is more linguistically diverse in terms of the Spanish abilities and their instructional programmes. Also, the Colombian sub-sample drawn on here consists of students from grades 1-

11. While this disallows for a meaningful direct comparison between Californian and Colombian students,

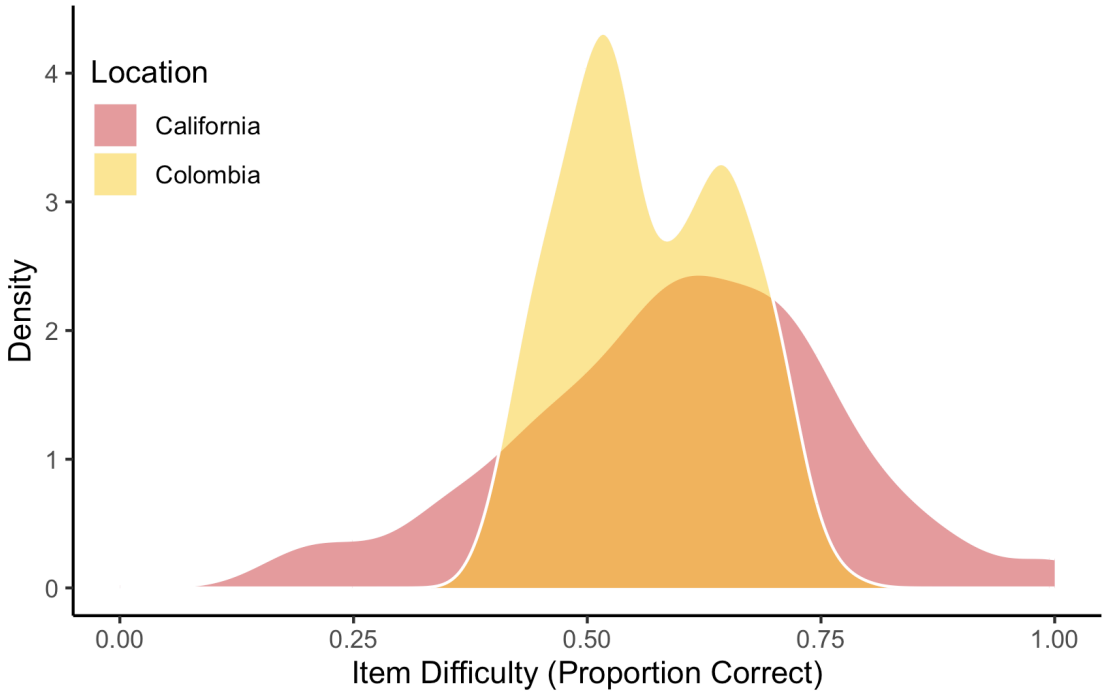


Figure 11.2: Distribution of item difficulty parameters, computed separately by location (California vs. Colombia) drawing only on students in grades K-2.

References

Keuleers, Emmanuel, and Marc Brysbaert. 2010. "Wuggy: A multilingual pseudoword generator." *Behavior Research Methods* 42 (3): 627–33. https://doi.org/10.3758/brm.42.3.627.

Solano-Flores, Guillermo, Eduardo Backhoff, and Luis Ángel Contreras-Niño. 2009. "Theory of Test Translation Error." *International Journal of Testing* 9 (2): 78–91. https://doi.org/10.1080/15305050902880835.

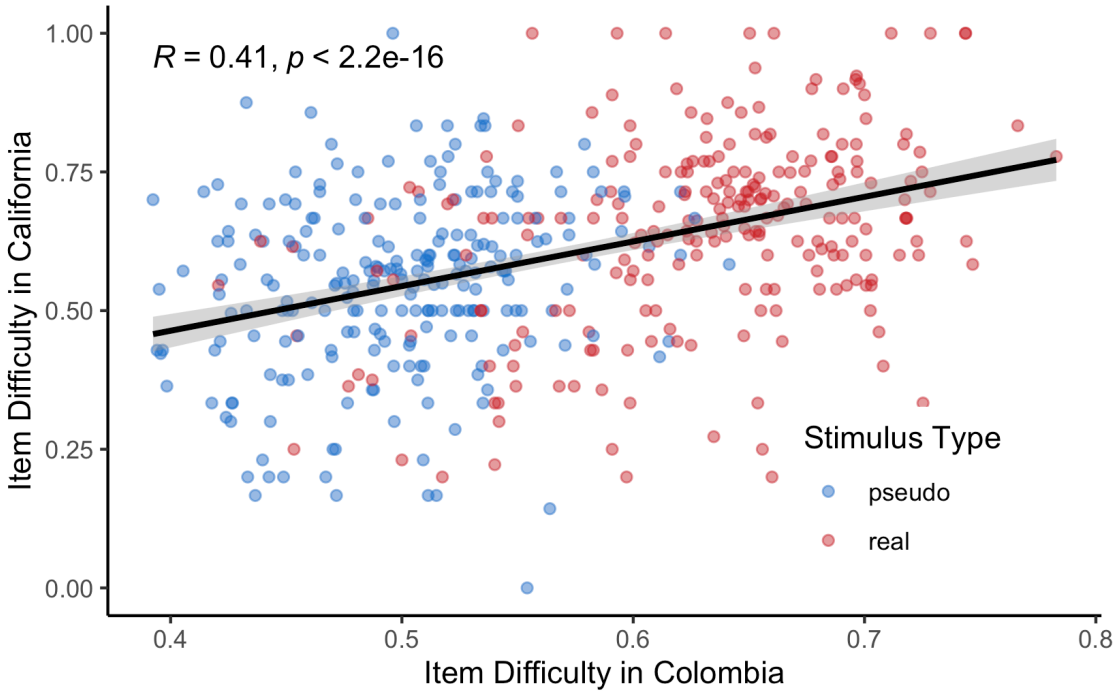


Figure 11.3: Correlation between item difficulty parameters, computed separately by location (California vs. Colombia) drawing only on students in grades K–2.

12 Eficiencia de lectura de frases (ROAR-Frase)

ROAR-Frase is a task designed to measure the speed with which a student can read and comprehend sentences (similar to Chapter 6). The student is presented with one sentence at a time and is tasked with labeling it as true or false as quickly as they can. The sentences are designed in a way that requires mininal background knowledge, use simple syntactic structure avoiding complex wording and structure, and have unambiguous answers. Importantly for Spanish speakers, careful attention was paid to ensure that sentence topics were diverse and not specific to one group of students' experiences. This is especially important for creating a measure that is universal for Spanish speakers across South and Central Latin America and bilingual students in the United States whose experiences might be very different. For example, sentences with topics about things like technology or big cities were avoided as there are several parts of Latin America where these sentences would effectively be testing students on their knowledge of ideas they were not used to interacting with (Rosario Basterra, Trumbull, and Solano-Flores 2011; Hambleton and Kanjee 1995).

As stated previously, the ability to efficiently sentences for understanding is paramount to reading development and is an area of reading that struggling readers, like those with dyslexia, often have a hard time with (Catts et al. 2024; Lyon, Shaywitz, and Shaywitz 2003). The structure of the sentences in ROAR-Frase are designed, as described above, to isolate a student's reading efficiency by minimizing comprehension demands while still assessing for understanding in each sentence.

12.1 Other measures of silent reading efficiency in Spanish

The Woodcock-Johnson Manufacturers, who created the Woodcock-Johnson Tests of Achievement Sentence Reading Fluency sub-test (F. A. Schrank et al. 2014) that is used as a metric of comparison to English ROAR-Sentence (see Section 23.1.1), also adapted a version for use in Spanish. The Spanish measure, Woodcock-Muñoz Batería de aprochevamiento fluidez en lecutra de frases (Fredrick A. Schrank et al. 2005), follows the same setup as the Sentence Reading Fluency measure in English. The student is given a booklet and a pencil and are instructed to read sentences silently and endorse as many as they can as true or false within three minutes. A proctor administers each assessment to a student individually, one on one, providing instructions for sample items and then guiding them through practice responses. The final score here is the total number endorsed correctly minus the total number endorsed incorrectly. In this assessment, skips do not count negatively.

12.2 Structure of the task and design of the items

As with ROAR-Sentence, ROAR-Frase uses a similar task design as the Woodcock-Muñoz fluidez en lecutra de frases. ROAR-Frase, however, is delivered online rather than through through a paper booklet, features gamification, items are designed specifically to tap into silent sentence reading efficiency, and items are delivered in a way that students are forced to endorse each item they see rather than being able to skip items. Importantly, items are designed and validated to ensure cultural relevance across different Spanish language variations spoken in the US and abroad.

As is discussed in the ROAR-Sentence section, the comprehension component that is often built into sentence reading tasks make for interpreting low scores difficult as disaggregating comprehension from efficiency for a struggling reader becomes very difficult. In line with this, there is yet another facet of comprehension to take into account for Spanish speakers as their experience with the language itself may vary so drastically from one student to another (Rosario Basterra, Trumbull, and Solano-Flores 2011; Hambleton and Kanjee 1995). By designing items that are unambiguously true or false, removing content knowledge as well specific experience based knowledge and maintaining simple sentence structures, we aim to target reading efficiency more directly.

12.3 Scoring

ROAR-Frase, like ROAR-Sentence, is a two alternative forced choice (2AFC) task and is scored as the total number of correct responses minus the total number of incorrect responses in the alloted time window. This method of scoring controls for students who were engaging in guessing behavior by accounting for the number of incorrect responses.

12.4 ROAR-Frase Norms for Monolingual Spanish Speakers

As described in Section 10.3, an initial norming study was completed in a primarily monolingual spanish speaking sample in Colombia. Partiipants completed two independent 90 second blocks of ROAR-Frase. ROAR-Frase was delivered to all students above 2nd grade in Colombia. We then ran another group of bilingual students in California in grades 1 and 2. Figure 12.1 displays the norms by grades and Figure 12.2 shows the norms by age. Scores for a represenative sample of multilingual learners in California are shown relative to the monolingual norms.

References

Catts, Hugh W, Nicole Patton Terry, Christopher J Lonigan, Donald L Compton, Richard K Wagner, Laura M Steacy, Kelly Farquharson, and Yaacov Petscher. 2024. "Revisiting the Definition of Dyslexia." *Ann. Dyslexia*, January.

Hambleton, Ronald K, and Anil Kanjee. 1995. "Increasing the Validity of Cross-Cultural Assessments: Use of Improved Methods for Test Adaptations." *European Journal of Psychological Assessment* 11 (3): 147–57.

Lyon, G Reid, Sally E Shaywitz, and Bennett A Shaywitz. 2003. "A Definition of Dyslexia." *Ann. Dyslexia* 53 (1):

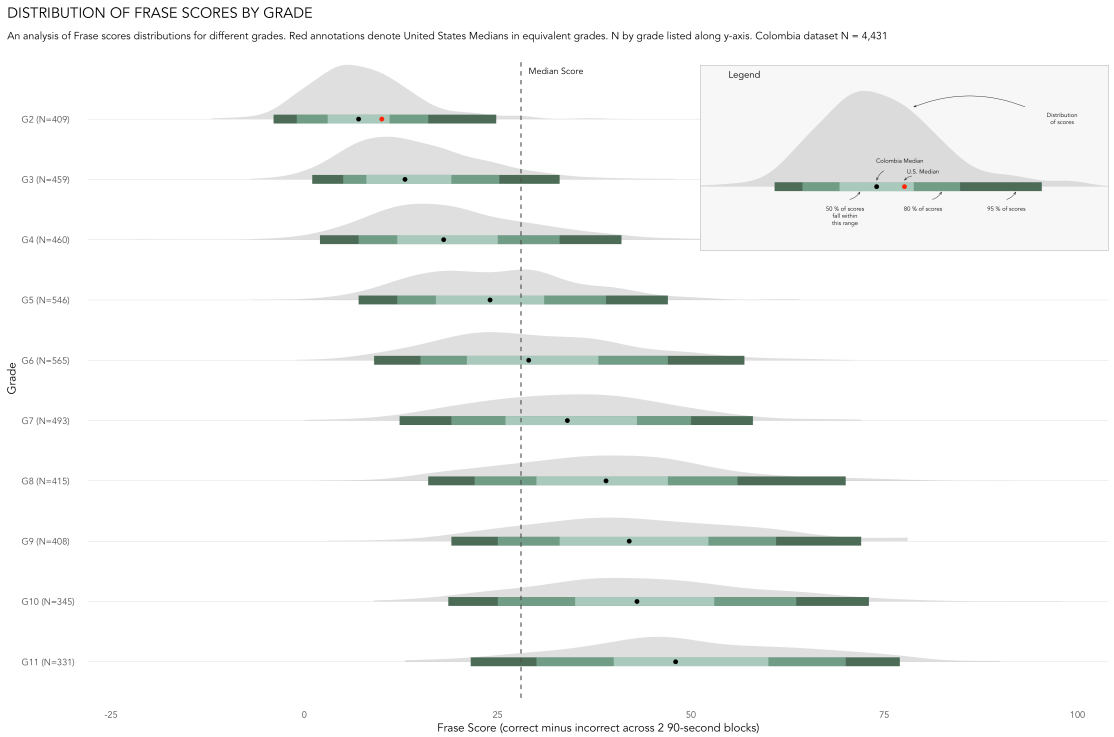


Figure 12.1: ROAR-Frase Colombia score distrubtion by grade. Frase score is the combination of correct minus incorrect across 2 90-second blocks. United States median is included as a red dot on the plot.

1–14.

Rosario Basterra, Marı□a del, Elise Trumbull, and Guillermo Solano-Flores. 2011. *Cultural Validity in Assessment*. Routledge New York, NY.

Schrank, F A, K S McGrew, N Mather, B J Wendling, and E M LaForte. 2014. "Woodcock-Johnson IV Tests of Achievement." Riverside Publishing Company.

Schrank, Fredrick A, Kevin S McGrew, Mary L Ruef, and Criselda G Alvarado. 2005. "Bateri□a III Woodcock-Muñoz." *Rolling Meadows: Riverside Publishing*.

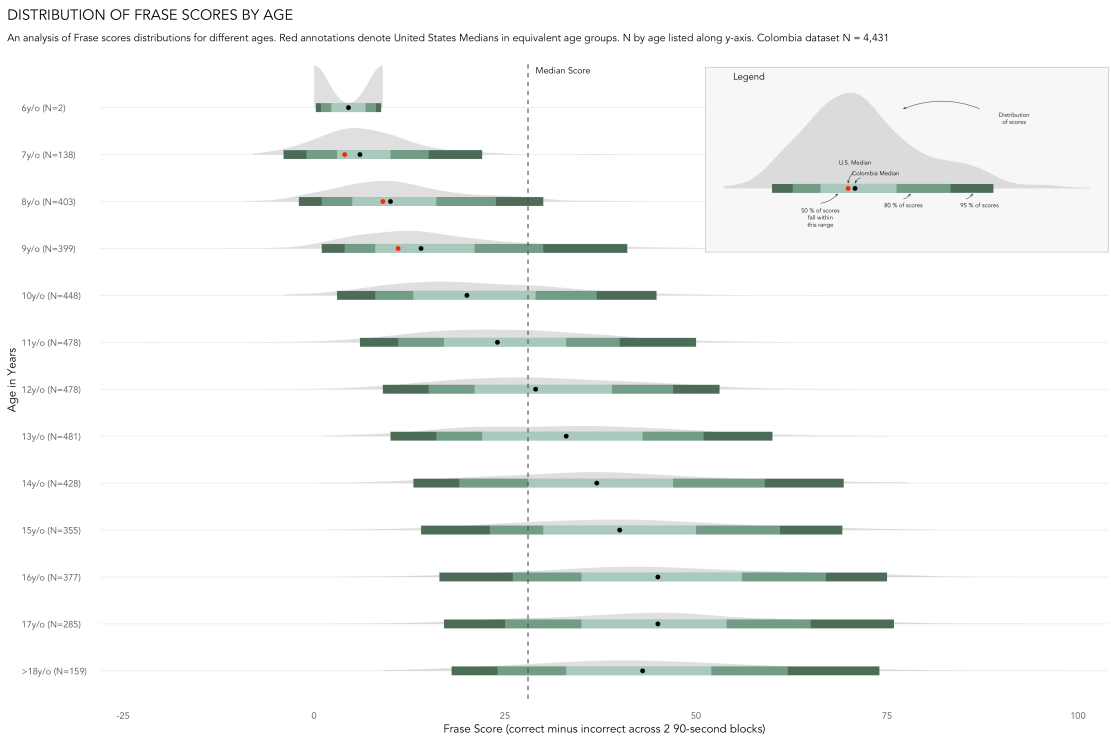


Figure 12.2: ROAR-Frase Colombia score distrubtion by age groups Frase score is the combination of correct minus incorrect across 2 90-second blocks. United States median is included as a red dot on the plot.

13 Conciencia fonológica (ROAR-Fonema)

The skill of phonological awareness (PA), as described in the ROAR-Phoneme section (Chapter 7), refers to the ability to manipulate the sound structure of spoken language at various levels (word, syllable, onset or rime, or phoneme). Spanish is a more transparent orthography than English, meaning that the correspondene between phonemes (sounds) and graphemes (letters) generally follows a one-to-one mapping, where each letter consistently represents a single sound. As a result, researchers hypothesized that phonological awareness might play a different role in learning to read Spanish versus English. However, research indicates that children with reading disabilities in Spanish exhibit phonemic processing difficulties similar to those observed in children learning to read in less transparent orthographies, such as English (Anthony and Francis 2005; Anthony and Lonigan 2004; Anthony et al. 2011, 2009, 2007; Mı□guez-Álvarez, Cuevas-Alonso, and Saavedra 2022; Jiménez González and Rosario Ortiz Gonzalez 1994). Moreover, researchers believe that phonological awareness skills developed in Spanish offer a window through which to aid English reading development in a cross-linguistic transfer for bilingual students where students with strong phonological skills in Spanish were better equipped to develop those same skills in English (Kremin et al. 2019; Manis, Lindsey, and Bailey 2004; Lindsey, Manis, and Bailey 2003). This highlights the critical role of early phonological awareness skills in a student's native language for overall reading development. Therefore, assessing phonological awareness in Spanish is crucial as an early indicator of reading development, helping guide instruction for bilingual learners who are also learning to read in English. ROAR-Fonema was designed to be an efficient and automated measure of phonological awareness. ROAR-Fonema, like ROAR-Phoneme does not rely upon verbal responses.

13.1 Structure of the task

ROAR-Fonema has 2 sub-tests that measure different dimensions of phonological awareness:

- First-Sound Matching (FSM)
- Last-Sound Matching (LSM)

ROAR-Fonema, like ROAR-Phoneme, employs a one-interval, three-alternative forced choice task. As with the entire ROAR suite, instructions are narrated by a character provided with light gamification. The student first interacts with practice items before each subtask where they are given feedback if they respond incorrectly. Practice items must be completed correctly for the student to begin the block. The student hears positive feedback from the characters along the way and is given breaks throughout each block where they can pause if

they would like to. Responses are not timed and a student can take as long as they would like to respond to any given item, but cannot have the item repeated as this is consistent with other tests of phonological awareness. As with ROAR-Phoneme, ROAR-Fonema follows the same stimuli presentation - the student hears a speaker give the prompt, sees the instruction image at the top, is read each of the three answer options at the bottom, and then selects their answer once the speaker finishes labeling each of the photos.

Adaptation of this task to Spanish took into account several important factors. Careful attention was given to ensuring that both items and images were representative of real students' experiences. Regionally specific words or images were avoided for confusion, for example there are several ways to say the word "car" in Spanish and some are more commonly heard in specific areas while others are more common broadly. We chose to use the words that were most commonly used across diverse sets of Spanish speakers. Additionally, great emphasis was placed on assuring that a wide range of phonemes with varying difficulty levels as well as frequencies were chosen. Similar to ROAR-Letra, attention was also paid to phonemes that would be confusing for bilingual versus monolingual Spanish speakers.

References

Anthony, Jason L, and David J Francis. 2005. "Development of Phonological Awareness." *Curr. Dir. Psychol. Sci.* 14 (5): 255–59.

Anthony, Jason L, and Christopher J Lonigan. 2004. "The Nature of Phonological Awareness: Converging Evidence from Four Studies of Preschool and Early Grade School Children." *Journal of Educational Psychology* 96 (1): 43.

Anthony, Jason L, Emily J Solari, Jeffrey M Williams, Kimberly D Schoger, Zhou Zhang, Lee Branum-Martin, and David J Francis. 2009. "Development of Bilingual Phonological Awareness in Spanish-Speaking English Language Learners: The Roles of Vocabulary, Letter Knowledge, and Prior Phonological Awareness." *Scientific Studies of Reading* 13 (6): 535–64.

Anthony, Jason L, Jeffrey M Williams, Lillian K Durán, Sandra Laing Gillam, Lan Liang, Rachel Aghara, Paul R Swank, Mike A Assel, and Susan H Landry. 2011. "Spanish Phonological Awareness: Dimensionality and Sequence of Development During the Preschool and Kindergarten Years." *Journal of Educational Psychology* 103 (4): 857.

Anthony, Jason L, Jeffrey M Williams, Renee McDonald, and David J Francis. 2007. "Phonological Processing and Emergent Literacy in Younger and Older Preschool Children." *Annals of Dyslexia* 57: 113–37.

Jiménez González, JE, and Maria del Rosario Ortiz Gonzalez. 1994. "Phonological Awareness in Learning Literacy." *Intellectica* 18 (1): 155–81.

Kremin, Lena V, Maria M Arredondo, Lucy Shih-Ju Hsu, Teresa Satterfield, and Ioulia Kovelman. 2019. "The Effects of Spanish Heritage Language Literacy on English Reading for Spanish–English Bilingual Children in the US." *International Journal of Bilingual Education and Bilingualism* 22 (2): 192–206.

Lindsey, Kim A, Franklin R Manis, and Caroline E Bailey. 2003. "Prediction of First-Grade Reading in Spanish-Speaking English-Language Learners." *Journal of Educational Psychology* 95 (3): 482.

Manis, Franklin R, Kim A Lindsey, and Caroline E Bailey. 2004. "Development of Reading in Grades k–2 in Spanish–Speaking English–Language Learners." *Learning Disabilities Research & Practice* 19 (4): 214–24.

Mı□guez-Álvarez, Carla, Miguel Cuevas-Alonso, and Ángeles Saavedra. 2022. "Relationships Between Phonological Awareness and Reading in Spanish: A Meta-Analysis." *Language Learning* 72 (1): 113–57.

14 Conocimiento de Sonidos de Letras (ROAR-Letra)

ROAR-Letra is the Spanish version of ROAR-Letter. This assessment measures knowledge of both upper-case and lower-case letter names and sounds in Spanish. ROAR-Letra has been adapted for Spanish by incorporating relevant letter names and sounds as well as adjusting distractors to be better suited for the Spanish language. ROAR-Letra can be used to assess comprehensive letter name and sound knowledge, or it can be shortened to evaluate a specific number of randomly selected items in each subskill. Additionally, ROAR-Letra can be used as a diagnostic tool to guide instruction as it returns specific information about the letter names and letter-sound correspondences that the individual student does and does not know.

14.1 Adaptation of the Task to Spanish

ROAR-Letra underwent rigous testing through an iterative research and development process to be universally suitable for Spanish speaking students from diverse linguistic backgrounds. Careful attention was paid to ensure the sounds and letters were understandable across different Spanish dialects by accounting for variations in pronunciation. For example, sounds like "ll" which may be pronounced differently in various dialects were chosen very carefully to ensure broad usability.

Additionally, important considerations for bilingual students were also kept in mind. For example, "x", "h", and "j" were not included as distractors for one another as bilingual students, who are accustomed to both English and Spanish, may have a harder time differentiating these sounds. This decision was influenced by research on bilingual children's spelling strategies and common mistakes that bilingual learners make when spelling. Helman (2004) highlights how Spanish-speaking students leverage their understanding of the Spanish sound system when learning to spell in English, which can lead to confusion with similar-sounding letters like "h" and "j" or "v" and "b". Zutell and Allen (1988) discuss the spelling strategies used by Spanish-speaking bilingual children and the challenges they face in distinguishing between English and Spanish phonemes that sound similar across languages. By avoiding such distractors, we aim to create an assessment that is fair to English language learners.

14.2 Structure of the task

ROAR-Letra, like ROAR-Letter, is a four alternative fourced choice (4AFC) task divided into 3 blocks: - Upper-case letter names (27 items) - Lower-case letter names (27 items) - Letter-sound correspondences (62 items) Like all ROAR measures, ROAR-Letra is lightly gamified.

The task begins with instructions and practice trials with feedback until the student understands the game. Then, in each block, the student is presented with the name or sound of a letter and asked to select the correct letter from the four choices. The student can replay letter names or sound if they need to. Figure 14.1 depicts ROAR-Letra.

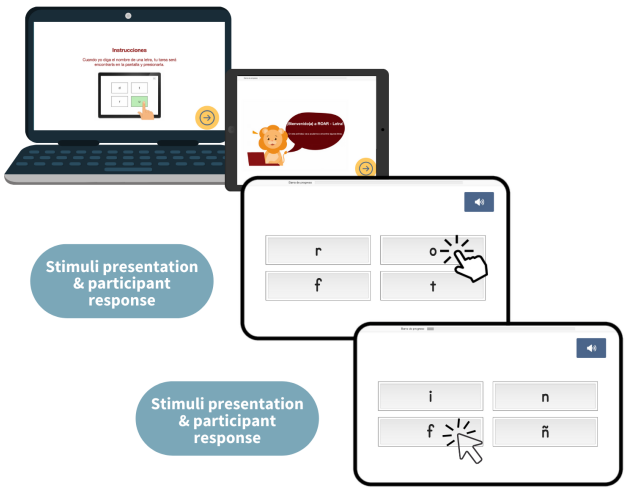


Figure 14.1: ROAR-Letra task. In the ROAR-Letra task, participants are first directed to choose the spoken letter name and then the letter sound in separate blocks. The task is divided into several blocks with encouragement throughout. Students have the opportunity to replay any letter name or sound audio they miss. There is no time limit for responses.

Part III

Reliability

References

Helman, Lori A. 2004. "Building on the Sound System of Spanish: Insights from the Alphabetic Spellings of English-Language Learners." *The Reading Teacher* 57 (5): 452–60.

Zutell, Jerry, and Virginia Allen. 1988. "The English Spelling Strategies of Spanish-Speaking Bilingual Children." *TESol Quarterly* 22 (2): 333–40.

15 Reliability of ROAR-Word

15.1 Background: Published studies

The first published version of ROAR-Word achieved exceptional alternate form reliability (r=0.95) using fixed forms that were equated based on item response theory (Yeatman et al. 2021). To improve efficiency of ROAR-Word, Ma et al. (2023) built the first, open-source, computer adaptive testing (CAT) algorithm in Javascript[1], and then ran a series of experiments to study how reliability and efficiency of ROAR-Word could be improved with CAT. Figure 15.1 reproduces a figure from Ma et al. (2023) showing an experiment comparing ROAR-CAT to a standard, non-adaptive testing approach. In this experiment, participant were randomly assigned to complete ROAR-Word with the trial order controlled by either a) jsCAT[2] (solid line) versus b) random item sampling (dotted line). ROAR-CAT achieved the same reliability in roughly 40% fewer trials.

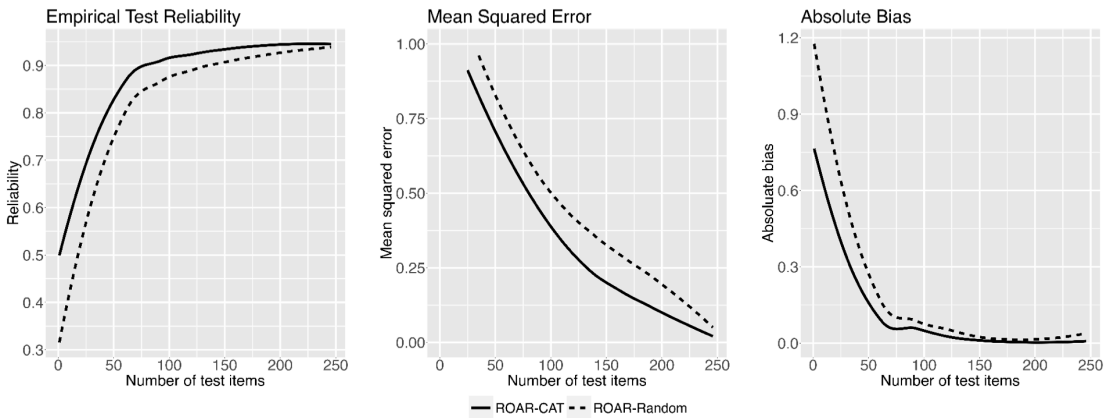


Figure 15.1: ROAR-Word is 40% more efficient when using computer adaptive testing.

This innovation has now been incorporated into all the ROAR measures to create quick and efficient, adaptive assessments that span broad age ranges.

[1] https://github.com/yeatmanlab/jscat
[2] https://github.com/yeatmanlab/jscat

15.2 Criteria for identifying disengaged participants and flagging unreliable scores

ROAR-Word is designed to be totally automated: words are read silently, responses are non-verbal, instructions and practice trials are narrated by characters, and scoring is done in real time after each response. This makes it possible to efficiently assess a whole school district simultaneously. However, a concern about automated assessments is that without a teacher to individually administer items, monitor, and score responses, some students might disengage and provide data that is not representative of their true ability. One benefit of a lexical decision task is that there is an extensive literature on the expected response time distribution (Balota, Yap, and Cortese 2006; Keuleers et al. 2012; Balota et al. 2007). Based on the amount of time it takes signals from the eye to reach the brain, for the visual features to be processed, the word to be recognized, and a motor response to be initiated, extremely fast response times are most likely due to rapid guessing behavior indicative of disengagement from the assessment (Ratcliff, McKoon, and Gomez 2004; Balota, Yap, and Cortese 2006). Our previous publications have validated fast response time as an indicator of participant disengagement (Ma et al. 2023; Yeatman et al. 2021). This effect can be seen in Figure 15.2 which shows a plot of median response time (RT) versus proportion correct for each participant. **None** of the participants with a median response time less than 450ms (horizontal black line in Figure 15.2) are accurate on ROAR-Word. Since ROAR-Word is run as a computer adaptive test (CAT), All participants should be around 75% correct: item difficulty changes adapatively based on participant responses. Participants that respond very quickly and inaccurately are disengaged and not providing data that is representative of their true ability.

> **! CRITERIA FOR FLAGGING UNRELIABLE SCORES**
>
> Participants with low accuracy (<65% correct) and a median response time <450ms are flagged in ROAR-Score reports and their data is excluded from analyses. Teachers can choose whether to re-administer ROAR or interpret data cautiously in relation to other data sources and contextual factors.

15.3 Reliability of computer adativer ROAR-Word

ROAR-Word runs as computer adaptive test based on a Rasch model. The current, default version of ROAR-Word takes about 4 minutes (84 items). More items can be administered for a more precise measure or fewer items can be administered as a quick screener. Table 15.1 reports marginal reliability computed based on data from 10294 students under the IRT model for the standard, 84 item version of ROAR-Word. Reliability ($\rho_{xx'}$) is computed based on the estimated variance of $\hat{\theta}$ relative to the estimated standard error ($\widehat{SE}(\hat{\theta})^2$) using Equation 20.1:

$$\hat{\rho}_{xx'} = \frac{\widehat{VAR}(\hat{\theta})}{\widehat{VAR}(\hat{\theta}) + \widehat{SE}(\hat{\theta})^2}, \tag{15.1}$$

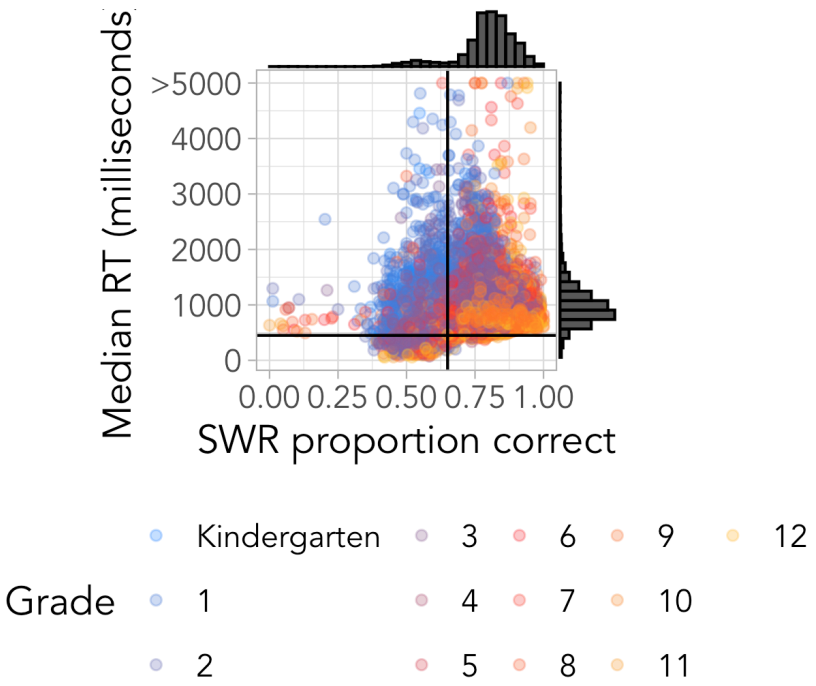


Figure 15.2: Criteria for identifying disengaged participants and flagging unreliable scores on ROAR-Word. Participants displaying extremely rapid responses performed near chance indicative of disengagement and/or rapid guessing behavior. Black lines indicate the cut off for flagging disengaged participants with unreliable scores.

Table 15.1: Reliability of ROAR-Word by Grade

Grade	Empirical Reliability	N
All	0.9415742	10294
K	0.8708076	131
1	0.9174180	1050
2	0.9295753	1123
3	0.9418479	572
4	0.9361429	320
5	0.9397313	315
6	0.9203889	1000
7	0.9114732	846
8	0.9195974	716
9	0.9093690	1347
10	0.9071286	1243
11	0.9082764	932
12	0.9081830	699

To ensure that ROAR-Word is a fair and equitable assessment across different demographic groups we also report reliability separately by gender (Table 15.2), eligibility for free and re-

duced price lunch (Table 15.3), English learner status as designated by the school district (Table 15.4), primary language (Table 15.5), special education (Table 15.6), ethnicity (Table 15.7), and race (Table 15.8)

Table 15.2: Reliability of ROAR-Word by Gender (F=female, M=male)

Gender	Empirical Reliability	N
All	0.9415742	3733
F	0.9442770	1800
M	0.9433320	1933

Table 15.3: Reliability of ROAR-Word by FRL (F=Free, P=Paid, R=Reduced)

Free/Reduced Lunch Status	Empirical Reliability	N
All	0.9415742	1949
F	0.9275092	409
P	0.9434327	1390
R	0.9271505	150

Table 15.4: Reliability of ROAR-Word by EL Status (EL=English Learner, EO=English Only, IFEP=Initial Fluent English Proficient, RFEP=Reclassified Fluent English Proficient)

English Learner Status	Empirical Reliability	N
All	0.9415742	2368
EL	0.9474195	897
EO	0.9432996	1180
IFEP	0.9356226	213
RFEP	0.9344000	76
TBD	0.5528642	2

Table 15.5: Reliability of ROAR-Word by Primary Language

Primary Language	Empirical Reliability	N
All	0.9415742	1916
English	0.9447252	1396
Other	0.9154780	188
Spanish	0.9175260	332

Table 15.6: Reliability of ROAR-Word by Special Education Status

Special Education Status	Empirical Reliability	N
All	0.9415742	2046

0	0.9456441	1874
1	0.9470536	172

Table 15.7: Reliability of ROAR-Word by Hispanic Ethnicity

Hispanic Ethnicity	Empirical Reliability	N
All	0.9415742	4246
0	0.9422047	2580
1	0.9436946	1666

Table 15.8: Reliability of ROAR-Word by Race

Race	Empirical Reliability	N
All	0.9415742	3187
American Indian or Alaska Native	0.9338252	3
Asian	0.9400807	439
Black or African American	0.9310394	37
Filipino	0.9404038	9
Hawaiian or Other Pacific Islander	0.9042327	8
Hispanic	0.9436946	1666
Multiracial	0.9392057	249
White	0.9488497	776

References

Balota, David A, Melvin J Yap, and Michael J Cortese. 2006. "Visual Word Recognition." In *Handbook of Psycholinguistics*, 285–375. Elsevier.

Balota, David A, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. "The English Lexicon Project." *Behavior Research Methods* 39: 445–59.

Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. "The British Lexicon Project: Lexical Decision Data for 28,730 Monosyllabic and Disyllabic English Words." *Behav. Res. Methods* 44 (1): 287–304.

Ma, Wanjing A, Adam Richie-Halford, Amy Burkhardt, Klint Kanopka, Clementine Chou, Benjamin Domingue, and Jason D Yeatman. 2023. "ROAR-CAT: Rapid Online Assessment of Reading Ability with Computerized Adaptive Testing."

Ratcliff, Roger, Gail McKoon, and Pablo Gomez. 2004. "A Diffusion Model Account of the Lexical Decision Task." *Psychol. Rev.* 111 (1): 159–82.

Yeatman, Jason D, Kenny An Tang, Patrick M Donnelly, Maya Yablonski, Mahalakshmi Ramamurthy, Iliana I Karipidis, Sendy Caffarra, et al. 2021. "Rapid Online Assessment of Reading Ability." *Sci. Rep.* 11 (1): 6396.

16 Reliability of ROAR-Sentence

ROAR-Sentence is a timed measure and the score is computed as the number of correct trials minus the number of incorrect trials in the alloted period time window. Originally, ROAR-Sentence was 3 minutes long but (Tran et al. 2023) demonstrated the cutting the time in half to 90 seconds had very little impact on reliability and validity of the measure. ROAR-Sentence consists of a collection of equated test forms where sentences are presented in a fixed order. We first report our methodology for equating test forms (Section 16.1) and then report alternate form reliability (Section 16.3).

16.1 Equating ROAR-Sentence test forms

16.2 Criteria for identifying disengaged participants and flagging unreliable scores

ROAR-Sentence is designed to be totally automated: reading is done silently, responses are non-verbal, instructions and practice trials are narrated by characters, and scoring is done automatically in real time. This makes it possible to efficiently assess a whole district simultaneously. A concern about automated assessments is that without a teacher to individually administer items, monitor, and score responses, some students might disengage and provide data that is not representative of their true ability. For a measure like ROAR-Sentence where items are designed and validated to have an unambiguous and clear answer, disengaged participants can be detected based on fast and innacurate responses. Our approach to identifying and flagging disegnaged participants with unreliable scores was published in (Tran et al. 2023). Figure 16.1 shows a plot of median response time (RT) versus proportion correct for each participant. Most participants were very accurate (>90% correct responses). However there was a bimodal distribution indicating a small group of participants who were performing around chance. These participants also had extremely fast response times.

! Criteria for flagging unreliable scores

Participants with a median response time <1,000ms AND low accuracy (<65% correct) are flagged as unreliable scores in ROAR score reports and are excluded from analyses since scores do not accurately represent the participant's ability. Teachers can choose whether to re-administer ROAR or interpret data cautiously in relation to other data sources and contextual factors.

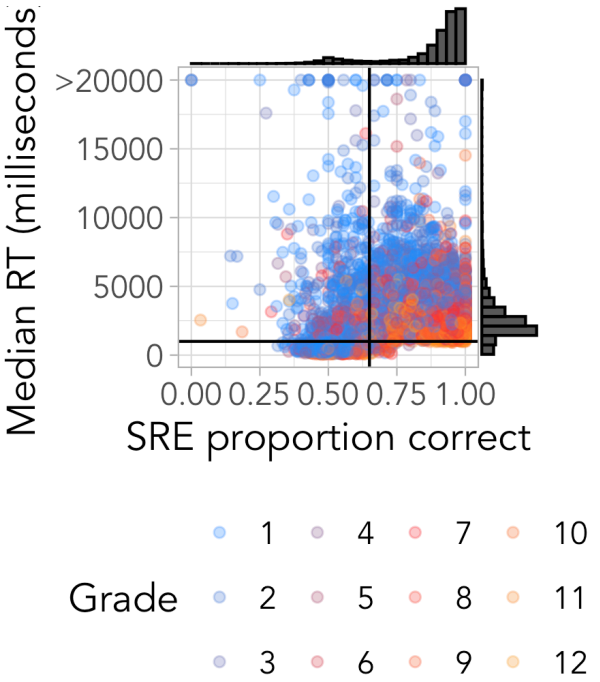


Figure 16.1: Criteria for identifying disengaged participants and flagging unreliable scores on ROAR-Sentence. Participants displaying extremely rapid responses performed near chance on ROAR-Sentence. This criteria is consistent across multiple studies (Tran et al. 2023). Black lines indicate the cut off for flagging disengaged participants with unreliable scores.

16.3 Alternate form reliability

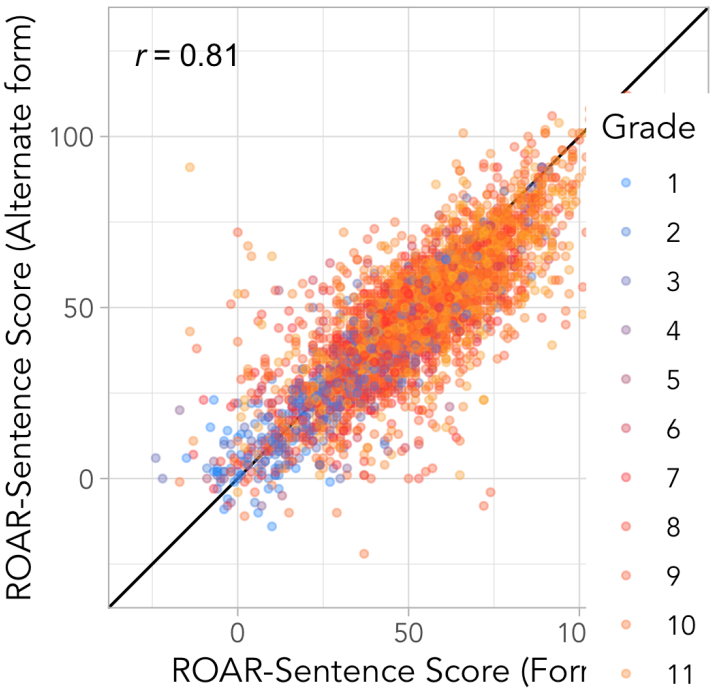
Alternate form reliability is computed as the Pearson correlation between scores on equated test forms that were administered during the same testing session. Figure 16.2a shows a plot of student scores on alternate test forms combining grades and Figure 16.2b shows separate plots for each grade. Table 16.1 reports alternate form reliability for the full sample and separately by grade.

Table 16.1: Alternate form reliability for ROAR-Sentence

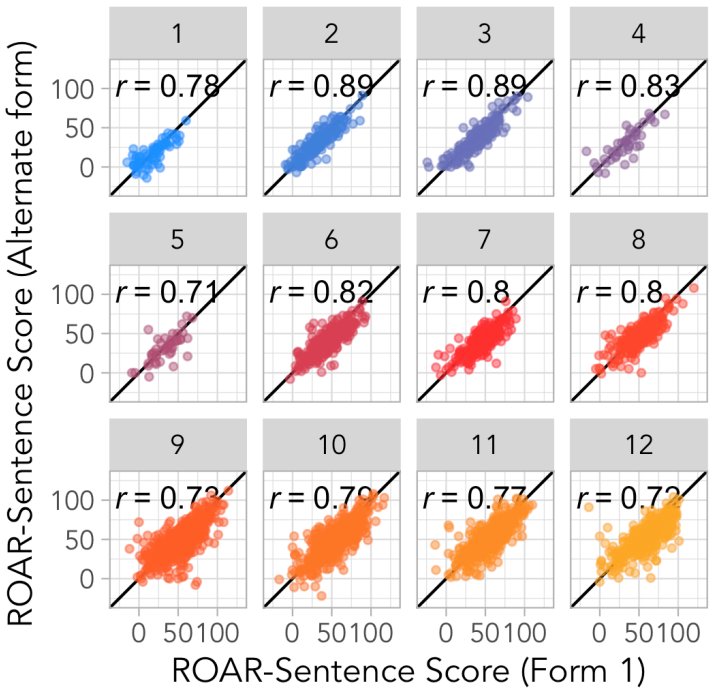
Grade	Alternate Form Reliability	N
All	0.8116780	3633
1	0.7810830	69
2	0.8890698	163
3	0.8943240	185
4	0.8320945	44
5	0.7115194	45
6	0.8184703	271
7	0.7958987	209
8	0.7963777	272
9	0.7307424	745
10	0.7894435	643
11	0.7669829	542
12	0.7175356	445

References

Tran, Jasmine E, Jason D Yeatman, Amy Burkhardt, Wanjing A Ma, Jamie Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. 2023. "Development and Validation of a Rapid Online Sentence Reading Efficiency Assessment."



(a) Alternate form reliability across grades



(b) Alternate form reliability separately by grade

Figure 16.2: Alternate form reliability for ROAR-Sentence

17 Reliability of ROAR-Phoneme

17.1 Background: Published studies

Gijbels et al. (2024) reported high correlations between ROAR-Phoneme and standardized measures of PA (CTOPP-2, r=.80) for children from Pre-K through fourth grade and exceptional reliability for the ROAR-Phoneme composite score ($\alpha = 0.96$). Each individual subtest was also highly reliable:

- First Sound Matching (FSM) $\alpha = 0.89$
- Last Sound Matching (LSM) $\alpha = 0.92$
- Deletion (DEL) $\alpha = 0.85$

References

Gijbels, Liesbeth, Amy Burkhardt, Wanjing Anya Ma, and Jason D Yeatman. 2024. "Rapid Online Assessment of Reading and Phonological Awareness (ROAR-PA)." *Sci. Rep.* 14 (1): 1–16.

18 Reliability of ROAR-Letter

18.1 Design and implimentation of computer-adaptive letter-sound assessment

A Rasch model was fit to the ROAR-Letter calibration sample (see Table 8.1 and Table 8.2). All ROAR-Letter items fit the model well (see Chapter 4 for fit criteria). Based on this IRT model, Figure 18.1 shows the upper and lower bounds on reliability as a function of the number of items that a participant completes. We then ran a CAT simulation as in (Ma et al. 2023) to determine the final item selection criteria that would maximize reliability in the fewest number of trials.

ROAR Letter CAT Parameters and Reliability

25 Trials

- 5 Upper case letter names
- 5 Lower case letter names
- 15 Letter-sound correspondences

Item selection: Fisher information

Marginal reliability = 0.85

References

Ma, Wanjing A, Adam Richie-Halford, Amy Burkhardt, Klint Kanopka, Clementine Chou, Benjamin Domingue, and Jason D Yeatman. 2023. "ROAR-CAT: Rapid Online Assessment of Reading Ability with Computerized Adaptive Testing."

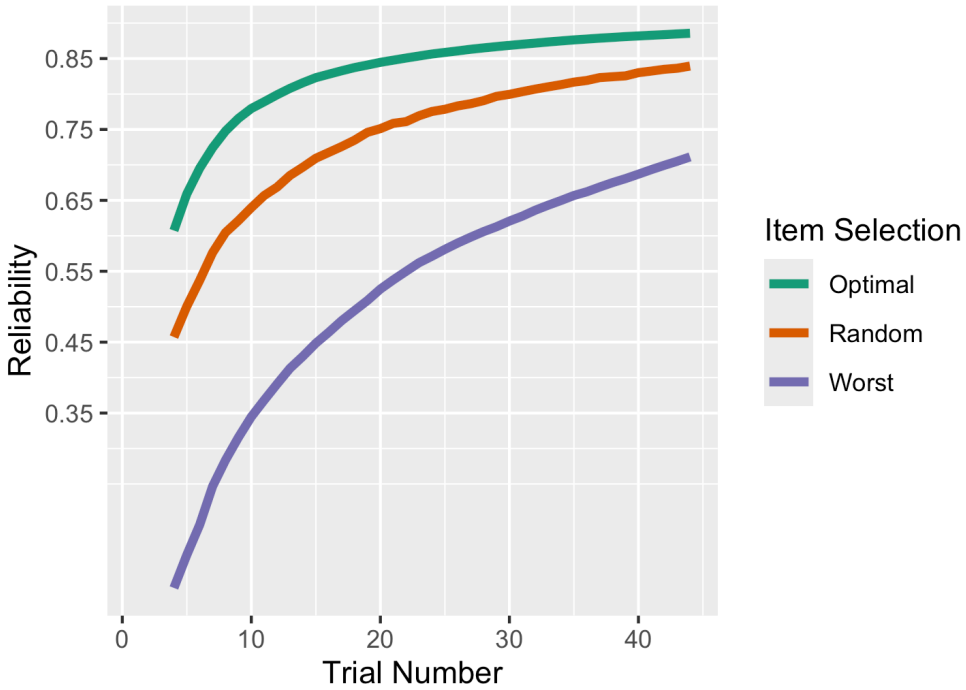


Figure 18.1: Letter-CAT simulation based on item-response data in 4,041 kindergarten and first grade students. Items were sampled in 3 different ways and marginal reliability was calculated as a function of the number of items that each participant completed. The simulation shows that the choice of items has a major impact on the reliability of the measure. For Optimal sampling (green) the N items with difficulty closest to the participant's $ heta$ estimate were used. For Random sampling (orange) a random sample of N items were taken for each participant. For Worst sampling (purple) the N items with difficulty furthest from the participant's $ heta$ estimate were used. This simulation highlights the massive efficiency gain that would be possible from an optimized CAT.

19 Reliability of Dyslexia Prediction and Subtyping

Many dyslexia screening initiatives require the use of specific measures such as Rapid Automatized Naming (RAN; see Section 9.2.1) and measures of visual processing (RVP; see Section 9.2.2). The following sections report the reliability of ROAR measures that were specifically designed for dyslexia prediction and subtyping (see Section 9.2 for more information and a theoretical background on the measures).

19.1 Reliability of ROAR Rapid Automatized Naming (ROAR-RAN)

Reliability and evidence of construct validity was assessed based on correlations among scores across RAN-Letters, RAN-Colors and RAN-Numbers. The assessment was conducted in two stages: first, measuring the reliability of the automated scoring system relative to the manual scoring method, and second, analyzing the correlation between the three RAN measures using the automated scores.

19.1.1 Automated Scoring Reliability

We assessed the reliability of our automated scoring system by comparing it to manual scoring in a sample of 100 participants. In the manual process, the duration of each task was measured by manually timing from the start of the first spoken word and the end of the last spoken word. In the automated process, the duration was determined using start and end timestamps generated by our automatic speech recognition model.

The correlation between manual and automated scoring was calculated for each RAN task, with each task achieving a Pearson correlation coefficient greater than 0.95 (see Figure 19.1). These strong correlations suggest that the automated scoring system reliably produces scores that are nearly identical to manual scoring.

19.1.2 Correlation Among ROAR-RAN Measures

Next, we examined the correlation between the three RAN measures—RAN-Letters, RAN-Colors, and RAN-Numbers—using the automated scores. This analysis aimed to confirm the construct validity of the ROAR-RAN tasks by evaluating the relationships among these measures. All three measures are designed to tap into the same latent construct, though color naming is not as automatized as letter naming and, thus, we expect a lower correlation for RAN-Colors.

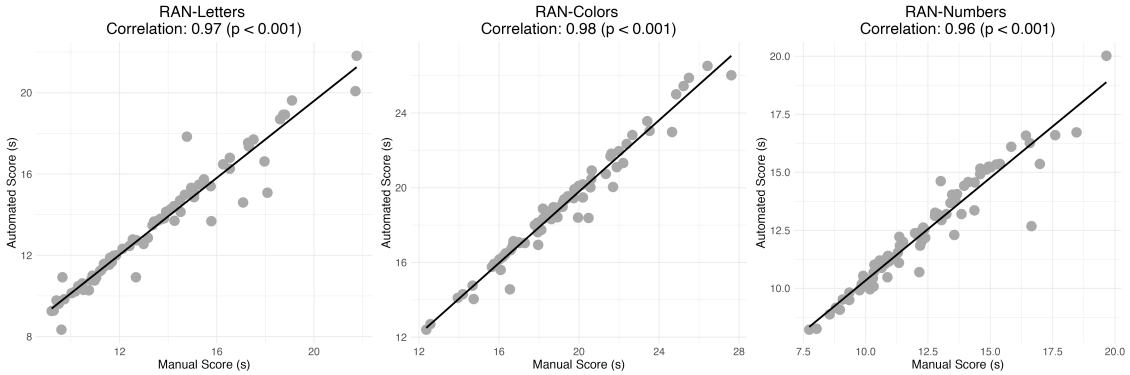


Figure 19.1: ROAR-RAN automated scoring algorithm is precise and reliable

We calculated Pearson correlations between the automated scores for all participants across the following pairs of tasks: RAN-Letters and RAN-Numbers, RAN-Letters and RAN-Colors, and RAN-Colors and RAN-Numbers. Figure 19.2 shows these correlations providing evidence that each measure is reliably tapping in to a similar latent construct.

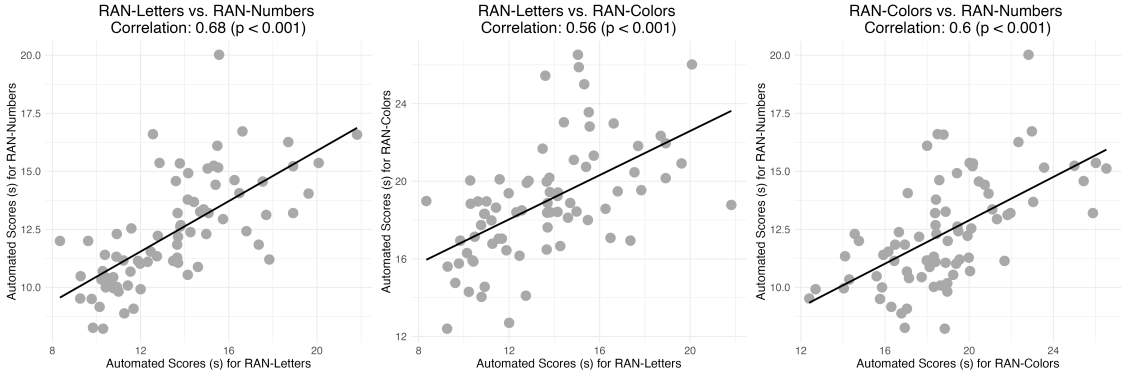


Figure 19.2: Evidence for reliability and construct validity of ROAR Rapid Automized Naming

19.2 Reliability of ROAR Rapid Visual Processing (ROAR-RVP)

19.2.1 Background: Published studies

The rapid visual processing paradigm that was administered to older children between 7-17 years (Ramamurthy, White, and Yeatman 2023) was translated for a younger population and a computer adaptive testing algorithm[1] was implemented see 4 in order to improve reliability and ensure the task spans a broad age range. We translated the task for a younger population by iteratively changing the task and the design while administering it to small groups of kindergarten and first graders within the Multitudes study[2] sample population. This work was done through collaboration between the ROAR team at Stanford University[3] and the Multi-

[1] https://github.com/yeatmanlab/jscat
[2] https://dyslexia.ucsf.edu/Multitudes
[3] https://roar.stanford.edu/

tudes Team at UCSF[4]. The final versions are tailored to Kindergarten and first grade children but spans through adulthood. Details of the design and validation process[5] are published in Ramamurthy et al. (2024).

19.2.2 Data informed design changes to achieve high reliability in young children

For the RVP measure item difficulty depends on a variety of task parameters. There are two important task parameters that have the potential to influence performance: 1) encoding time and 2) string length.

19.2.2.1 Study 1 (N = 56)

We first administered the task with similar parameters as were used in our previous study of older children and adults (Ramamurthy, White, and Yeatman 2023) (encoding time of 120ms and string length of 6 letters). We observed that K/1 children's task performance was significantly lower with an encoding time of 120ms (16.813% +/- 7.654) with very low reliability (Spearman Brown corrected split half reliability: 0.075) compared to the older cross-sectional population reported in our previous work (37.148% +/- 1.191; reliability: 0.8). However, performance increased with an encoding time of 240 (21.125%+/- 8.371) and 480 ms (24.107% +/- 9.851; d': 0.269). Notably, even at 480ms many participants still performed at chance and reliability was still low (Spearman Brown corrected split half reliability was 0.306).

19.2.2.2 Study 2 (N = 86)

We tested how performance changes in trials with four elements (2 on either side of fixation) and six elements (3 on either side of fixation) with encoding times of 240 ms and 480 ms. We observed that there was an overall improvement in task performance in Study 2 [Mean accuracy: 30.025 +/- 1.344] compared to the overall task performance from Study 1 [Mean accuracy: 20.685 +/-0.777; Mean d': 0.159+/-0.032].

19.2.2.3 Study 3 (N= 175)

In the next iteration, we added a 2-element string in addition to 4- and 6- element strings. We further reduced redundancy by removing an encoding time of 480 ms that did not increase accuracy. An encoding period 240 ms ensures that encoding occurs without making a saccadic eye movement (Li, Hanning, and Carrasco 2021). Overall task performance increased significantly compared to Studies 1 and 2. Performance in Study 3 (40.429% +/- 1.0368) is comparable to the overall performance reported in a previous study with cross-sectional data (n=185) (Ramamurthy, White, and Yeatman 2023), where overall task performance for 6 to

[4] https://dyslexia.ucsf.edu/Multitudes
[5] https://osf.io/preprints/psyarxiv/x6ysc

17 yr olds in the MEP task with an encoding time of 120ms and a string length of 6 elements was 37.148% +/- 1.191. Overall task reliability was comparable between Study 3 (r = 0.8) and Study 2 (r = 0.802) see Figure Figure 19.3 below.

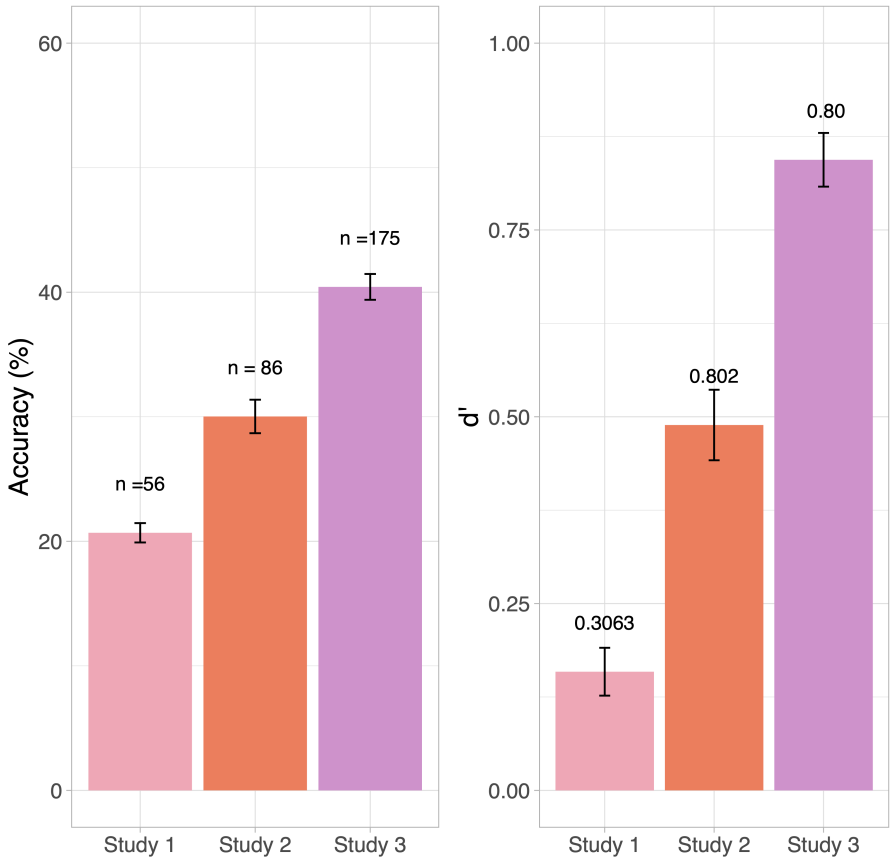


Figure 19.3: Studies to optimize reliability of ROAR Rapid Visual Processing

19.2.2.4 IRT to model item difficulty levels and to optimize task parameters

Data from Study 3 (N = 175) was used to calibrate an Item Response Theory (IRT) model. Trials with different string lengths were blocked (twelve 2-letter trials, twenty four 4-letter trials and thirty six 6-letter trials) respectively. The goal of IRT is to place item difficulty (blocks of different string lengths) on an interval scale. The Rasch model (1 parameter logistic with a guess rate fixed at 0.167) was fit to the response data for the 3 item types (constraining difficulty for repeated trials with the same string lengths) for all 175 participants using the MIRT package in R (Philip Chalmers 2012). Figure 19.4 shows item difficulty for each block (a) and item response functions for all three blocks of different string lengths (b).

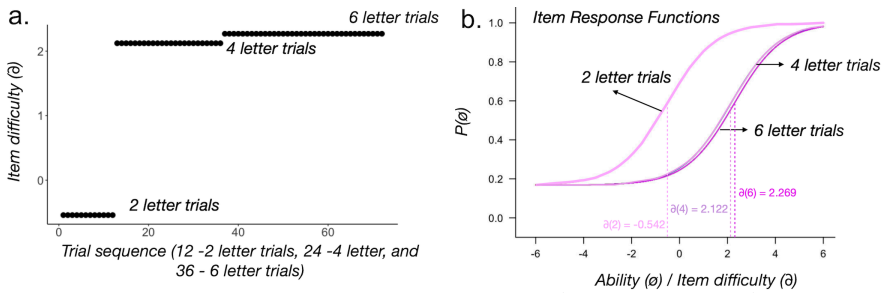


Figure 19.4: Calibration of item response theory (IRT) model for ROAR Rapid Visual Processing

19.2.2.5 Final Optimized version

As a first step towards reducing redundancy, we can shorten the task by eliminating the thirty six 6-element trials completely and used 2- and 4- element trials with an encoding time to 240ms. Further, for efficient task administration we built a simple transition rule and a termination rule. At the end of the 2-element block if children get 4 or more trials correct (>=4/24) then they transition to the next item difficulty and complete eight 4-letter trials. If children perform at or below chance then the task terminates after twenty four 2-letter trials.

19.2.3 Construct validity: performance on RVP-Letters and RVP-Symbols is highly correlated

We used the optimized version of the RVP task described above and created two versions 1) letters and 2) symbols (pseudo-letters). These measures were administered to 1457 children in K/1 across the state of California by the UCSF Dyslexia Center[6] as part of an initiative to develop a universal dyslexia screener, Multitudes[7]. As an initial proof-of-concept study we compared ability in the RVPL task and RVPS task and found a high correlation (r = 0.73; disattenuated r = 0.9125 given task reliability of 0.80, Figure 19.5). This provides evidence that both RVPL and RVPS reliably tap into the same latent construct.

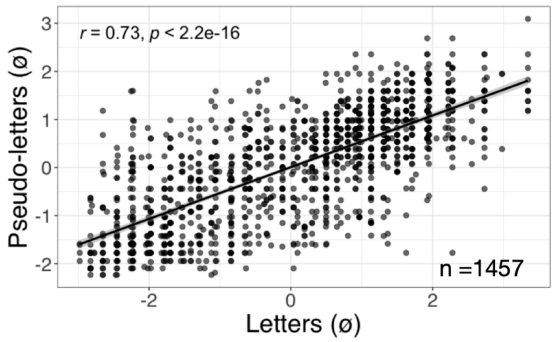


Figure 19.5: Evidence for reliability and construct validity of ROAR Rapid Visual Processing

[6] https://dyslexia.ucsf.edu/
[7] https://multitudesinfo.ucsf.edu/

Part IV

Reliability of ROAR-Español

References

Li, Hsin-Hung, Nina M Hanning, and Marisa Carrasco. 2021. "To Look or Not to Look: Dissociating Presaccadic and Covert Spatial Attention." *Trends Neurosci.* 44 (8): 669–86.

Philip Chalmers, R. 2012. "Mirt: A Multidimensional Item Response Theory Package for the R Environment." *J. Stat. Softw.* 48 (May): 1–29.

Ramamurthy, Mahalakshmi, Klint Kanopka, Adam Richie-Halford, Benjamin Domingue, Francesca Pei, Phaedra Bell, Lucy Yan, Andrea Hartsough, Maria L Gorno-Tempini, and Jason D Yeatman. 2024. "Design and Validation of a Rapid Visual Processing Measure for Screening Reading Difficulties in Early Childhood," February.

Ramamurthy, Mahalakshmi, Alex White, and Jason D Yeatman. 2023. "Children with Dyslexia Show No Deficit in Exogenous Spatial Attention but Show Differences in Visual Encoding."

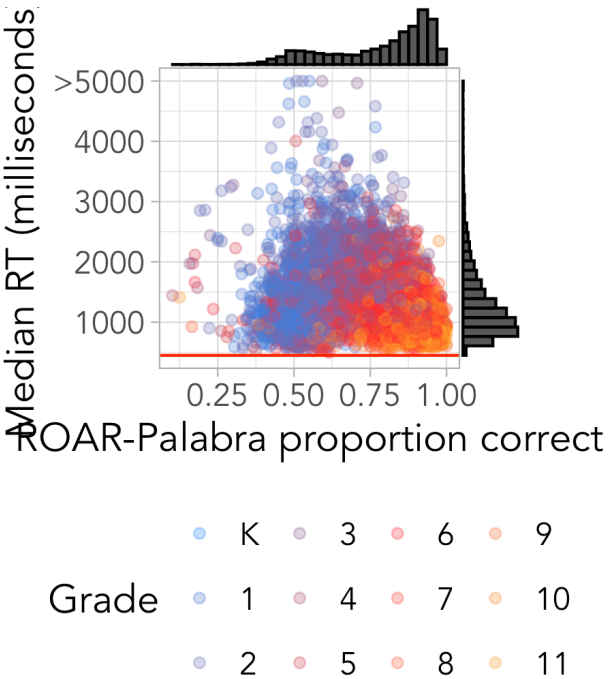
20 Reliability of ROAR-Palabra

20.1 Background: Published studies

Bhat et al. (2024) reported a large study (N=1,337) examining the relationship between reading skills, math skills, and phonological awareness in Spanish speaking students in Colombia. In this sample, the reported marginal reliability of ROAR-Palabra was 0.92, reliability of ROAR-Frase was 0.82, and ROAR-Fonema was 0.85.

20.2 Criteria for identifying disengaged participants and flagging unreliable scores

To account for unreliable scores and disengaged participation (as discussed in Chapter 15 and shown in Chapter 11), participants with a median response time <450ms are flagged in ROAR-Score reports and their data is excluded from analyses.



20.3 Reliability of fixed-length ROAR-Palabra

ROAR-Palabra runs as fixed-length test and scores are computed based on a Rasch model. The current version of ROAR-Palabra takes about 5 minutes (70 items). In the near future, a computer-adaptive version will become available leveraging the full item bank. Then, more items can be administered for a more precise measure or fewer items can be administered as a quick screener. Table 20.1 reports marginal reliability computed based on data from 5408 students under the IRT model for the standard, 70 item version of ROAR-Palabra. Reliability ($\rho_{xx'}$) is computed based on the estimated variance of $\hat{\theta}$ relative to the estimated standard error ($\widehat{SE}(\hat{\theta})^2$) using Equation 20.1:

$$\hat{\rho}_{xx'} = \frac{\widehat{VAR}(\hat{\theta})}{\widehat{VAR}(\hat{\theta}) + \widehat{SE}(\hat{\theta})^2}, \tag{20.1}$$

Table 20.1: Reliability of ROAR-Word by Grade

Grade	Empirical Reliability	N
All	0.9352133	5408
K	NA	1
1	0.7495653	412
2	0.8847775	638
3	0.9147536	515
4	0.9126525	510
5	0.9133743	580
6	0.8850746	589
7	0.8439554	509
8	0.8345813	423
9	0.8145449	412
10	0.7855567	421
11	0.7853985	398

To ensure that ROAR-Palabra is fair and equitable for different demographic groups, we also report reliability by gender (Table 20.2), eligibility for free and reduced price lunch (Table 20.3), English learner status based on state of California designations (Table 20.4), primary langauge spoken (Table 20.5), special education (Table 20.6), ethnicity (Table 20.7), and race (@Table 20.8)

Table 20.2: Reliability of ROAR-Word by Gender

Gender	Empirical Reliability	N
All	0.9352133	4349
F	0.9387689	2189

M	0.9326198	2160

Table 20.3: Reliability of ROAR-Word by FRL (California Sub-sample Only)

Free/Reduced Lunch Status	Empirical Reliability	N
All	0.9352133	250
F	0.8552586	120
P	0.8620204	84
R	0.8713555	46

Table 20.4: Reliability of ROAR-Word by EL Status (California Sub-sample Only)

English Learner Status	Empirical Reliability	N
All	0.9352133	250
EL	0.8535509	142
EO	0.8599942	70
IFEP	0.8590431	22
RFEP	0.9009931	15
TBD	NA	1

Table 20.5: Reliability of ROAR-Word by Primary Language (California Sub-sample Only)

Primary Language	Empirical Reliability	N
All	0.9352133	239
English	0.8741980	123
Spanish	0.8439935	116

Table 20.6: Reliability of ROAR-Word by Special Education Status (California Sub-sample Only)

Special Education Status	Empirical Reliability	N
All	0.9352133	250
0	0.8596218	234
1	0.8255854	16

Table 20.7: Reliability of ROAR-Word by Hispanic Ethnicity (California Sub-sample Only)

Hispanic Ethnicity	Empirical Reliability	N
All	0.9352133	232
0	0.7928431	13

1	0.8529172	219

Table 20.8: Reliability of ROAR-Word by Race (California Sub-sample Only)

Race	Empirical Reliability	N
All	0.9352133	232
Asian	0.8859822	2
Black or African American	0.0016945	2
Filipino	NaN	NaN
Hawaiian or Other Pacific Islander	NA	1
Hispanic	0.8529172	219
White	0.8274541	8

References

Bhat, Kruttika G., Alexa Mogan, Ana Saavedra, Mia Fuentes-Jimenez, Julian M. Siebert, Wanjing Anya Ma, Carrie Townley-Flores, et al. 2024. "Shared and Unique Influences of Phonological Processing on Reading and Math[a]."

[a]

21 Reliability of ROAR-Frase

ROAR-Frase is a timed Spanish reading measure where the student reads sentences and decides if the statement is true or false. The score is computed as the number of correct trials minus the number of incorrect trials in the alloted period time window. Each participant completed 2 90-second blocks which randomly sampled from a large item bank. We had two administrations, one in 3 different regions in Colombia and one in a region of California where a majority of the students speak Spanish. Colombian students were primarily monolingual Spanish speakers and students in California were bilingual, but many entered school with Spanish as their primary language.

21.1 Criteria for flagging unreliable scores

ROAR-Frase is designed to be totally automated where the student can complete the assessment independent of any assistance from an educator or adult. Instructions are delivered through headphones with engaging an story-line. Additionally, students complete practice trials with feedback to ensure the task instructions are clear. Sentences are presented onscreen and reading is done silently. Students respond with their keyboards. Items are designed in a way that does not require background information to discern if a sentence is true or false.

A potential concern with automated assessments is that, in the absence of a teacher to administer items individually, monitor responses, and score them, some students may disengage from the task, leading to data that does not accurately reflect their actual abilities. ROAR-Frase, having items that are unambiguous and clear, can detect students who were not engaged during the assessment. Our approach to identifying and highlighting disengaged participants with scores that are thought to be unreliable can be seen below. Figure 21.1 shows a plot of median response time (RT) for each participant against the proportion correct on the assessment, collapsed across both 90-second blocks. It is clear that there is a bimodal distrubition that indicates a group of paritipants who were performing at chance and responding very quickly. Participants with a median response time <1,000ms and proportion correct <0.65 are flagged as unreiliable scores in the ROAR score report and removed from the following analyses as it is believed that these scores do no represent a participant's true ability.

21.2 Alternate form reliability - Colombia

Alternate form reliability for Frase is computed as the Pearson correlation adjusted with the Spearman-Brown formula between scores on the two 90-second blocks that were completed

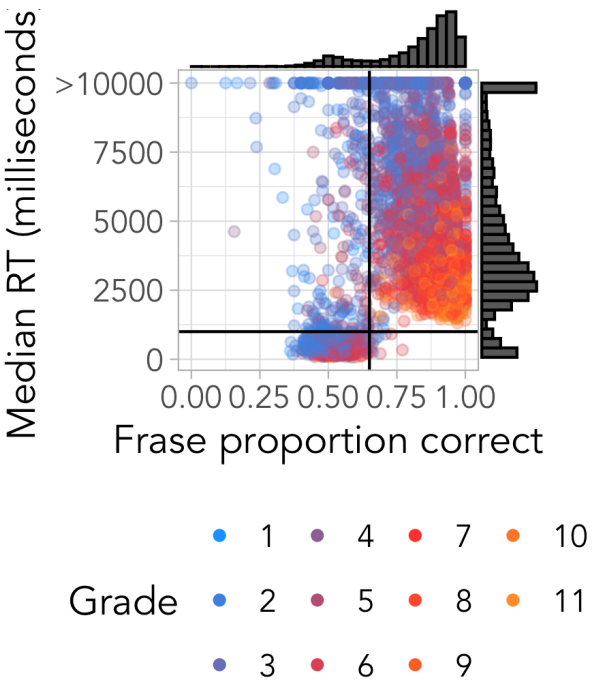


Figure 21.1: Criteria for identifying disengaged participants and flagging unreliable scores on ROAR-Frase. Participants displaying extremely rapid responses performed near chance on ROAR-Frase.

during the same testing session. Figure 21.2 shows a plot of student scores on alternate test forms combining grades and Figure 21.3 shows separate plots for each grade. Table 21.1 reports alternate form reliability for the full Colombian sample and separately by grade. Table 21.2 depicts alternate form reliability for the full Colombian sample separated by gender.

Table 21.1: Alternate form reliability for ROAR-Frase in Colombia split by student grade

Grade	Alternate Form Reliability	N
All	0.9039315	4512
2	0.6908008	413
3	0.7360957	490
4	0.7867389	475
5	0.8000892	561
6	0.7886965	569
7	0.7663706	496
8	0.8227222	417
9	0.8198782	408
10	0.8203578	346
11	0.8363014	337

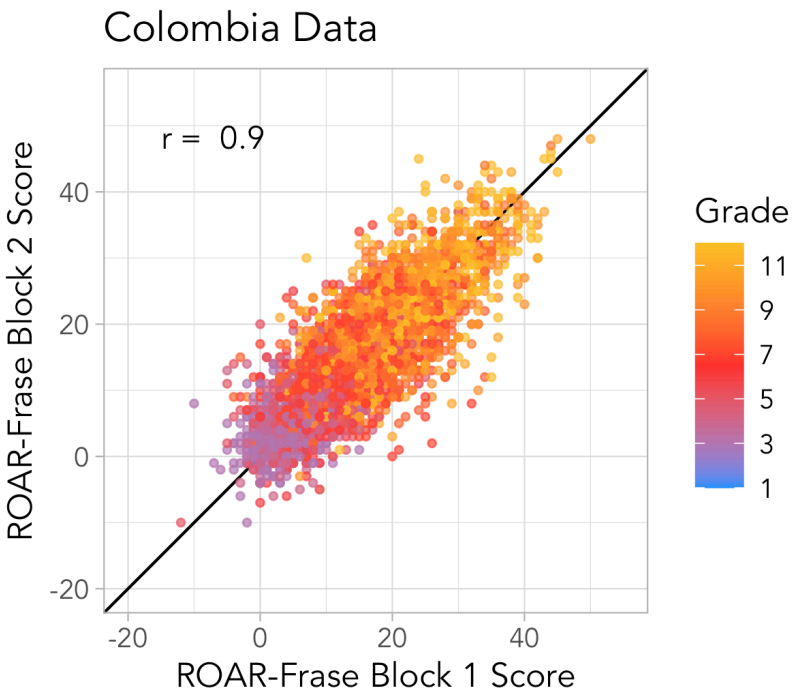


Figure 21.2: ROAR-Frase Colombia alternate form reliability across grades. Alternate form reliability is calculated as the Pearson correlation between scores on the two 90-second blocks that were completed in one sitting and adjusted by the Spearman-Brown formula.

Table 21.2: Alternate form reliability for ROAR-Frase in Colombia split by student gender

Gender	Alternate Form Reliability	N
All	0.9039315	4512
Female	0.9035495	1888
Male	0.9055784	1849
NA	0.9009216	775

21.3 Alternate form reliability - United States

Here we show reliability between blocks 1 and 2 for all United States data (California). As with Colombia data, alternate form reliability for Frase is computed as the Pearson correlation adjusted with the Spearman-Brown formula between scores on the two 90-second blocks that were completed during the same testing session. Figure 21.4 shows a plot of student scores on alternate test forms combining grades and Figure 21.3 shows separate plots for each grade.

Table 21.3 reports alternate form reliability for the California sample and separately by grade, Table 21.4 shows the breakdown of alternate form reliability by gender, Table 21.5 depicts alternate form reliability for the full California sample separated by English Learner Status, Table 21.6 shows alternate form reliability separated by primary language, Table 21.7 shows

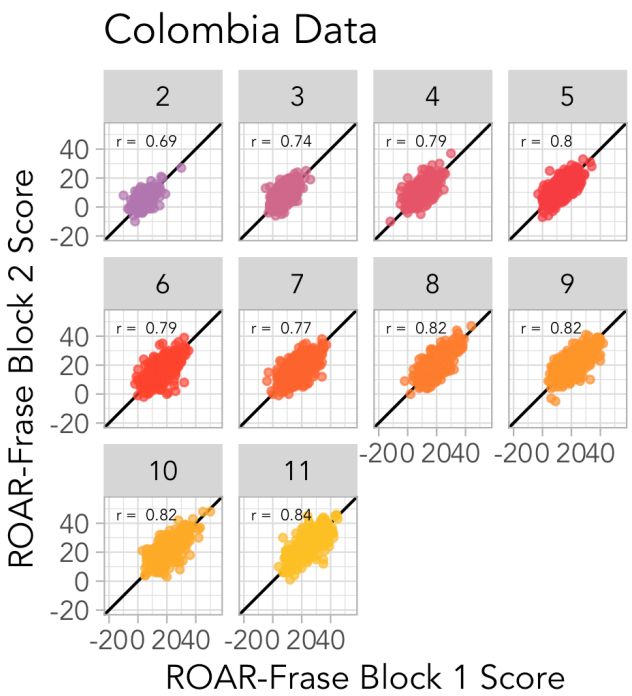


Figure 21.3: ROAR-Frase Colombia alternate form reliability within grades. Alternate form reliability is calculated as the Pearson correlation between scores on the two 90-second blocks that were completed in one sitting and adjusted by the Spearman-Brown formula.

breakdown by special education status, and finally, Table 21.8 shows breakdown of reliability by free and reduced lunch status.

Table 21.3: Alternate form reliability for ROAR-Frase in U.S. by student grade.

Grade	Alternate Form Reliability	N
All	0.7567431	256
1	0.6922893	106
2	0.7316300	150

Table 21.4: Alternate form reliability for ROAR-Frase in U.S. by student gender

Gender	Alternate Form Reliability	N
All	0.7567431	256
Female	0.7847464	118
Male	0.7271188	134
NA	0.9835240	4

Table 21.5: Alternate Form Reliability reliability for ROAR-Frase in U.S. by English Learner Status

English Learner Status	Alternate Form Reliability	N
All	0.7567431	256
English Learner	0.7096593	144
English Only	0.7932300	73
Initial Fluent English Proficiency	0.7421303	22
Reclassified Fluent English Proficiency	0.7493917	13
NA	0.9835240	4

Table 21.6: Alternate Form Reliability reliability for ROAR-Frase in U.S. by Primary Language

Primary Language	Alternate Form Reliability	N
All	0.7567431	256
English	0.7817044	117
Spanish	0.7292383	125
NA	0.8662868	14

Table 21.7: Alternate Form Reliability reliability for ROAR-Frase in U.S. by Special Education Status

Special Education Status	Alternate Form Reliability	N
All	0.7567431	256
Yes	0.7913782	15
No	0.7498295	237
NA	0.9835240	4

Table 21.8: Alternate Form Reliability reliability for ROAR-Frase in U.S. by Free and Reduced Lunch Status

Free and Reduced Lunch	Alternate Form Reliability	N
All	0.7567431	256
Pays	0.8066802	82
Reduced	0.8148188	45
Free	0.6528948	125
NA	0.9835240	4

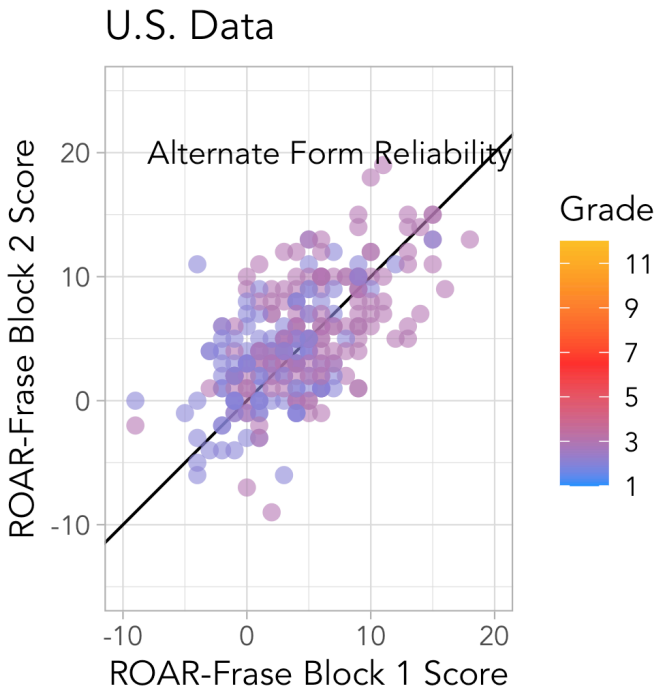


Figure 21.4: ROAR-Frase U.S. alternate form reliability across grades. Alternate form reliability is calculated as the pearson correlation between scores on the two 90-second blocks that were completed in one sitting and adjusted by the Spearman-Brown formula.

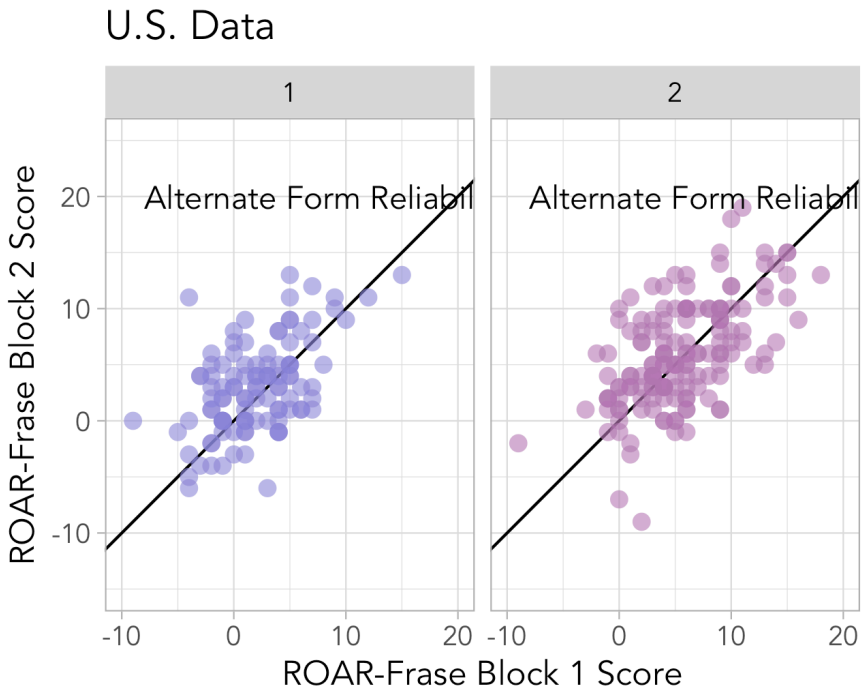


Figure 21.5: ROAR-Frase U.S. alternate form reliability within grade. Alternate form reliability is calculated as the pearson correlation between scores on the two 90-second blocks that were completed in one sitting and adjusted by the Spearman-Brown formula.

Part V

Construct Validity: Evidence that ROAR subtests reliably measure the intended constructs

22 Single Word Recognition (ROAR-Word) Concurrent Validity

ROAR-Word is designed to measure the latent construct of single word reading. Traditionally, single word reading is measured by having children read lists of real words and pseudo words of increasing complexity and scoring them based on their accuracy of pronunciation. Thus we first establish that the silent, lexical decision task in ROAR-Word taps into the same latent construct by comparing ROAR-Word scores to a variety of other standardized measures of single word reading (Section 22.1). After establishing convergent validity with carefully selected criterion measures, we next establish divergent validity from measures of receptive vocabulary (**?@sec-divergent-validity-swr**). Finally, since single word reading is the foundation upon which reading fluency and comprehension is built, we also examine how well ROAR-Word scores predict more distal reading measures (**?@sec-validity-with-distal-measures-swr**).

22.1 Convergent validity with oral measures of single word reading

22.1.1 Woodock Johnson Basic Reading Skills

22.1.1.1 Background: Published studies

In an initial proof-of-concept study we compared proportion correct on a pilot version of ROAR-Word to individually administered Woodcock-Johnson Letter Word Identification (WJ-Word-ID) scores and found an exceptionally high correlation (r = 0.91, disattenuated r = 0.94; (Yeatman et al. 2021); Figure 22.1). Moderation analysis confirmed that ROAR-Word is equally valid for children with dyslexia and typical readers (6-18 years of age). Different measures of single word reading, such as the Woodcock-Johnson (WJ) and Test of Word Reading Efficiency (TOWRE), are highly correlated, and a variety of standardized measures largely tap into the same latent construct. For example, it is common to use a threshold on WJ or TOWRE to group research participants into dyslexic versus control groups (Fletcher et al. 2006). The correlation between timed, untimed, real word, and pseudoword reading measures across these assessments ranged from r = 0.72 to 0.93; ROAR-Word is similarly correlated with each measure (Figure 22.1). Thus, in terms of convergent validity, ROAR-Word is highly correlated with myriad measures that are often used interchangeably in reading and dyslexia research.

In a second proof-of-concept study we optimized the measurement scale with IRT, added additional items tailored to younger participants, and deployed a new version of ROAR-Word that

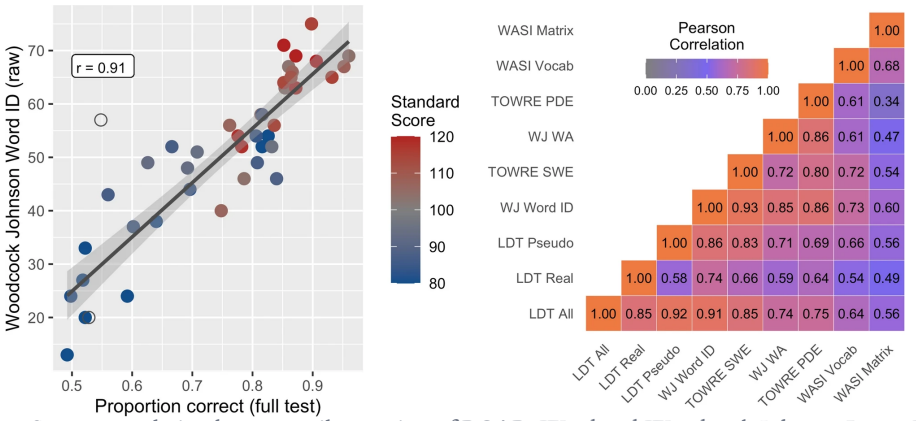


Figure 22.1: Strong correlation between pilot version of ROAR-Word and Woodcock Johnson Letter Word ID

was half the length. Data from kindergarten, first, and second grade students revealed an exceptionally high correlation of r=0.97 between ROAR-Word and WJ-Word-ID (Figure 22.2) (Yeatman et al. 2021).

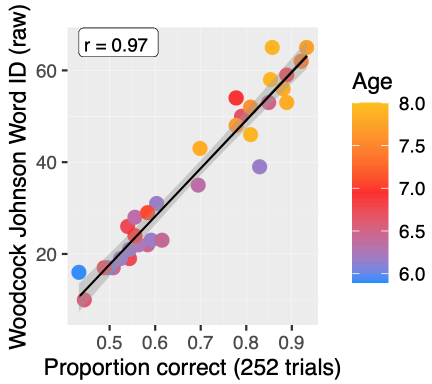


Figure 22.2: Strong correlation between pilot version of ROAR-Word and Woodcock Johnson Letter Word ID in grades K-2

22.1.1.2 Additional validation against Woodcock Johnson Basic Reading Skills

The initial validation studies published in (Yeatman et al. 2021) provided strong initial evidence that ROAR-Word accurately tapped into the construct of single word reading ability. However, the sample published in (Yeatman et al. 2021) was recruited to participate in research studies in the Yeatman Lab and was not, therefore, representative of the diversity of students in the United States. Hence we undertook a series of additional validation studies in collaboration with school districts that had adopted ROAR. Figure 22.3 shows the age distribution and Table 22.1 shows the demographics of the students that participated in these validation studies.

Table 22.1: Demographics of participants in concurrent validity study of Woodcock Johnson Basic Reading Skills (WJ BRS) and ROAR-Word

	N	%	% Missing
Female	185	45.45	7.13
Free or Reduced Lunch	51	12.53	84.77
Race/Ethnicity			
Hispanic Ethnicity	42	10.32	0.00
White	193	47.42	0.00
Black or African American	22	5.41	0.00
Asian	89	21.87	0.00
American Indian or Alaska Native	5	1.23	0.00
Hawaiian or Other Pacific Islander	2	0.49	0.00
Multiracial	51	12.53	0.00
Total	407		

Figure 22.4 shows the relationship between ROAR-Word raw scores and Woodock Johnson Letter Word Identification (WJ LWID) assessed at the same time point and Figure 22.5 shows this relationship broken down by grade. The strongest relationships are in early elementary school but that is because single word reading is at ceiling by late elementary school for most students. Table 22.2 reports the correlations separately for each grade.

When comparing the correlation between ROAR-Word scores and individually administered Woodcock Johnson (WJ) assessments, the correlation is strongest in early elementary school and decreases in middle school and high-school (Table 22.2). But this decrease in correlation is likely driven by the fact that most middle school and high school students are at ceiling in single word reading restricting the range of WJ scores. We used the correlation between the two WJ subtests (Word Identification and Word Attack) as a reliability estimate for WJ subtest scores in each age range and found, as expected, that a decrease in reliability of the WJ explained the lower correlation between ROAR and WJ in the older grades. Table 22.3 shows disattenuated correlations that correct for differences in reliability in each sample.

Table 22.2: Pearson correlations (ρ) between ROAR-Word and Woodcock Johnson Scores by grade. Basic Reading Skills is calculated by summing Word ID and Word Attack subtests. WJ Reliability is calculated as the Pearson correlation between Word ID and Word Attack subtests. This reliability metric gives an upper bound on the correlation for ROAR-Word with either subtest

Grade	Word ID	Word Attack	Basic Reading Skills	WJ Reliability	N
1	0.81	0.81	0.81	0.84	43
2	0.87	0.87	0.88	0.82	48
3	0.77	0.77	0.77	0.82	226
4	0.72	0.72	0.71	0.88	68
5	0.69	0.69	0.62	0.85	48
6–8	0.64	0.64	0.63	0.78	83
9–12	0.53	0.53	0.59	0.65	76

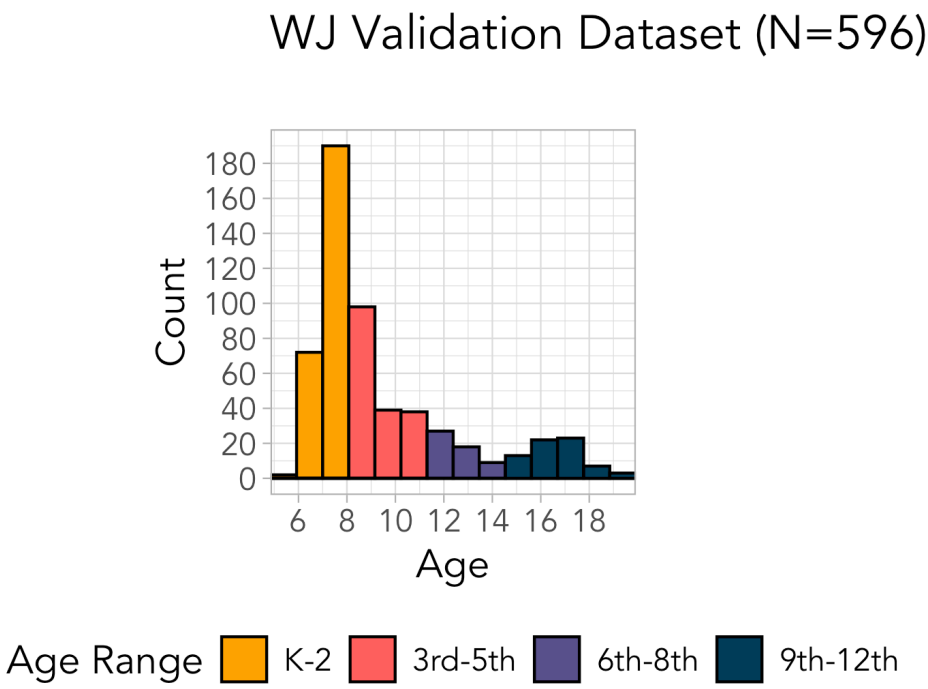


Figure 22.3: Age distribution of concurrent validity study of Woodcock Johnson Basic Reading Skills (WJ BRS) and ROAR-Word

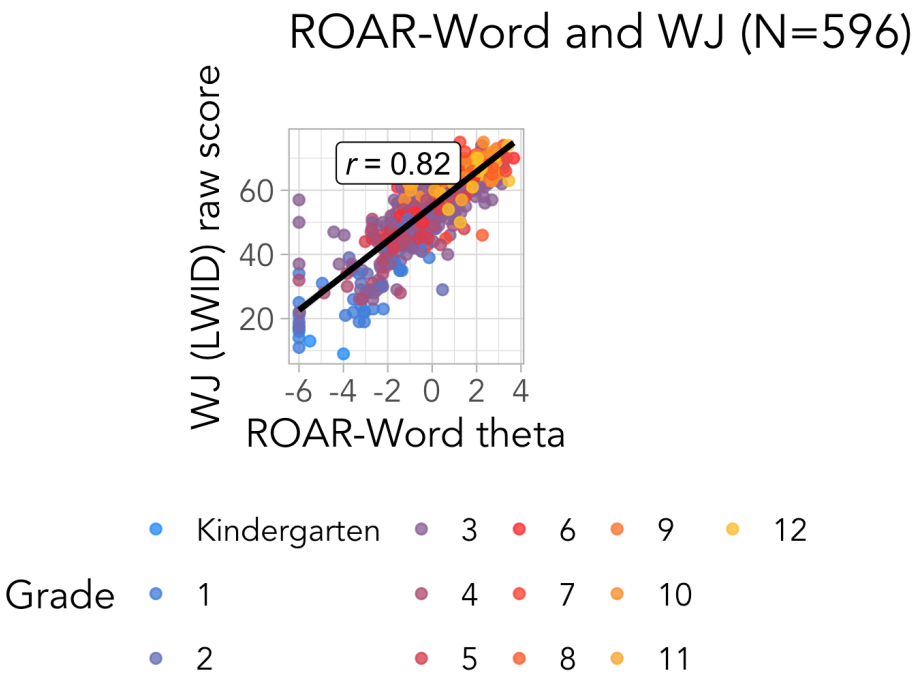


Figure 22.4: ROAR-Word is highly correlated with Woodcock Johnson Letter Word Identification (WJ LWID)

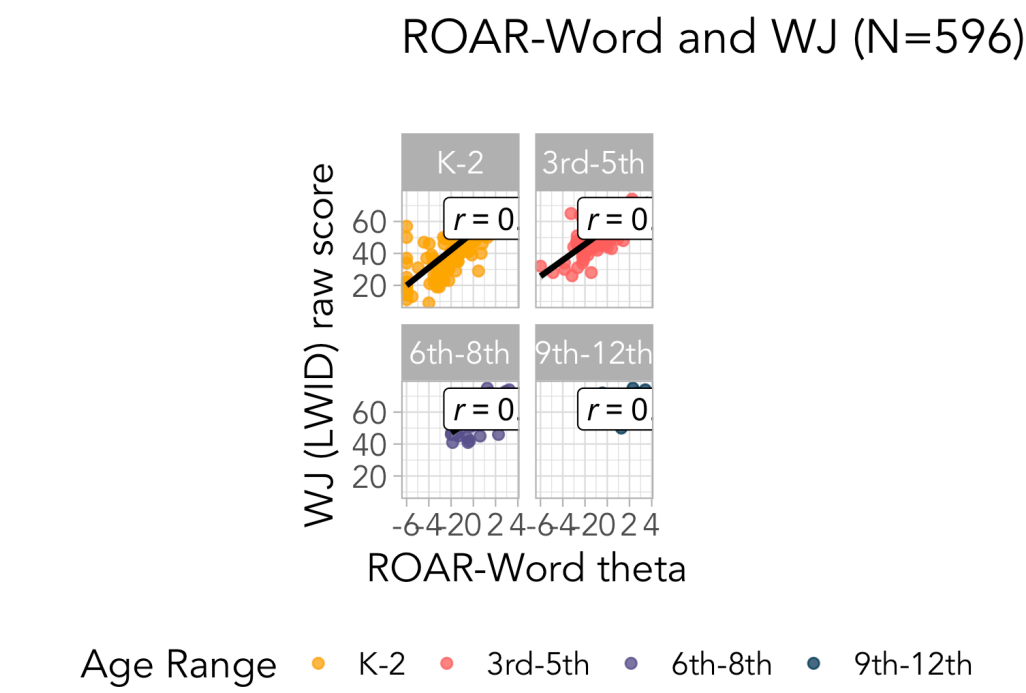
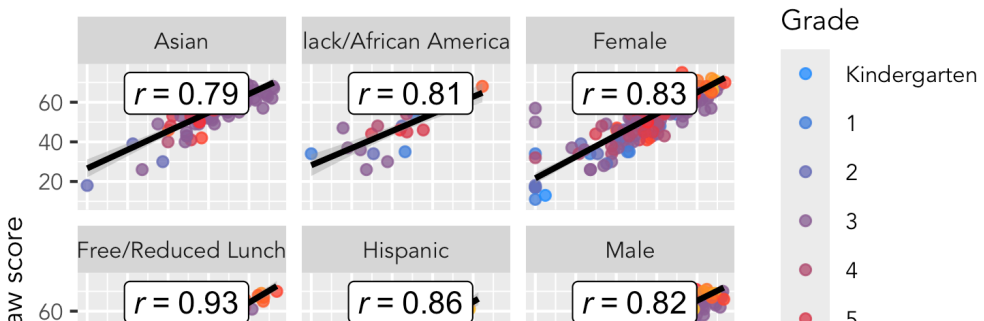


Figure 22.5: Correlations between ROAR-Word and Woodcock Johnson Letter Word Identification across different grade bands.

Table 22.3: Disattenuated correlations between ROAR-Word and Woodcock Johnson Scores by grade. The correlation between WJ Word ID and WJ Word Attack was used as an estimate of WJ reliability in each sample and ROAR-Word reliability was taken from Chapter 15.

Grade	Word ID	Word Attack	N
1	0.92	0.92	43
2	0.99	0.99	48
3	0.88	0.88	226
4	0.79	0.79	68
5	0.77	0.77	48
6–8	0.75	0.75	83
9–12	0.69	0.69	76

Figure 22.6 shows concurrent validity data comparing ROAR-Word and WJ split by demographic groups. The relationship between ROAR-Word and WJ is similar across different demographics.



22.1.2 Fastbridge

22.1.2.1 Background: Published studies

The Formative Assessment System for Teachers (FAST) from FastBridge Learning[1], is a screener and curriculum based measure widely used across many schools in the United States. In a published validation study of the new, shortened computer adaptive version of ROAR-Word which is now the current standard of practice (ROAR-CAT), we compared the θ estimates from ROAR-Word against the individually-admininstered FAST™ earlyReading measure and found a correlation of r=0.89 in 1st grade and r=0.73 in 2nd grade. This initial published study was a small sample size but provided strong evidence of construct validity for the new computer adaptive measure. Figure 22.7 reproduces Figure 10 from Ma et al. (2023) which shows convergent validity of the shortened CAT version of ROAR-Word with FastBridge and Fountas & Pinnell. The following sections undertake similar analyses with a much larger sample including multiple school districts

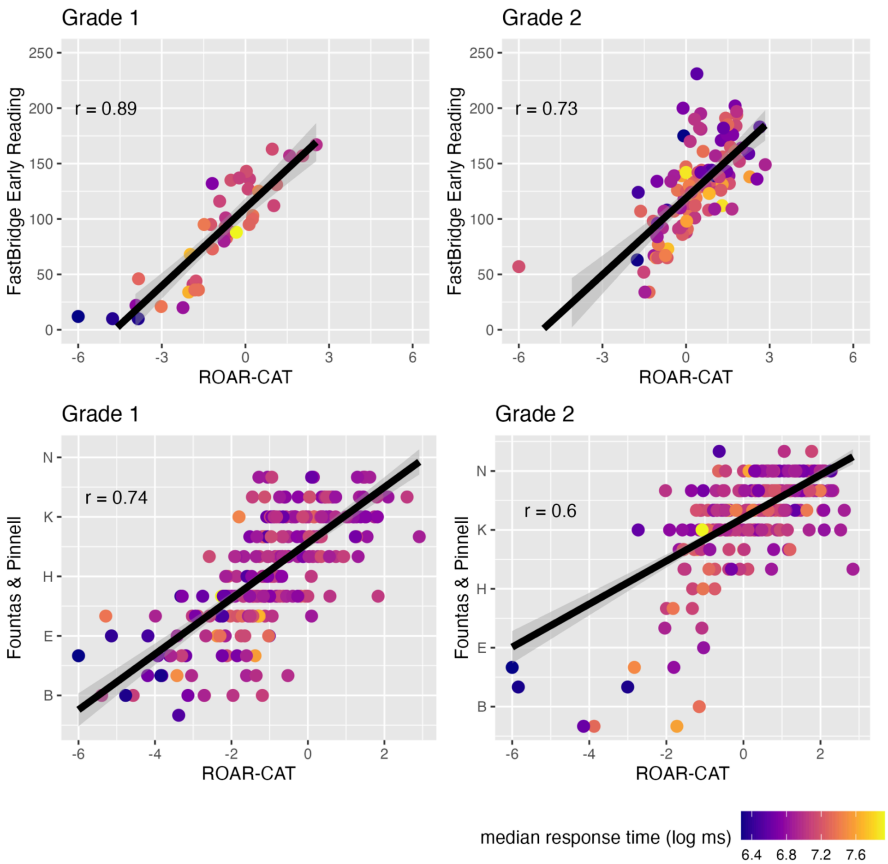


Figure 22.7: Strong correlation between Computer Adaptive ROAR-Word (ROAR-CAT) and FastBrdige in grades 1-2.

[1] https://support-content.fastbridge.org/FAST_Research/FAST_Technical_Manual_Version_FINAL.pdf

22.1.2.2 Additional validation against Fastbridge

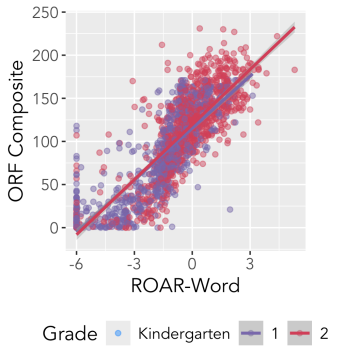
In collaboration with two large and diverse school districts in the State of California, we ran a study of concurrent validity to compare ROAR against FastBridge. Table ?? shows the demographics of the sample.

	N	%	% Missing
Female	1692	50.24	0.65
Free or Reduced Lunch	520	15.44	11.13
English Learner	494	14.67	11.13
Special Education Status	113	3.36	11.07
Race/Ethnicity			
Hispanic Ethnicity	703	20.87	0.03
White	1210	35.93	0.03
Black or African American	76	2.26	0.03
Asian	927	27.52	0.03
American Indian or Alaska Native	20	0.59	0.03
Hawaiian or Other Pacific Islander	18	0.53	0.03
Multiracial	540	16.03	0.03
Total	3368		

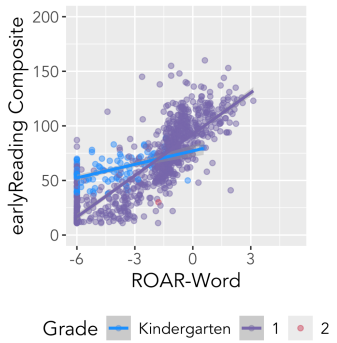
We compared ROAR-Word scores against following FastBridge measures administered within a month (concurrent validity):

- FastBridge Curriculum Based Measurement for Reading (FAST™ CBMreading) is an Oral Reading Fluency (ORF) measure where students read a leveled passage out loud for one minute. The FastBridge technical manual states "CBMreading is a simple, efficient, evidence-based assessment used for universal screening in grades 1 through 8, and progress monitoring for grades 1-12"(Christ and Colleagues 2018, 14). Words Read Correct, or WRC, "is the primary metric used in reporting student performance on FAST™ CBMreading" (Christ and Colleagues 2018, 19). This measure includes scores on three separate passages as well as a composite score.
- FAST™ earlyReading is designed to measure component skills of reading in kindergarten and first grade (Christ and Colleagues 2018, 30). It includes Sight Word (real word list) and Nonsense Word (pseudoword list) decoding measures.

Figure 22.8 shows the relationship between FAST™ CBMreading, FAST™ earlyReading and ROAR-Word scores. Table 22.5 reports the Pearson correlations between each measure. Correlations between all the measures were exceptionally high and the correlation between ROAR-Word and FastBridge was almost as high as the internal consistency of FastBridge measures.



(a) ROAR-Word is correlated with FAST™ CBMreading Oral Reading Fluency



(b) ROAR-Word is correlated with FAST™ earlyReading Composite

Figure 22.8: ROAR-Word is highly correlated with FAST™ CBMreading and FAST™ earlyReading

Table 22.5: Convergent validity of ROAR-Word: Comparision to FastBridge

	ROAR-Word	ORF Passage 1	ORF Passage 2	ORF Passage 3	ORF Composite	Nonser
ROAR-Word	**1.00**	**0.82**	**0.82**	**0.82**	**0.83**	
ORF Passage 1	0.82	1.00	0.96	0.96	0.98	
ORF Passage 2	0.82	0.96	1.00	0.96	0.99	
ORF Passage 3	0.82	0.96	0.96	1.00	0.98	
ORF Composite	0.83	0.98	0.99	0.98	1.00	
Nonsense Words	0.75	0.86	0.84	0.84	0.85	
Sight Words	0.78	0.87	0.87	0.87	0.87	
earlyReading Composite	0.76	0.90	0.90	0.91	0.91	

Table 22.6 Shows the correlation between FAST™ earlyReading ROAR-Word in kindergarten (ORF is not typically administered until first grade). Table 22.7 shows the correlations for first grade and, Table 22.8 shows the correlations for second grade.

Table 22.6: Convergent validity of ROAR-Word: Comparision to FAST™ earlyReading in kindergarten

	ROAR-Word	Nonsense Words	Sight Words	earlyReading Composite
ROAR-Word	**1.00**	**0.60**	**0.63**	**0.58**
Nonsense Words	0.60	1.00	0.81	0.90
Sight Words	0.63	0.81	1.00	0.90
earlyReading Composite	0.58	0.90	0.90	1.00

Table 22.7: Convergent validity of ROAR-Word: Comparision to FAST™ earlyReading in

	ROAR-Word	ORF Passage 1	ORF Passage 2	ORF Passage 3	ORF Composite	Nonser
ROAR-Word	**1.00**	**0.81**	**0.83**	**0.82**	**0.83**	
ORF Passage 1	0.81	1.00	0.96	0.96	0.98	

ORF Passage 2	0.83	0.96	1.00	0.97	0.99
ORF Passage 3	0.82	0.96	0.97	1.00	0.99
ORF Composite	0.83	0.98	0.99	0.99	1.00
Nonsense Words	0.73	0.86	0.84	0.84	0.85
Sight Words	0.77	0.87	0.87	0.87	0.87
earlyReading Composite	0.79	0.90	0.90	0.91	0.91

Table 22.8: Convergent validity of ROAR-Word: Comparision to FAST™ earlyReading in second grade

	ROAR-Word	ORF Passage 1	ORF Passage 2	ORF Passage 3	ORF Composite	earlyRe
ROAR-Word	**1.00**	**0.80**	**0.79**	**0.78**	**0.80**	
ORF Passage 1	0.80	1.00	0.95	0.95	0.98	
ORF Passage 2	0.79	0.95	1.00	0.95	0.98	
ORF Passage 3	0.78	0.95	0.95	1.00	0.98	
ORF Composite	0.80	0.98	0.98	0.98	1.00	
earlyReading Composite	NA	NA	NA	NA	NA	

References

Christ, and Theodore Colleagues. 2018. *Formative Assessment System for Teachers™ Technical Manual*. Author; FastBridge Learning.

Fletcher, J M, G R Lyon, L S Fuchs, and M A Barnes. 2006. *Learning disabilities: From identification to intervention*. New York: Guilford PRess.

Ma, Wanjing A, Adam Richie-Halford, Amy Burkhardt, Klint Kanopka, Clementine Chou, Benjamin Domingue, and Jason D Yeatman. 2023. "ROAR-CAT: Rapid Online Assessment of Reading Ability with Computerized Adaptive Testing."

Yeatman, Jason D, Kenny An Tang, Patrick M Donnelly, Maya Yablonski, Mahalakshmi Ramamurthy, Iliana I Karipidis, Sendy Caffarra, et al. 2021. "Rapid Online Assessment of Reading Ability." *Sci. Rep.* 11 (1): 6396.

23 Sentence Reading Efficiency (ROAR-Sentence) Concurrent Validity

23.1 Convergent validity with silent sentence reading fluency

23.1.1 ROAR-Sentence (ROAR-SRE) correlation with Woodcock-Johnson Sentence Reading Fluency (WJ-SRF)

ROAR-Sentence is designed to measure the latent construct of silent sentence reading efficiency, which represents the speed or efficiency with which a student can read simple sentences for understanding. The goal of the ROAR-Sentence task is to isolate reading efficiency by minimizing comprehension demands while maintaining checks for understanding.

We establish concurrent validity for ROAR-Sentence through a large-scale validation study that compares student performance on ROAR-Sentence to performance on the Woodcock-Johnson Sentence Reading Fluency (WJ-SRF) subtest, (Schrank et al. 2014). The development of ROAR-Sentence and the results of the validation study are detailed in study 3 of (Tran et al. 2023).

23.1.1.1 Background

Our goal in designing a new silent sentence reading efficiency measure was to more directly target reading efficiency by designing simple sentences that are unambiguously true or false and have minimal requirements in terms of vocabulary, syntax and background knowledge.

Traditional measures that are most similar to ROAR-Sentence are sometimes referred to as sentence reading fluency tasks, and while they are not administered online, they do elicit silent responses from students. The Woodcock Johnson (WJ) Tests of Achievement "Sentence Reading Fluency" subtest (Schrank et al. 2014), relies on an established design: A student reads a set of sentences and endorses whether each sentence is true or false. A student endorses as many sentences as they can within a fixed time limit (usually three minutes). The final score is the total number of correctly endorsed sentences minus the total number of incorrectly endorsed sentences.

The WJ-SRF is standardized to be administered in a one-on-one setting and the stimuli consist of printed lists of sentences which students read silently and mark Yes/No with a pencil to endorse the sentences as true or false (Schrank et al. 2014). Even though the criteria for item development on these assessments is not specified in detail, there is a growing literature showing the utility of this general approach. A similar paper-based silent reading assessment, the Test of

Silent Reading Efficiency and Comprehension (TOSREC), also involves endorsing sentences as TRUE/FALSE during a 3 minute time period. It is straightforward to administer and score and has exceptional reliability (Johnson, Pool, and Carter 2011; Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. 2010; Wagner 2011).

23.1.1.2 Participants

Participants for the validation study were recruited through two methods. The first validation sample was obtained from a longitudinal study of children with dyslexia (ages 8–14; grades 2–8; and adults ages 19–34) and a study of visual attention (children ages 7–17; grades 1–11). In these studies trained researcher coordinators individually administered standardized assessments and participants then completed ROAR-Sentence (Tran et al. 2023).

The second validation sample comprised 3rd grade students from a local school district that agreed to participate in the validation study (see Table S1 in (Tran et al. 2023) for school demographics). 3rd grade was selected for validation because it is the most common age for a dyslexia diagnosis. To conduct in-person validations in schools, a team of 7 researcher coordinators administered assessments to the students. All research coordinators completed human subjects research training, practiced extensively, and shadowed senior administrators before conducting assessments on students. Each research coordinator completed training with feedback until they were able to reliably administer each assessment. The selection of students was based on the interest of parents and teachers. Prior to the research, parents and guardians were given the opportunity to opt their students out of the research. Teachers were also informed, and their interest in the research was conveyed to the district superintendent, who then notified the research team.

Research into ROAR-Sentence is ongoing, and a third validation sample was collected after (Tran et al. 2023) was submitted for preprint. Sample 3 includes students in grades 1–8. It was collected from a private school in a low-income urban neigborhood in California by the same team of research coordinators who collected Sample 2.

Demographics for the sample are shown in #tbl-sre-wj-demographics. The distribution of participants by age is shown in #fig-sentence-age-histogram.

	N	%	% Missing
Female	125	49.21	4.72
Free or Reduced Lunch	14	5.51	94.49
Race/Ethnicity			
Hispanic Ethnicity	20	7.87	4.72
White	97	38.19	4.72
Black or African American	10	3.94	4.72
Asian	79	31.10	4.72
American Indian or Alaska Native	2	0.79	4.72
Hawaiian or Other Pacific Islander	0	0.00	4.72

Multiracial	37	14.57	4.72
Total	254		

Table 23.1: Demographics for ROAR-Sentence-WJ-SRF Validation

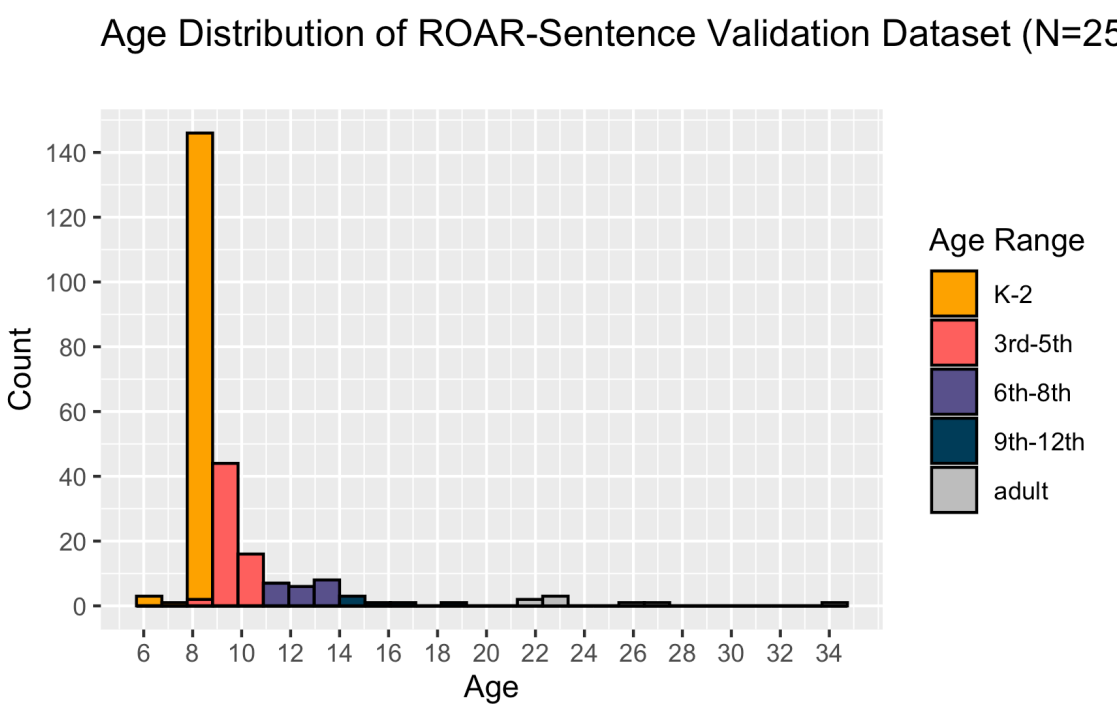


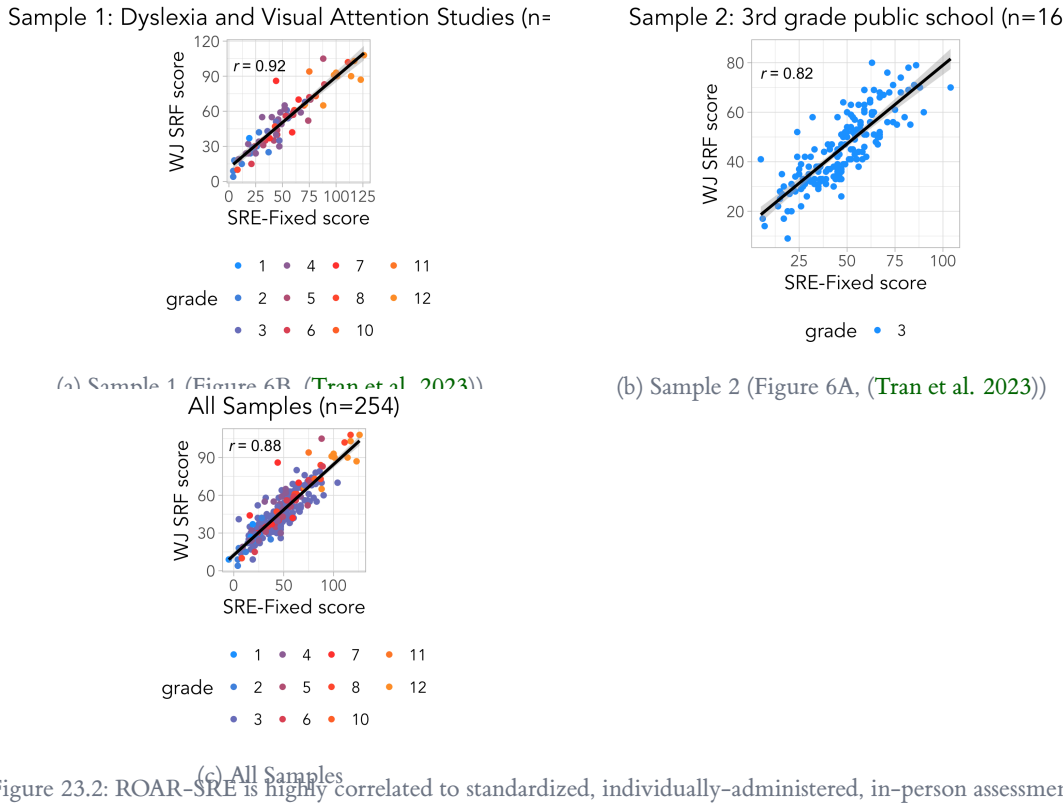
Figure 23.1: Distribution of ROAR-Sentence validation data by age

23.1.1.3 Measures

Students in Sample 1 completed individually-administered reading assessments in the course of intake screening for a longitudinal study of children with dyslexia and a separate study of visual attention. Students in Sample 2 and Sample 3 were pulled out of their classrooms to complete individually-administered reading assessments. Testing for both samples included (1) Woodcock Johnson IV Tests of Achievement Sentence Reading Fluency (WJ-SRF), in which participants silently read sentences on paper in a one-on-one setting as quickly as possible and endorse them as true or false; (2) Letter Word Identification (WJ-LWID) in which participants read words out loud and are scored for accuracy; (3) Word Attack (WJ-WA) in which participants read pseudowords out loud and are scored for accuracy (Schrank et al. 2014); (4) Test of Word Reading Efficiency Sight Word Efficiency (TOWRE-SWE) in which participants read lists of real words as quickly and accurately as possible; (5) Phonemic Decoding Efficiency (TOWRE-PDE) in which participants read lists of pseudowords as quickly and accurately as possible (Torgesen, Wagner, and Rashotte 2011). Each student completed ROAR-Sentence as part of their regular school day without the presence of researchers within 2 months prior to the in-person validation.

23.1.1.4 Results

We found a strong correlation between ROAR-SRE and WJ-SRF in Samples 1 and 2 (r=0.92, r=0.91), and across all samples (r=0.88) (#fig-sample-1-3).



(a) Sample 1 (Figure 6B, (Tran et al. 2023))

(b) Sample 2 (Figure 6A, (Tran et al. 2023))

(c) All Samples

Figure 23.2: ROAR-SRE is highly correlated to standardized, individually-administered, in-person assessments of reading fluency.

Strong correlation between ROAR-SRE and WJ-SRF was observed across all demographic groups (#fig-sre-srf-corr-by-demo).

Across all the samples, ROAR-SRE was moderately correlated with untimed single word reading accuracy (WJ-LWID, r=0.68), untimed pseudoword reading accuracy (WJ-WA, r=0.56), real word list reading speed (TOWRE-SWE, r=0.66), and pseudoword list reading speed (TOWRE-PDE, r=0.57) (#fig-sre-corr-matrix). This pattern of correlations supports the notion that sentence reading efficiency is a separable, yet highly related construct, to single word reading speed and accuracy.

In addition to examining the correlation between SRE and WJ-SRF, the study used precise timing data to investigate the optimal length for the assessment. Many assessments of sentence reading fluency/efficiency are 3 minutes by convention but previous work has not systematically analyzed the relationship between assessment length and reliability. Precise timing information collected by the application was used to calculate each participants' ROAR-SRE score at 10 second time intervals which was then correlated against the full 3 minute WJ-SRF

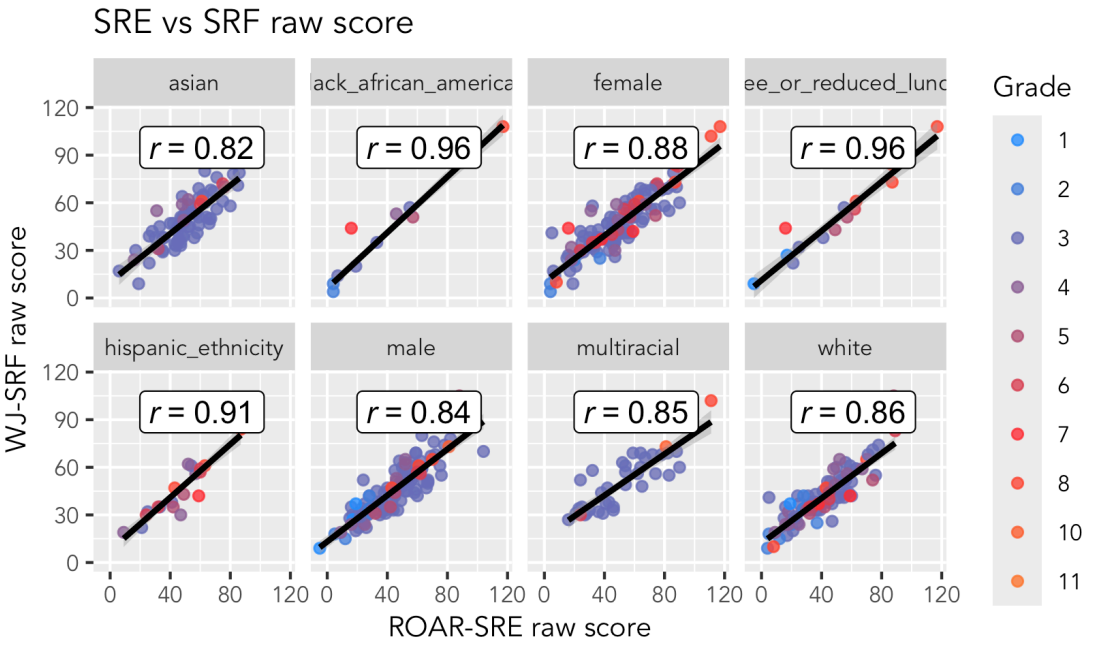


Figure 23.3: Distribution of ROAR-Sentence validation data by demographic

Figure 23.4: ROAR-SRE is moderately correlated with untimed single word reading accuracy (WJ-LWID), untimed pseudoword reading accuracy (WJ-WA), real word list reading speed (TOWRE-SWE), and pseudoword list reading speed (TOWRE-PDE),

scores. The correlation between ROAR-SRE and WJ-SRF increased as a function of assessment length. However, the correspondence between the two measures hit a peak between 60 and 90 seconds (#fig-timing) indicating that the remaining assessment time did not further contribute to the reliability of the measure.

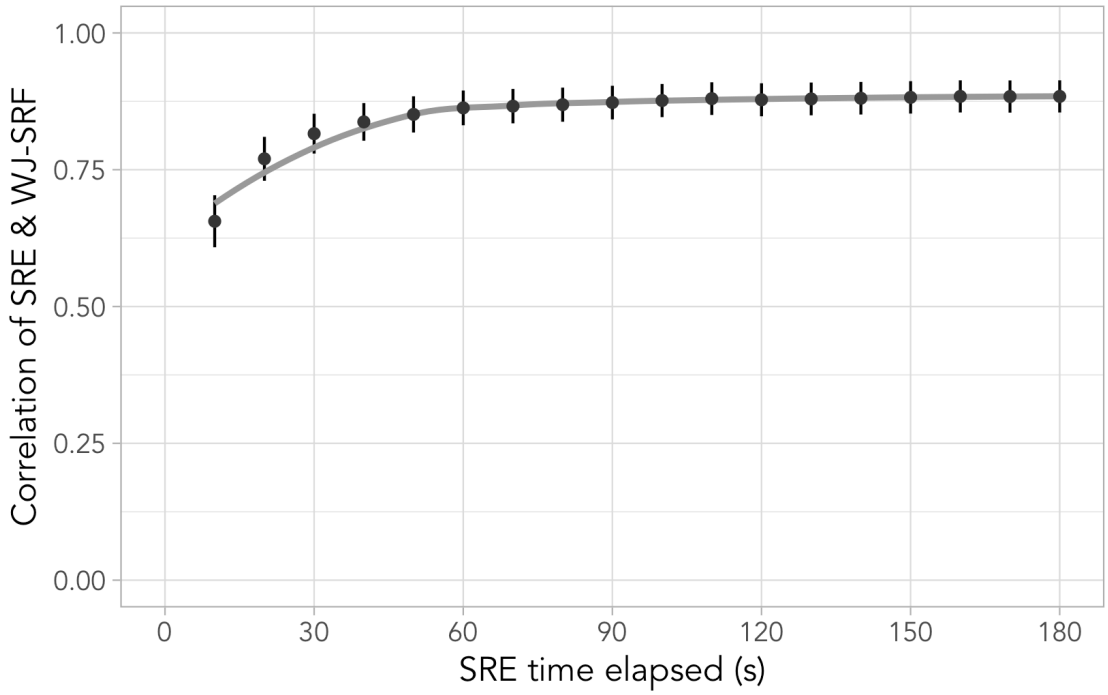


Figure 23.5: Correlation between ROAR-SRE and WJ-SRF as a function of assessment time ((Tran et al. 2023), Figure 6C)

23.1.1.5 Discussion

The study demonstrated that the unproctored, online ROAR-Sentence (ROAR-SRE) assessment was highly correlated with a similar, standardized measure delivered one-on-one in person (WJ-SRF). This provides strong evidence for the concurrent validity of an online measure. Moreover, the stronger correspondence between sentence reading (WJ-SRF) versus single word decoding (WJ-LWID and WJ-WA) and single word reading efficiency (TOWRE-SWE and TOWRE-PDE) measures demonstrated that sentence and word reading are related but dissociable constructs as highlighted in other work (Silverman et al. 2013). Finally, the analysis of assessment length demonstrated that a one minute sentence reading efficiency measure achieves high reliability. This finding opens the possibility of more regular progress monitoring with a quick and automated one minute assessment.

23.2 Convergent validity with oral reading fluency (ORF)

In collaboration with two large and diverse school districts in the State of California, we ran a study of concurrent validity to compare ROAR against FastBridge earlyReading. The Formative Assessment System for Teachers (FAST) from FastBridge Learning[1], is a screener and curriculum based measure widely used across many schools in the United States.

23.2.1 ROAR-Sentence (ROAR-SRE) validation against FastBridge earlyReading

We compared the raw from ROAR-Sentence against the individually-admininstered FAST™ earlyReading measure and found a correlation for the Oral Reading Fluency Composite of 0.85 across grades 1-3.

23.2.2 Participants

The demographics of the sample that completed both ROAR-SRE and FastBridge are displayed in Table 23.2

	N	%	% Missing
Female	902	46.78	1.24
Free or Reduced Lunch	513	26.61	20.59
English Learner	403	20.90	20.59
Special Education Status	131	6.79	20.59
Race/Ethnicity			
Hispanic Ethnicity	626	32.47	0.10
White	432	22.41	0.10
Black or African American	8	0.41	0.10
Asian	360	18.67	0.10
American Indian or Alaska Native	1	0.05	0.10
Hawaiian or Other Pacific Islander	6	0.31	0.10
Multiracial	178	9.23	0.10
Total	1928		

Table 23.2: Demographics for ROAR-Sentence-Fastbridge-Validation

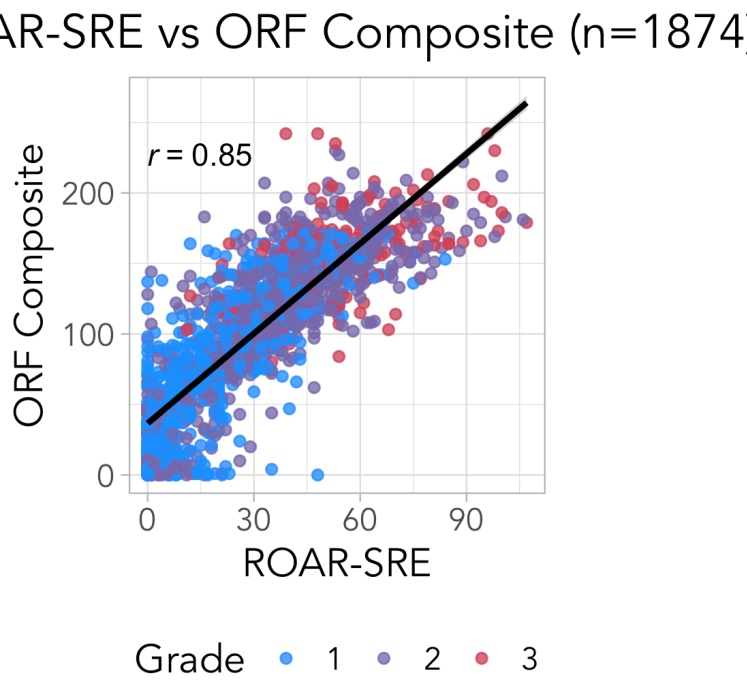


Figure 23.6: ROAR-SRE is strongly correlated with Fastbridge ORF Composite.

23.2.2.1 Results

References

Johnson, Evelyn S, Juli L Pool, and Deborah R Carter. 2011. "Validity Evidence for the Test of Silent Reading Efficiency and Comprehension (TOSREC)." *Assess. Eff. Interv.* 37 (1): 50–57.

Schrank, F A, K S McGrew, N Mather, B J Wendling, and E M LaForte. 2014. "Woodcock-Johnson IV Tests of Achievement." Riverside Publishing Company.

Silverman, Rebecca D, Deborah L Speece, Jeffrey R Harring, and Kristen D Ritchey. 2013. "Fluency Has a Role in the Simple View of Reading." *Sci. Stud. Read.* 17 (2): 108–33.

Torgesen, Joseph K, Richard Wagner, and Carl Rashotte. 2011. *TOWRE 2: Test of word reading efficiency*. Pearson Clinical Assessment.

Tran, Jasmine E, Jason D Yeatman, Amy Burkhardt, Wanjing A Ma, Jamie Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. 2023. "Development and Validation of a Rapid Online Sentence Reading Efficiency Assessment."

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. 2010. *Test of Silent Reading Efficiency and Comprehension*. Pro Ed.

Wagner, Richard K. 2011. "Relations Among Oral Reading Fluency, Silent Reading Fluency, and Reading Comprehension: A Latent Variable Study of First-Grade Readers." *Sci. Stud. Read.* 15 (4): 338–62.

[1]https://support-content.fastbridge.org/FAST_Research/FAST_Technical_Manual_Version_FINAL.pdf

24 Phonological Awareness (ROAR-Phoneme) Concurrent Validity

24.1 Background: Published studies

24.1.1 Evolution of the Design of ROAR-Phoneme items and subtests

The original ROAR-Phoneme subtests were selected based on well-known standardized in-person PA tasks (e.g., Wagner, Torgesen, and Rashotte (1999)):

- **First Sound Matching (FSM)** the participant has to find a word with the same **first sound** as the target word
- **Last Sound Matching (LSM)** the participant has to find a word with the same **last sound** as the target word
- **Rhyming (RHY)** the participant has to find the word that **rhymes** with the target word
- **Blending (BLE)** the participant has to merge parts of a word together and select the appropriate target word
- **Deletion (DEL)** the participant has to determine what is left after a section of the word is omitted.

Items for ROAR-Phoneme were designed such that selecting the correct answer would require the same cognitive operation as a traditional PA assessment with verbal responses. To achieve this, each item requires the participant to perform the same operation in their mind (e.g., determining if the first/last sound of two words matches; removing phonemes from a word), but the answer is selected from a set of alternatives rather than verbalized.

In the original design, FSM, LSM and RHY each consisted of 25 trials, divided into 2 blocks (16 and 9 items). The difference between blocks of these 3 subtests was finding the first sound (FSM), last sound (LSM), or word that rhymed (RHY) of a CVC word (difficulty level 1, 16 items) or a (C)CVC(C) word (difficulty level 2, 9 items). Thus, for the easier items (i.e., difficulty level 1) children had to identify a single phoneme (e.g., of FSM: Q: "Which picture starts with the same sound as pin?" A: "pup"), whereas for the more difficult items (i.e., difficulty level 2), children had to identify a consonant sound within a phoneme cluster (e.g., of FSM: Q: "Which picture starts with the same sound as clown?" A: "crab"). For FSM the three answer options were either the target (i.e., same first sound), a foil that started with the last sound of the provided word (Foil 1), or a foil with the same vowel (Foil 2). For LSM the same reasoning was made, but for the last sound of the word. For RHY the target word would rhyme, whereas Foil 1 would have the same vowel but would not rhyme and Foil 2 would have the same first sound. BLE and DEL each consisted of 24 items, divided into 3 difficulty levels (i.e., syllable

level, onset or rime level, phoneme level) with each 8 items. These difficulty levels were based on a suggested hierarchy within PA skills (Stanovich 2017; Treiman and Zukowski 1991; Anthony and Lonigan 2004) . For example, for the subtest DEL an item of difficulty level 1 could be: Q: "What is lipstick without stick?" A: "lip", for difficulty level 2: Q: "What is farm without 'f'?" A: "arm", and for difficulty level 3: Q: "What is snail without 'n'?" A: "sail". For both the BLE and DEL subtests, all additions and omissions led to lexical changes rather than morphological changes of the word structure. An item was either scored as correct (i.e., target selected) or as incorrect (i.e., foil selected). No distinction was made in the scores based on which foil was selected.

24.1.2 Proof-of-concept: Validation of items and composite scores

To validate the feasibility of a web-browser based PA task (containing 5 subtests: FSM, LSM, RHY, BLE, and DEL) that only required clicks/touchscreen responses, we tested 143 participants (Age: 3.87–13.00, μ=7.13, σ=1.89; Sex: 67 F, 76 M) and performed a correlation analysis between each ROAR-PA subtest and the well-established standardized CTOPP-2. The results (Fig. 1, left panel) revealed strong correlations between the CTOPP-2 and all ROAR-Phoneme subtests: LSM (r=0.65), DEL (r=0.62), FSM (r=0.61), RHY (r=0.60), and BLE (r=0.55). Each subtest, except for BLE, showed high internal consistency based on Cronbach's α (LSM: α=0.92, CI95=[0.89; 0.93], FSM: α=0.90, CI95=[0.87; 0.93], RHY: α=0.86, CI95=[0.81; 0.89], DEL: α=0.84, CI95=[0.77; 0.88], BLE: α=0.70, CI95=[0.57; 0.78]) and the composite scores of both CTOPP-2 (α=0.88, CI95=[0.85 ; 0.91]) and ROAR-PA (α=0.85, CI95=[0.80; 0.89]) had good (0.8$\leq$ α<0.9) internal consistency.

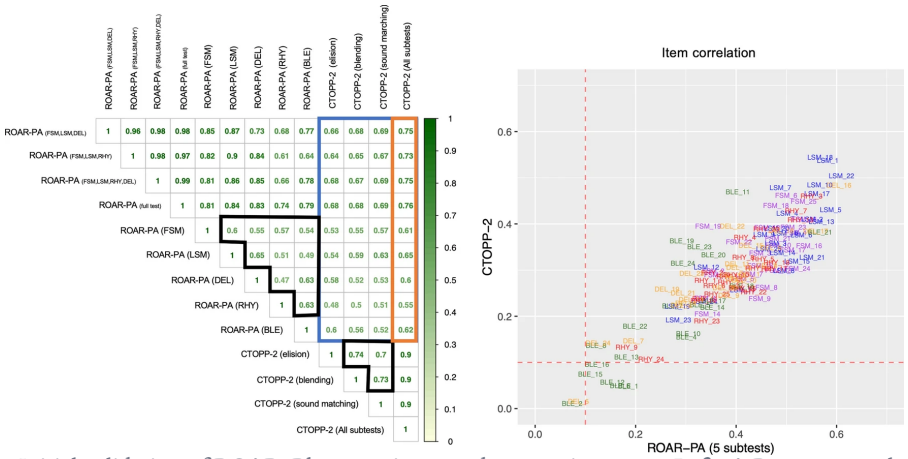


Figure 24.1: Initial validation of ROAR-Phoneme items and composite scores. **Left**: A Pearson correlation matrix between the 5 ROAR-PA subtests (% correct), CTOPP-2 subtests (raw scores), and overall (raw) scores on both tasks, of the original cohort that completed all 5 subtests (N=143). Orange box: correlation coefficients between ROAR-PA and overall CTOPP-2; Blue Box: correlation coefficients between ROAR-PA and subtests of CTOPP-2; Black boxes: correlation coefficients within subtests of ROAR-PA and subtests of CTOPP-2. **Right**: Item Correlation Analysis: For each stimulus, we plot the point-biserial correlation between performance on the item and ROAR-PA accuracy (x-axis) as well as the correlation between performance on the item and CTOPP-2 raw score (y-axis). Items with low correlations with overall test performance or CTOPP-2 performance (threshold r□.10; dotted red line) were removed from the test.

24.1.3 Optimization of ROAR-PA as a screening tool

To optimize ROAR-PA as a valid screening tool we sought to create a composite score that best approximated the CTOPP-2 composite index. To do so, we created a linear model with the CTOPP-2 scores as the dependent variable and the scores of each individual subtest as predictor variables. This model (CTOPP-2~FSM+LSM+RHY+BLE+DEL) showed that the subtests FSM (β = 0.79; t=2.33; p=0.02), LSM (β = 1.07; t=3.92; p<0.001), and DEL (β = 1.27; t=3.13; p=0.002) were significant predictors of the CTOPP-2 scores, but the subtests RHY (β = 0.44; t=1.13; p>0.10), and BLE (β = 0.55; t=0.90; p>0.10, were not. We then used a Likelihood Ratio test to determine the influence of these non-significant subtests in our CTOPP-2 prediction, by comparing the full model, as described above, to a model without BLE (4 ROAR-PA subtests: FSM, LSM, RHY, DEL); and a model with only the 3 significant subtests (FSM, LSM, DEL). We found no significant differences in model predictions between the full model and the 4 subtest model (χ^2=0.85; p>0.10), nor the full model and the 3 subtest model (χ^2=2.05; p>0.10), suggesting that the three subtests (FSM, LSM, DEL) are sufficient to obtain an accurate PA composite that approximates the CTOPP (R^2=0.57).

These findings are corroborated by interpreting the Pearson correlation coefficients between ROAR-PA and CTOPP-2. Although the highest correlation was reported by summing the scores on all 5 ROAR-PA subtests (r=0.76), a composite score based on 4 (FSM, LSM, DEL,RHY) or 3 ROAR-PA subtests (FSM, LSM, DEL) was equally correlated with CTOPP-2 (r=0.75). The 3-subtest composite and 4-subtest composite both achieved good reliability as well: Cronbach's alpha of $\alpha_{4subtests}$=0.84, CI95=[0.77; 0.88] and $\alpha_{3subtests}$=0.78, CI95=[0.67; 0.84] respectively. As convergent validity greater than r=0.70 is recommended to reflect whether two measures capture a common construct, it can be concluded that all possible composite scores (5, 4, and even 3 subtests) suffice to capture PA skills.

Furthermore, an item analysis comparing the correlations between the item responses of ROAR-PA for each of the 123 test items and CTOPP-2 scores showed that performance on items from the subtest LSM were especially highly correlated with overall ROAR-PA performance and CTOPP-2 performance (Figure 24.1). This suggests LSM items are most informative about overall PA abilities. Items from the BLE subtest were least informative: the correlation between most blending items and ROAR-PA total score and CTOPP-2 total score was close to zero.

24.1.4 Ideal age range for ROAR-Phoneme

After selecting 3 subtests that make an efficient and reliable ROAR-PA composite score, we collected ROAR-PA data for an additional group of 127 participants, including mostly older children, resulting in a total of 270 participants (Age: 3.87–14.92, μ=9.12, σ=2.71; Sex: 125 F, 145 M) who completed ROAR-PA FSM, LSM, and DEL subtests. Of these participants, 266 were also administered the CTOPP-2 PA assessment. The Pearson correlation analysis with the CTOPP-2 for this extended group of participants resulted in an overall correlation between CTOPP-2 and ROAR-PA composite (3 subtests) of r=0.70 (as opposed to r=0.75 in the initial sample of participants). The correlation between the CTOPP-2 and the individual

subtests also went down for FSM and LSM (Fig. 2. Left top). The decrease in correlation likely reflected ceiling effects in older participants (Figure 24.2).

To examine the effect of age on the correlation between ROAR-PA and CTOPP-2, we split our sample into 3 different age bins (3.87–6.99 years old (N=71), 7.00–9.99 years old (N=91), 10.00–14.92 years old (N=104)). We found a correlation coefficient between the composite scores of ROAR-PA and CTOPP-2 of r=0.79 (CI95=[0.68; 0.86], Cronbach's α=0.88) for the youngest group, r=0.69 (CI95=[0.56; 0.78], Cronbach's α=0.79) for the middle group, and r=0.31 (CI95=[0.13; 0.48], Cronbach's α=0.65) for the oldest group of children. Further correlation and Rasch analyses provided an ideal age range of up to 9.50 years old for the ROAR-PA (Figure 24.2), leading to a Pearson correlation coefficient of r=0.80 (CI95=[0.73; 0.85], Cronbach's α of 0.80) between the ROAR-PA composite and CTOPP-2, and an increase of the correlations for individual subtests (FSM, LSM, DEL) to the CTOPP-2. This indicates that the ROAR-PA in its current form is predictive of PA skills for children in pre-kindergarten through fourth grade (Figure 24.2) but has ceiling effects above fourth grade. Interestingly, the correlation analysis in our sample shows a similar effect for the CTOPP-2 scores, indicating that both PA tasks (ROAR-PA and CTOPP-2) are most suited for younger children.

24.1.5 Factor structure of Phonological Awareness

To evaluate the dimensionality of the ROAR-PA assessment we used exploratory FA with oblique rotation. FA poses the question of whether there is evidence that all of these items are measuring the same underlying phonological processing ability, or whether the items of these subtests better represent separable (but correlated) dimensions of PA.

Our results suggest a multi-dimensional framework. First, the scree plot (Figure 24.3) of the different items (N=74) on these three subtests (FSM, LSM, DEL) indicate three factors before the rate of decrease flattens. Second, the magnitudes of the loadings for the three-factor model are larger than the one-factor model. Finally, examining the factor loadings, the items from each of the three subtests cleanly separate into separate factors, with the exception of a single item: FSM_13.

24.1.6 Item Response Theory analysis: Rasch model

In a second step we identified a subset of items from ROAR-PA to remove in order to both improve model fit and reduce the length of the assessment. Given the evidence for a multi-dimensional framework, we proceeded by calibrating a Rasch Model separately for each of the subtests (FSM, LSM and DEL). In this IRT analysis we included data for all participants between 3.87 and 9.50 years old. For each subtest we reviewed four criteria, compiled from both the factor analysis and Rasch Model item fit statistics, to determine the best subset of items: (1) Does the item load on the subtest factor with a relationship > .30? (Tabachnick et al. 2012) (2) Does the item resemble a functional form when looking at empirical plots? ((Allen and Yen 2001) (3) Is the item flagged based on Rasch model fit statistics (Wright 1994)? (4) Finally, as we want items to be informative and not redundant, is the item located near two or

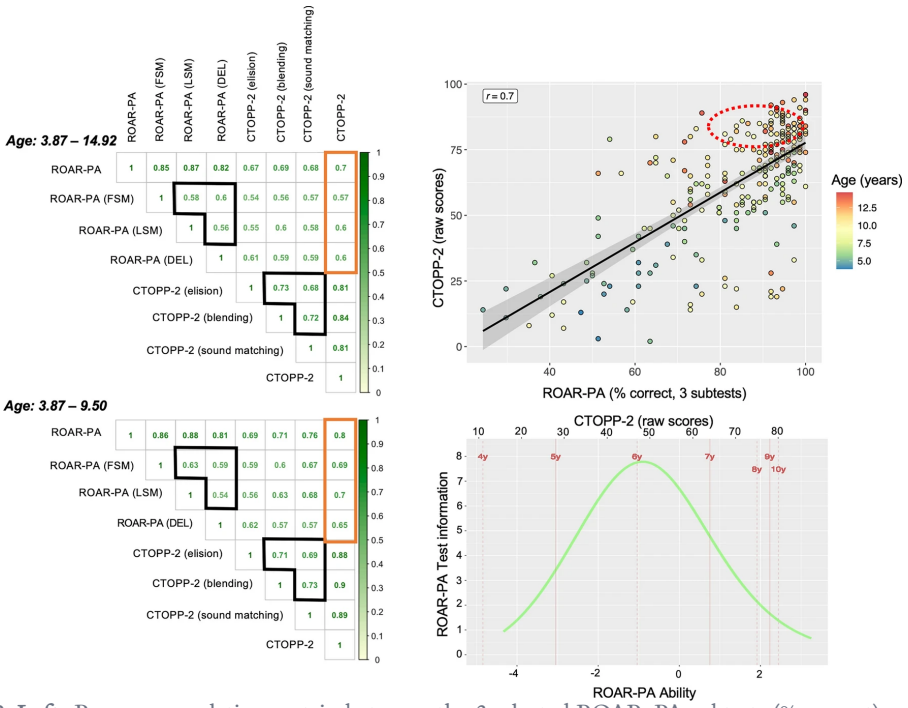


Figure 24.2: **Left**: Pearson correlation matrix between the 3 selected ROAR-PA subtests (% correct) and between the CTOPP-2 subtests (raw scores) and overall scores on both tasks for all children (age 3.87—14.92 years; N=266—top) and for a subset of children, based on the limited age-selection (age 3.87–9.50 years; N=145—bottom). Orange box: correlation coefficients between ROAR-PA and overall CTOPP-2; Black boxes: correlation boxes within subtests of ROAR-PA and subtests of CTOPP-2. **Right top**: Pearson correlation plot between the ROAR-PA (% correct) and between the CTOPP-2 (raw scores), for all children (N=266) age 3.87—14.92 years. The red dotted oval points to the ceiling effect of the oldest children. **Right bottom**: Test information functions for the ROAR-PA. The x-axis shows ability estimates based on the Rasch model. The upper x-axis shows the estimated CTOPP-2 raw score equivalent based on the linear relationship between ability estimates and CTOPP-2 scores. The pink lines indicate age equivalents for CTOPP-2 scores (based on the CTOPP-2 manual). Test information is high for participants scoring between 4.5 and 9.5 years age equivalent on the CTOPP-2.

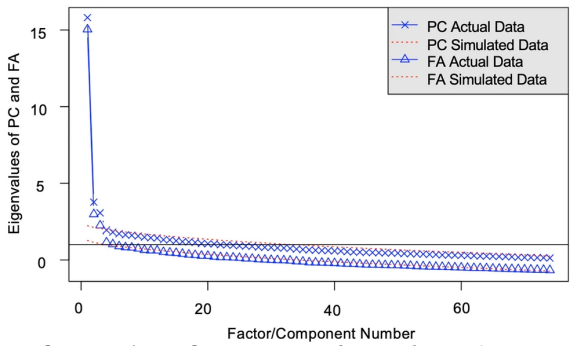


Figure 24.3: Parallel Analysis of Scree plots of 74 items on three subtests (FSM, LSM, DEL). Inspection of the scree plot suggests three factors before the amount of variation represented by the eigenvalues flattens out. The dotted red lines represent extracted eigenvalues from data sets that are randomly created; there are three factors with observed eigenvalues that are larger than those extracted from the simulated data.

more items based on difficulty distribution, to create a test length that seems appropriate for children's attention spans (Figure 24.5)?

Analysis of the FSM subtest (Figure 24.5) suggested removing 6 of the 25 items. After removing these items, no major degradation or change in the key item statistics for this assessment was observed. Cronbach's α remained high ($\alpha_{FSMallitems}$ = .90, CI95 = [.87 ; .93] & $\alpha_{FSMadjusted}$ = .89, CI95 = [.85 ; .92]), and the distributions of the proportion-correct values and the point-biserial correlations for all items remained similar. The correlation (Fig. 2, left bottom) between FSM total scores and CTOPP-2 stayed about the same ($r_{FSMallitems}$ = .69, $r_{FSMadjusted}$ = .67). Analysis of the LSM subtest (Fig 4., left bottom) also suggested removing 6 out of 25 items. Similar to FSM, Cronbach's α of LSM remained high ($\alpha_{LSMallitems}$ = .92, CI95 = [.90 ; .94] & $\alpha_{LSMadjusted}$ = .92, CI95 = [.90 ; .93]), the distributions of the proportion-correct values, the point-biserial correlations, and the correlation between the total scores and the CTOPP-2 remained similar ($r_{LSMallitems}$ = .70, $r_{LSMadjusted}$ = .70). Analysis of the DEL subtest (Figure 24.5) indicates removal of 5/24 items. Again, Cronbach's α remained high ($\alpha_{DELallitems}$ = .86, CI95 = [.79 ; .89] & $\alpha_{DELadjusted}$ = .85, CI95 = [.78 ; .89]), the distributions of the proportion-correct values, the point-biserial correlations, and the correlation between the total scores and the CTOPP-2 remained similar ($r_{DELallitems}$ = .65, $r_{DELadjusted}$ = .63).

	Subtest item	Criterium 1	2	3	4	Subtest item	Criterium 1	2	3	4	Subtest item	Criterium 1	2	3	4
1	FSM 2			X	X	LSM 3		X		X	DEL 2		X		X
2	FSM 3		X		X	LSM 9		X		X	DEL 12	X	X		X
3	FSM 9		X		X	LSM 12	X	X	X		DEL 19	X	X		
4	FSM 13	X	X			LSM 14				X	DEL 23	X	X		X
5	FSM 19		X	X	X	LSM 23	X	X			DEL 24	X	X		X
6	FSM 20	X	X		X	LSM 25			X	X					

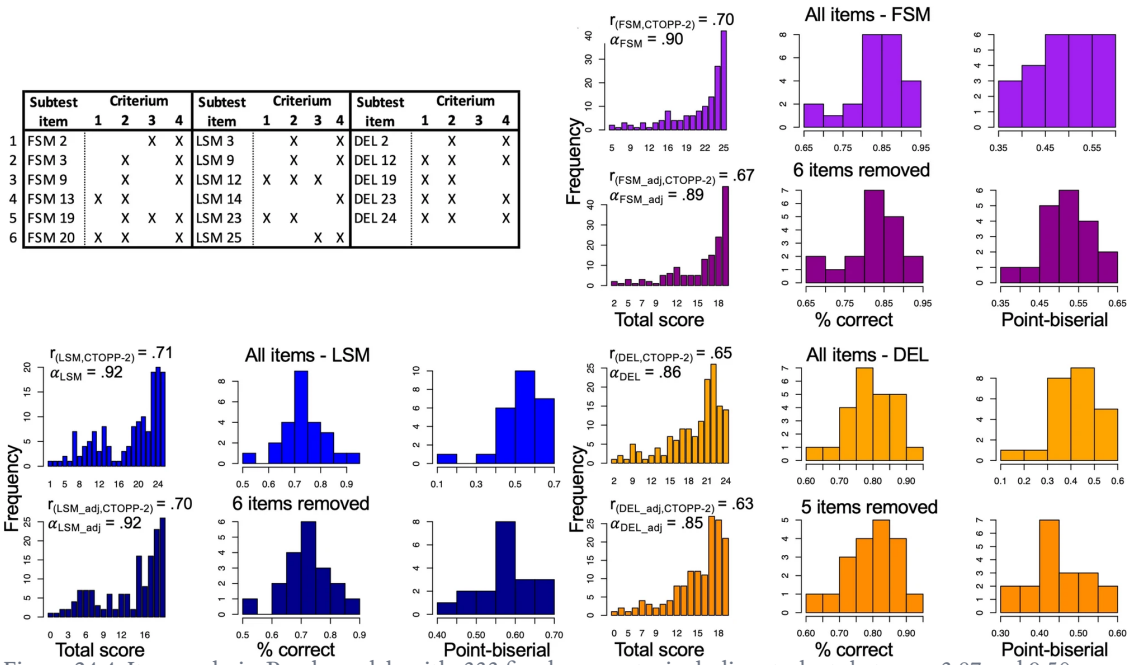


Figure 24.4: Item analysis: Rasch models with .333 fixed guess rate, including students between 3.87 and 9.50 years old, that were not excluded based on clicking or foil patterns for the three remaining subtests (FSM, LSM and DEL). **Left Top**: Deleted items based on a minimum of meeting 2/4 criteria. **Right top** & **Bottom**: Representation of correlation with CTOPP-2, Cronbach's α, Distributions of %-correct values and the point-biserial correlations for all items per subtest and for the subtests with items deleted based on the four criteria.

This Rasch item analysis suggests that every subtest of this ROAR-PA task has a good (DEL) to excellent (FSM, LSM) internal consistency, based on Cronbach's α, with a strong correlation

of every subtest (r > .65) to the overall CTOPP-2 scores. Item analysis based on meeting at least 2/4 suggested criteria, results in 19 items per subtest, and an overall task of 57 items + 2 practice items per subtest.

ROAR-Phoneme items were designed to span different theoretical levels of difficulty (e.g., Stanovich 2017; Treiman and Zukowski 1991). For the DEL subtest, difficulty levels were based on manipulation of (1) words and syllables (item 1-8), (2) onset and rimes (item 9-16), or (3) phonemes in the middle of the word (item 17-24). For FSM and LSM we can not follow these levels, as the task itself focuses on the first or last phoneme(s) of the word. We tried to create difficulty levels by manipulating single phonemes (level 1: item 1-16) or a single phoneme in a phoneme cluster (level 2: item 17-25). Surprisingly, based on the Rasch Model item-person maps for the three subtests (Fig. 5), we only found that the subtest DEL approximately follows the expected difficulty pattern. This analysis also showed that for FSM most items are closer to the lower-range of ability. For LSM and DEL, most items are close to the mid-range of ability.

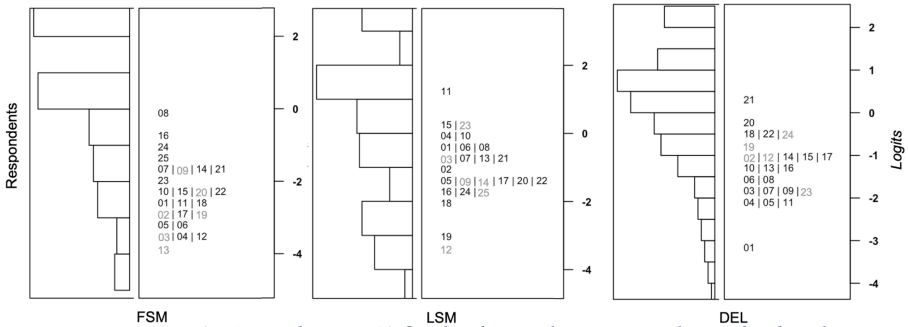


Figure 24.5: Item-person maps (or "Wright Maps") for the three subtests. Per subtest, the distribution of the ability of the students is plotted on the left-hand side; higher ability is closer to the top of the map. The distribution of the item difficulty is plotted on the right-hand side. The deleted items from the item analysis are grayed out.

24.2 Correlations between ROAR-Phoneme and ROAR-Word

PA is assessed at the beginning of reading instruction because of the relationship between PA and decoding skills. Students who struggle with PA tend to also struggle to learn decoding skills. There have been hundreds of studies documenting the relationship between PA and reading skills early in elementary school and the expected correlation ranges from r=0.3 to r=0.5 depending on the details of the measure and the sample (Scarborough 1998; Swanson et al. 2003). Table 24.1 shows the correlation between ROAR-Phoneme and ROAR-Word for kindergarten through 12th grade. The correlation is right in the expected range providing additional evidence for the validity of ROAR-Phoneme as a measure of PA skills.

Table 24.1

Grade	Correlation between ROAR-Phoneme and ROAR-Word	N
Kindergarten	0.44	254
1	0.53	3369

2	0.50	2014
3	0.41	956
4	0.42	588
5	0.38	400
6	0.19	1342
7	0.27	927
8	0.23	1442
9	0.32	1948
10	0.25	1756
11	0.33	1325
12	0.42	1053

Part VI

Construct Validity: ROAR-Español

References

Allen, Mary J, and Wendy M Yen. 2001. *Introduction to Measurement Theory*. Waveland Press.

Anthony, Jason L, and Christopher J Lonigan. 2004. "The Nature of Phonological Awareness: Converging Evidence from Four Studies of Preschool and Early Grade School Children." *Journal of Educational Psychology* 96 (1): 43.

Scarborough, Hollis S. 1998. "Predicting the Future Achievement of Second Graders with Reading Disabilities: Contributions of Phonemic Awareness, Verbal Memory, Rapid Naming, and IQ." *Annals of Dyslexia* 48: 115–36.

Stanovich, Keith E. 2017. "Speculations on the Causes and Consequences of Individual Differences in Early Reading Acquisition." In *Reading Acquisition*, 1st Edition, 307–42. Routledge.

Swanson, H Lee, Guy Trainin, Denise M Necoechea, and Donald D Hammill. 2003. "Rapid Naming, Phonological Awareness, and Reading: A Meta-Analysis of the Correlation Evidence." *Review of Educational Research* 73 (4): 407–40.

Tabachnick, BG, LS Fidell, BG Tabachnick, and LS Fidell. 2012. "Chapter 13 Principal Components and Factor Analysis." *Using Multivariate Statistics* 6: 612–80.

Treiman, Rebecca, and Andrea Zukowski. 1991. "Levels of Phonological Awareness." *Phonological Processes in Literacy: A Tribute to Isabelle Y. Liberman*.

Wagner, R K, J K Torgesen, and C A Rashotte. 1999. *Comprehensive test of phonological processes (CTOPP)*. Austin, TX: Pro-Ed.

Wright, Benjamin D. 1994. "Reasonable Mean-Square Fit Values." *Rasch Meas Transac* 8: 370.

25 Spanish Single Word Recognition (ROAR-Palabra) Concurrent Validity

ROAR-Palabra is designed to measure the latent construct of single word reading. Analogous to the concurrent validity analyses for ROAR-Word, we first establish that the silent, lexical decision task in ROAR-Palabra taps into the same latent construct by comparing ROAR-Palabra scores to a variety of other standardized measures of single word reading (see Section 22.1).

Figure 25.1 shows the correlation between ROAR-Palabra raw scores (θ) and a composite of Woodcock Muños (WM) Letter-Word Identification (real word reading) and Word Attack (pseudoword reading) raw scores. The correlation between ROAR-Palabra and WM is not as strong as the correlation between ROAR-Word and WJ (Figure 22.1; Figure 22.2; Figure 22.4) but this likely has to do with ceiling and floor effects in the sample: in transparent orthographies like spanish there is substantial less variation in single word reading skills. Figure 25.2 and Figure 25.3 show correlations for WM subtests separately.

25.1 Convergent validity with oral measures of single word reading

25.1.1 Woodock Muñoz

25.1.2 Growth Over Time

Another source of validation is examining growth trajectories of ROAR-Palabra scores over time. Figure 25.4 shows how ROAR-Palabra score steadily increase in each grade.

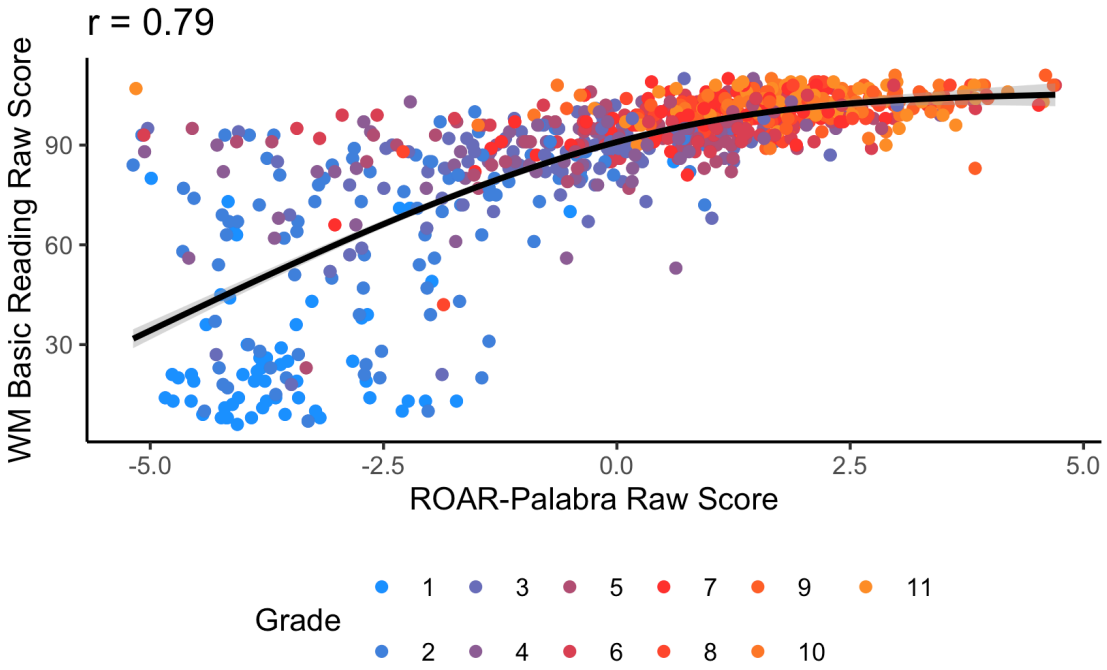


Figure 25.1: Correlations between ROAR-Palabra raw scores (theta) and Woodcock-Muñoz Basic Reading Skills raw scores.

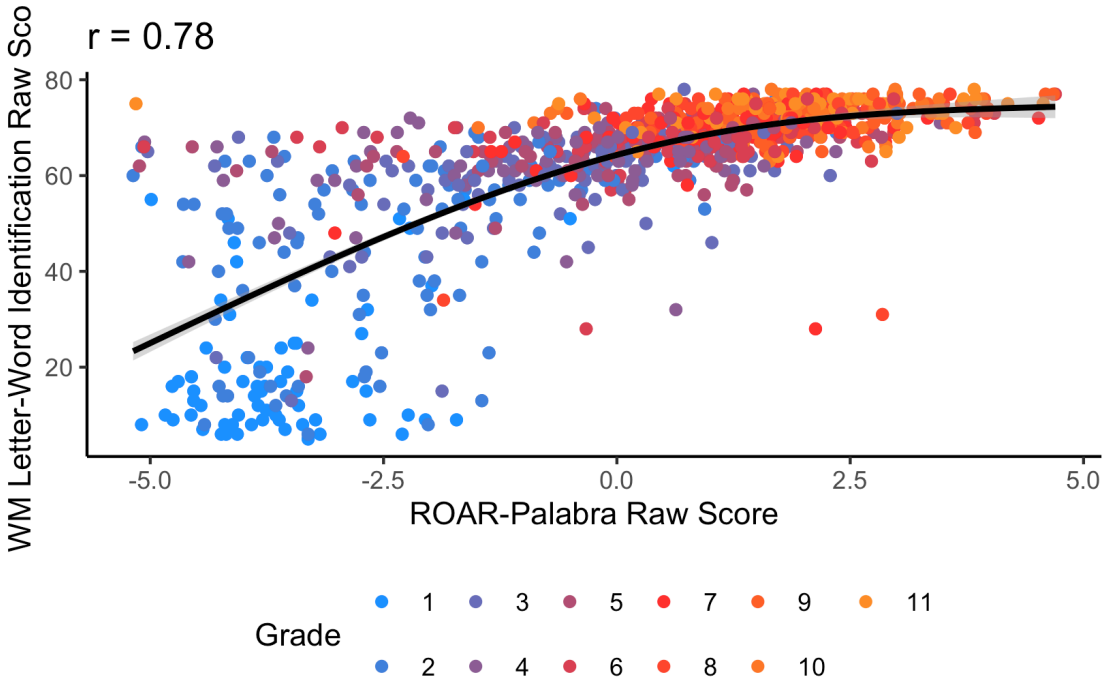


Figure 25.2: Correlations between ROAR-Palabra raw scores and Woodcock-Muñoz Letter-word Identification raw scores.

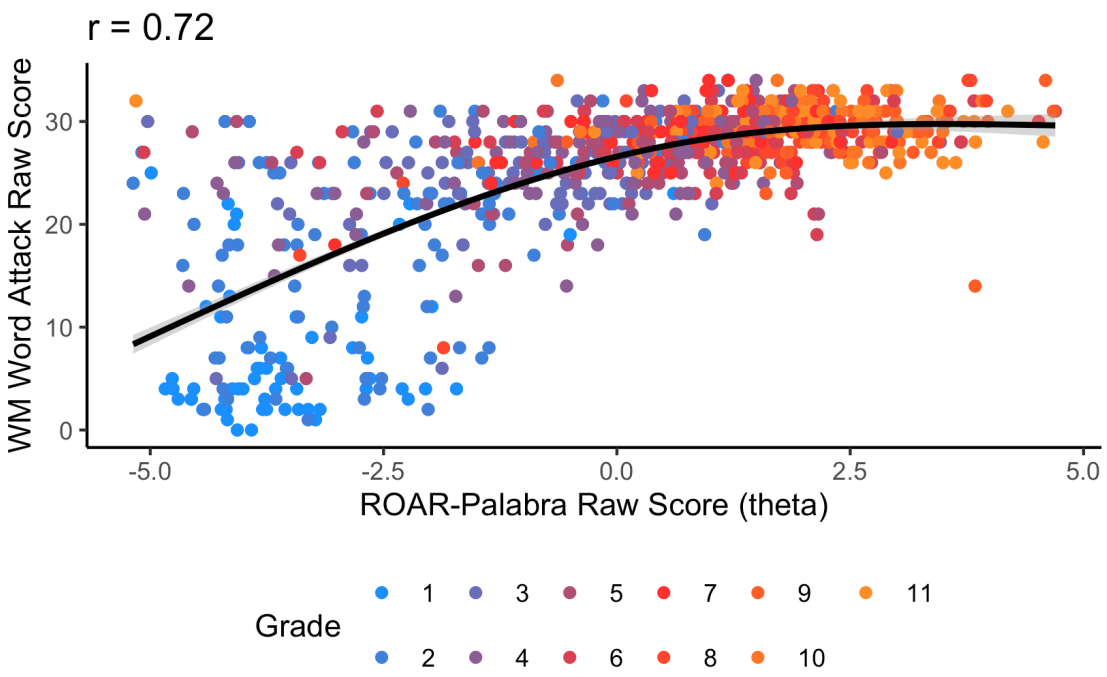


Figure 25.3: Correlations between ROAR-Palabra Raw Scores and Woodcock-Muñoz Word Attack raw scores.

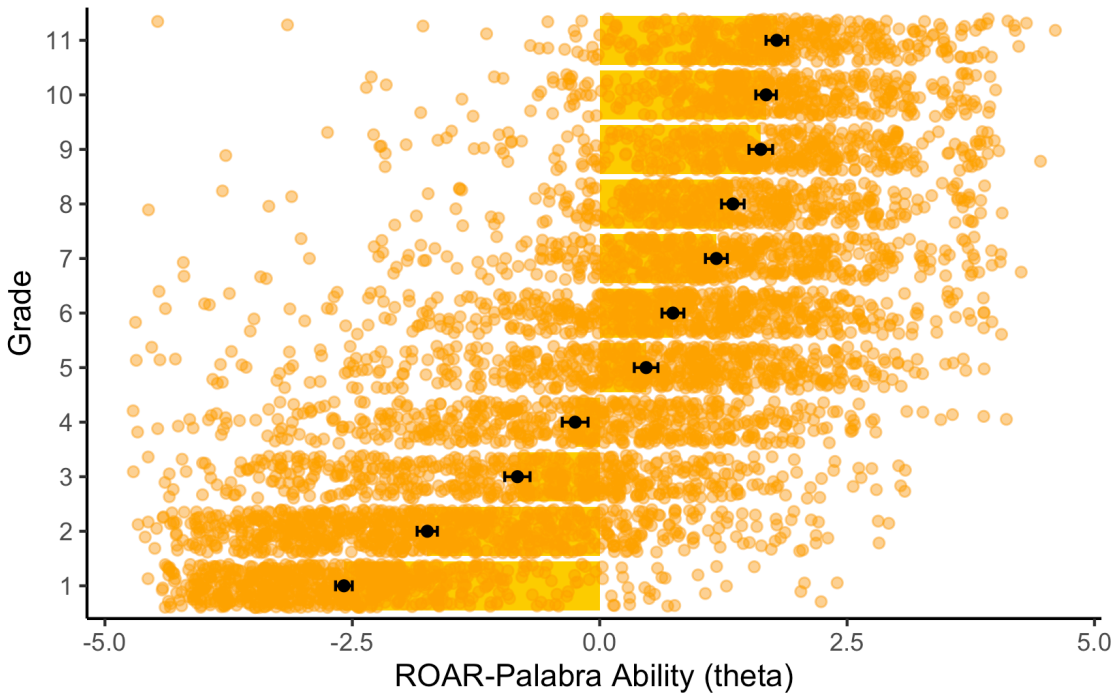


Figure 25.4: ROAR-Palabra ability estimates (theta) by grade

26 ROAR-Frase Concurrent Validity

26.1 Convergent validity with [Insert name of WM measure]

Part VII

Criterion Validity: Evidence for ROAR as a dyslexia screener

27 Validity: Dyslexia Screening and Sub-typing

From the perspective of neuroscience, written language is an incredible feat. Prompted by reading instruction, the brain constructs specialized circuits to translate visual symbols into their sounds and meanings (Yeatman 2022; Yeatman and White 2021). The brain has evolved dedicated circuits for spoken language and visual recognition processes because these skills have been integral to survival for eons. Written language, however, was invented by human societies only a few thousand years ago. It is unlikely that the brain evolved dedicated circuits for written language. The brain develops the circuitry for literacy through experiences with written language beginning in infancy thanks to the brain's capacity to change in response to new experiences, a principle known as "plasticity". This means that a child's experiences in the classroom sculpt their neural circuitry of literacy.

However, this circuitry is not built from scratch. Literacy is grounded in circuits that evolved for component processes, such as spoken language and visual recognition. As a child begins to learn to read, brain circuits that evolved for visual recognition are reorganized to process text and route this information to the brain's spoken language network. This process depends on instruction and practice. But for some children, the process of learning to read presents a substantial struggle. For children with **Developmental Dyslexia**, struggles with foundational reading skills—decoding, word recognition and reading speed/efficiency specifically—tend to persist throughout schooling unless they receive additional support and/or evidence-based intervention. The goal of a dyslexia screener is to identify students who would benefit from additional support in foundational reading skills. The promise of plasticity is that once they are identified and provided with intensive, targeted, systematic support in foundational reading skills, children with Developmental Dyslexia can develop the ability to decode and read efficiently.

There are a variety of definitions of dyslexia, but they all share this characteristic: persistent struggles with decoding, word recognition and establishing fluent reading.

International Dyslexia Association (IDA) Definition of Dyslexia

The International Dyslexia Association published one of the most widely used definitions of dyslexia which was developed through a consensus building process[a] in partnership with the National Center for Learning Disabilities (NCLD)[b], and the National Institute of Child Health and Human Development (NICHD)[c] (Lyon, Shaywitz, and Shaywitz 2003).

From the IDA website[d]: "Dyslexia is a specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities.

These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge."

[a]https://dyslexiaida.org/definition-consensus-project/
[b]https://www.ncld.org/
[c]https://www.nichd.nih.gov/
[d]https://dyslexiaida.org/definition-of-dyslexia/

While this definition has been widely used for the past 2 decades, there has been a recent push to revise the defintion of dyslexia to a) make diagnosis simpler, b) make it easier to get support to the students that need it and c) acknowledge heterogeneity and the multifactorial nature of dyslexia.

Proposed revisions to the definition of dyslexia

Catts et al. (2024) argue for an alternative, "prevention-based approach that focuses on the early identification of children at risk for dyslexia and the provision of instruction/intervention that is matched to their needs." Catts et al. (2024) specifically propose revisions to the definition that incorporate other, known causal factors beyond phonological awareness.

Snowling and Hulme (2024), on the other hand, argue that a revised definition is unnecessary and propose that causal arguments need not go into the definition.

Finally, Elliott and Grigorenko (2024) propose a "simpler definition that describes the primary difficulty, avoids reference to causal explanation, unexpectedness, and secondary outcomes, and redirects practitioner and policymaker focus to the importance of addressing and meeting the needs of all struggling readers."

The proposed revision of Elliott and Grigorenko (2024) only references challenges with word reading accuracy and speed, making Dyslexia more straightforward to diagnose and intervene.

27.1 Dyslexia screening based on foundational reading skills: Criterion validity

To assess sensitivity and specificity of **ROAR Foundational Reading Skills** (see Section 9.1) as an indicator of dyslexia risk, we ran two studies of criterion validity—one with a reading assessment that is among the most commonly used in schools, and one with the most widely-used measure in dyslexia research:

Criterion validity

1. A study in collaboration with two, large and diverse California school districts that uses FAST™ earlyReading and FAST™ CBMreading[1] risk categories as the criterion measures. FAST™ earlyReading and FAST™ CBMreading are individually administered

[1]https://support-content.fastbridge.org/FAST_Research/FAST_Technical_Manual_Version_FINAL.pdf

screeners that classify students into three different risk levels for reading difficulties: "Low Risk", "Some Risk", and "High Risk". For kindergarten we calculate prediction accuracy, sensitivity and specificity of ROAR Foundational Reading Skills relative to FAST™ earlyReading. For first grade, we calculate prediction accuracy, sensitivity and specificity of ROAR Foundational Reading Skills relative to FAST™ earlyReading and FAST™ CBMreading. For second grade we calculate prediction accuracy, sensitivity and specificity of ROAR Foundational Reading Skills relative to FAST™ CBMreading.

2. A study with participants recruited from around the United States that uses the Woodcock Johnson Basic Reading Skills Composite Index (WJ BRS) as the criterion measure. WJ BRS is the most widely used measure in dyslexia research for identifying characteristics of dyslexia and is one of the most widely used measures in special education and clinical practice for diagnosing dyslexia. For this study of criterion validity, we use a threshold of the 25th percentile based on national norms to define students at risk or with indications of dyslexia and we calculate prediction accuracy, sensitivity and specificity of ROAR Foundational Reading Skills relative to this criterion.

27.1.1 Criterion Validity Study 1: FastBridge

27.1.1.1 Sample demographics

This study was carried out in collaboration with two California school districts. Demographics of the sample are provided in Table ??.

Table 27.1 and Table 27.2 show the distribution of students in the sample across FAST™ earlyReading and FAST™ CBMreading risk categories. Note that FAST™ CBMreading categories of "College Pathway" and "Exceeding Expectations" have been included in the category "Low Risk" for the sake of this analysis.

Table 27.1: Distributions of FAST™ earlyReading risk categories

Grade	Early Reading Risk Level	N	Proportion of Risk Level
Kindergarten	High Risk	36	35.6%
Kindergarten	Some Risk	22	21.8%
Kindergarten	Low Risk	43	42.6%
1	High Risk	222	26%
1	Some Risk	177	20.8%
1	Low Risk	454	53.2%

Table 27.2: Distributions of FAST™ CBMreading risk categories

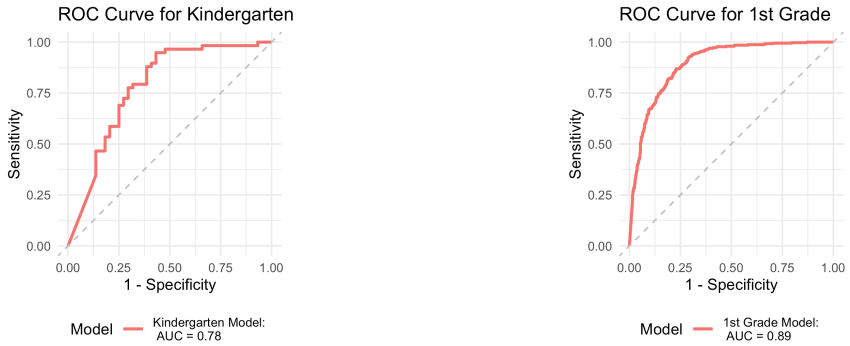
Grade	CBMreading Risk Level	N	Proportion of Grade
1	High Risk	201	22.5%
1	Some Risk	163	18.3%

1	Low Risk	528	59.2%
2	High Risk	187	19.3%
2	Some Risk	151	15.6%
2	Low Risk	633	65.2%

27.1.1.2 ROAR-Word

Since dyslexia is identified based on persistent difficulties with word reading accuracy and fluency, word reading measures are generally the most efficient screeners though additional measures of Letter Sound Knowledge, Phonological Awareness, Rapid Automatized Naming and Visual Processing can also improve sensitivity/specificity, particularly for younger students at the early stages of learning to read. Thus, we begin by computing prediction accuracy, sensitivity and specificity for ROAR-Word. We then examine whether additional measures lead to more accurate predictions. Finally, we examine each additional measure in isolation.

Figure 27.1 shows an ROC curve for kindergarten and 1st grade computed from a logistic regression model with ROAR-Word as a predictor of the FAST™ earlyReading "High Risk" category. Figure 27.2 shows and ROC curve for 1st and 2nd grades computed from a logistic regression model with ROAR-Word as a predictor of the FAST™ CBMreading "High Risk" category. All models in 1st and 2nd grade achieved exceptional accuracy with area under the curve (AUC) greater than 0.9 for both criterion measures. In kindergarten accuracy was lower, which is expected for a model that does not include other screening measures. Table 27.3 and Table 27.4 report sensitivity, specificity and accuracy by each demographic. Table 27.5 and Table 27.6 report sensitivity, specificity and accuracy by each demographic.



(a) Kindergarten prediction of FAST™ earlyReading (b) 1st grade prediction of FAST™ earlyReading

Figure 27.1: Prediction of FAST™ earlyReading risk categories based on a logistic regression model with ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds.

Table 27.3: Kindergarten area under the curve, best sensitivity and specificity, and specificity when sensitivity is held closest to 0.9 for FastBridge Early Reading.

Demographic Group	Grade	AUC	Best Specificity	Best Sensitivity	Specificity (Sensitivity at 0.9)
English Learner	Kindergarten	0.7142857	0.7500000	0.7619048	0.2500000
Female	Kindergarten	0.7701754	0.7894737	0.7666667	0.5263158

Male	Kindergarten	0.7693452	0.5833333	0.9285714	0.5416667
White	Kindergarten	0.7619048	0.5925926	0.9523810	0.5925926
Hispanic Ethnicity	Kindergarten	0.8072917	0.6666667	0.9687500	0.6666667
Free or Reduced Lunch	Kindergarten	1.0000000	1.0000000	1.0000000	1.0000000
All	Kindergarten	0.7825235	0.5681818	0.9482759	0.5681818

Table 27.4: 1st grade area under the curve, best sensitivity and specificity, and specificity when sensitivity is held closest to 0.9 for FastBridge Early Reading.

Demographic Group	Grade	AUC	Best Specificity	Best Sensitivity	Specificity (Sensitivity at 0.9)	Se
English Learner	1	0.8665179	0.7500000	0.8785714	0.5714286	
Female	1	0.9024567	0.8090452	0.8722222	0.7487437	
Male	1	0.8899066	0.6803069	0.9348837	0.6930946	
White	1	0.8441987	0.6636364	0.9279279	0.6636364	
Hispanic Ethnicity	1	0.8455505	0.7750000	0.7963801	0.5833333	
Black or African American	1	0.8194444	0.6666667	1.0000000	NA	
Multiracial	1	0.8390533	0.7769231	0.8461538	0.5153846	
SPED	1	0.9466403	1.0000000	0.8260870	0.7272727	
Free or Reduced Lunch	1	0.8842728	0.9041096	0.7213115	0.6849315	
All	1	0.8948850	0.7698113	0.8696742	0.7207547	

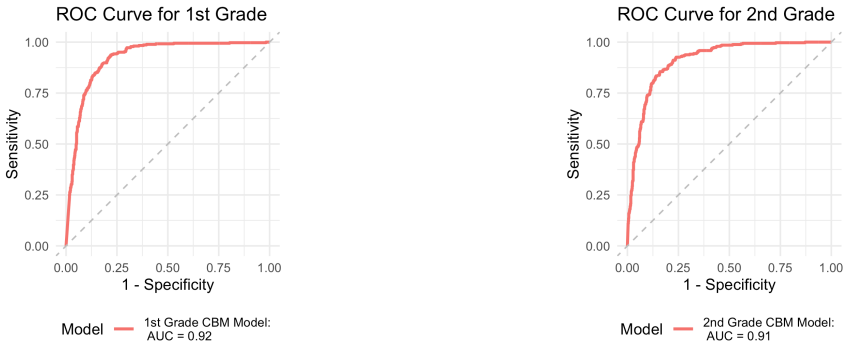
Table 27.5: 1st grade area under the curve, best sensitivity and specificity, and specificity when sensitivity is held closest to 0.9 for FastBridge CBM Reading.

Demographic Group	Grade	AUC	Best Specificity	Best Sensitivity	Specificity (Sensitivity at 0.9)	Sensit
English Learner	1	0.8613520	0.7321429	0.8785714	0.6428571	
Female	1	0.9245473	0.8096386	0.9447853	0.8216867	
Male	1	0.9207303	0.8166259	0.8934010	0.7750611	
White	1	0.8937006	0.7500000	0.9381443	0.7587209	
Hispanic Ethnicity	1	0.8608712	0.7539683	0.8651163	0.6746032	
Asian	1	0.9237395	0.7285714	1.0000000	0.7285714	
Multiracial	1	0.8986711	0.7984496	1.0000000	0.7984496	
SPED	1	0.9146341	0.9375000	0.9024390	0.8125000	
Free or Reduced Lunch	1	0.8880505	0.7532468	0.8826816	0.7012987	
All	1	0.9212068	0.7807229	0.9368132	0.8000000	

Table 27.6: 2nd grade area under the curve, best sensitivity and specificity, and specificity when sensitivity is held closest to 0.9 for FastBridge CBM Reading.

Demographic Group	Grade	AUC	Best Specificity	Best Sensitivity	Specificity (Sensitivity at 0.9)	Sensit
English Learner	2	0.8081314	0.7297297	0.8103448	0.5270270	
Female	2	0.9291882	0.7743363	0.9652778	0.8119469	
Male	2	0.8991029	0.8659091	0.8000000	0.7181818	

White	2	0.8852896	0.7398374	1.0000000	0.7926829
Hispanic Ethnicity	2	0.8916598	0.8717949	0.8082192	0.6752137
Asian	2	0.8841069	0.8354430	0.8666667	0.5443038
Multiracial	2	0.8273299	0.7735849	0.9090909	0.3836478
SPED	2	0.9068323	0.7826087	0.9047619	0.6521739
Free or Reduced Lunch	2	0.9117904	0.8953488	0.8225806	0.7558140
All	2	0.9116782	0.8400000	0.8550296	0.7777778



(a) 1st grade prediction of FAST™ CBMreading (b) 2nd grade prediction of FAST™ CBMreading

Figure 27.2: Prediction of FAST™ CBMreading risk categories based on a logistic regression model with ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds.

27.1.1.3 ROAR Foundational Reading Skills Composite

We next examine model accuracy based on a logistic regression model with all three ROAR measures of foundational reading skills: ROAR-Phoneme, ROAR-Letter and ROAR-Word. Because model accuracy was already near perfect for 1st and 2nd grade we would not expect a large improvement. However in kindergarten, when foundational reading skills are still being established, we expect measures of Phonological Awareness and Letter Sound knowledge to improve prediction accuracy. Figure 27.3 shows an ROC curve for the full model with all the ROAR measures (Phoneme, Letter, and Word) compared to models with each individual measure in kindergarten. ROAR-Letter and ROAR-Phoneme both achieved exceptional accuracy and the full model performed marginally better. Figure 27.4 shows ROC curves for the four models in 1st grade. In 1st grade ROAR-Word is the best single predictor and the full model (ROAR-Letter, ROAR-Phoneme, and ROAR-Word) performs marginally better.

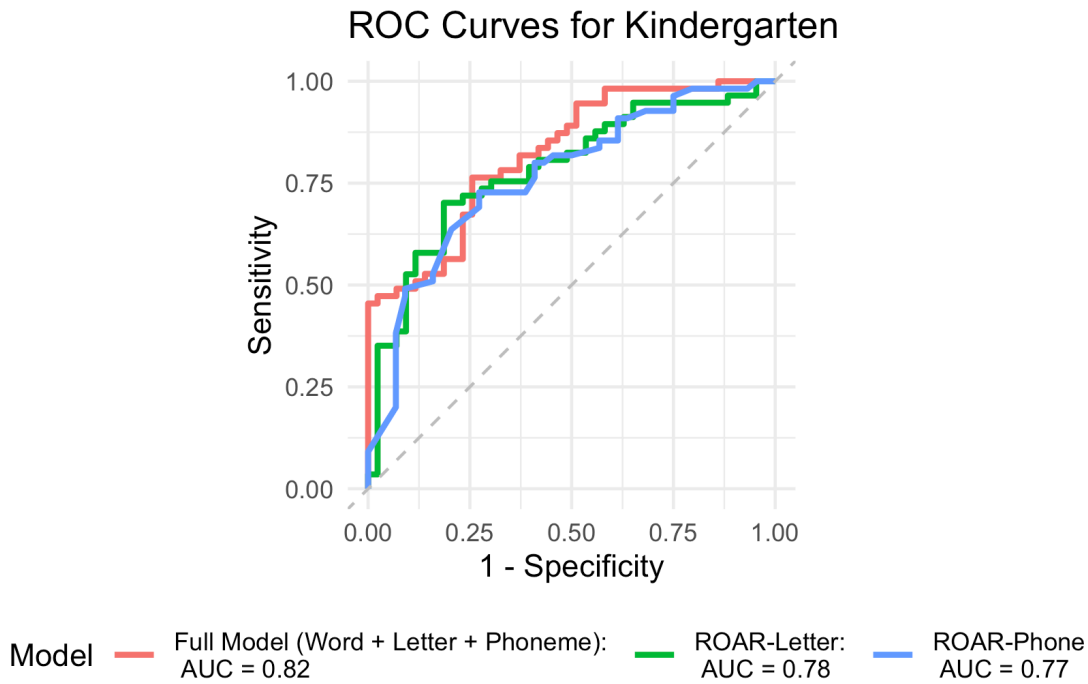


Figure 27.3: Prediction of FAST™ earlyReading risk categories in kindergarten based on a logistic regression model with ROAR-Letter, ROAR-Phoneme, and ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds. Full Model refers to the logistic regression with all three predictors and models of individual ROAR measures are shown for comparison.

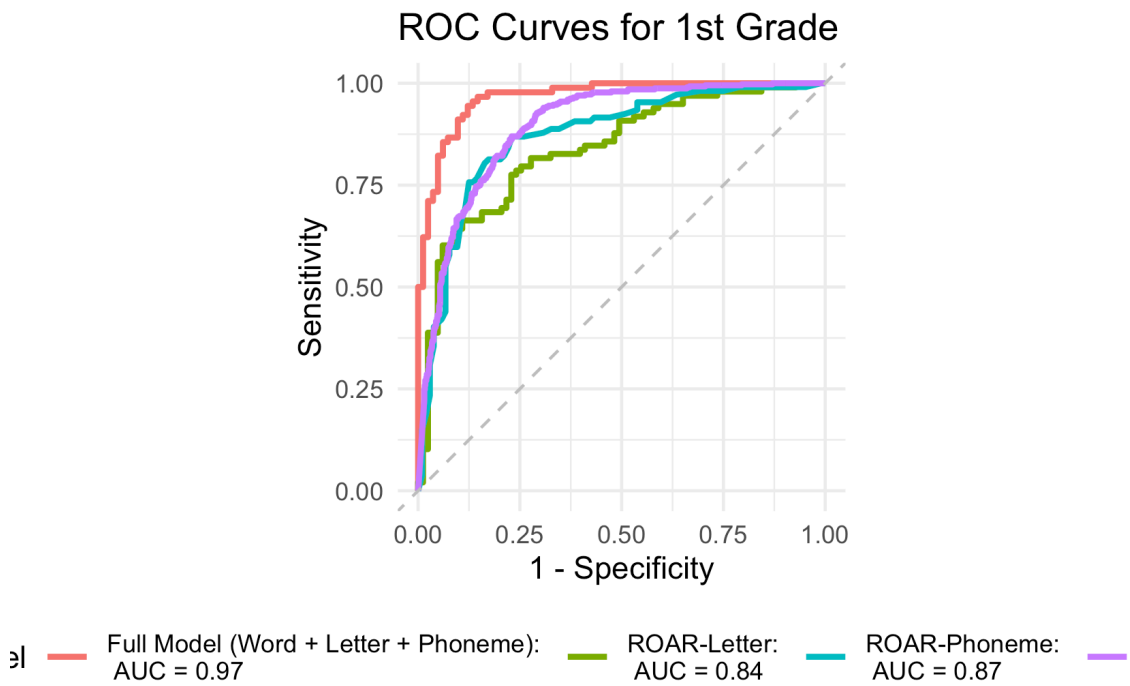


Figure 27.4: Prediction of FAST™ earlyReading risk categories in 1st grade based on a logistic regression model with ROAR-Letter, ROAR-Phoneme, and ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds. Full Model refers to the logistic regression with all three predictors and models of individual ROAR measures are shown for comparison.

27.1.2 Criterion Validity Study 2: Woodcock Johnson Basic Reading Skills (WJ BRS)

27.1.2.1 Sample demographics

This study included participants recruited from all around the United States for research studies in the Brain Development & Education Lab[2]. Figure 22.3 shows the age distribution and Table 22.1 shows the demographics of the students that participated in this validation study.

Table 27.7 shows the distribution of students in the sample across Woodcock Johnson Basic Reading Skills (BRS) risk categories. Note that the original risk categories for Woodcock Johnson BRS were not used, rather, we determined the three level risk categories. Low risk included students who were greater than the 50th percentile, some risk included students who were between the 25th and 50th percentiles, and high risk included students who were below the 25th percentile.

Table 27.7: Distributions of Woodcock Johnson risk categories

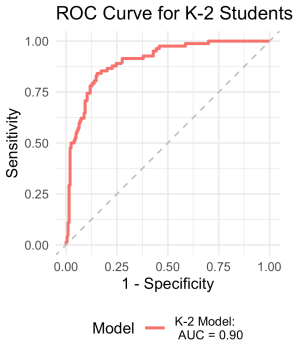
Age Range	WJ Reading Risk	N	Proportion of Risk Level
K-2	Low Risk	203	71.2%
K-2	Some Risk	45	15.8%
K-2	High Risk	37	13%
3rd-5th	Low Risk	76	40.6%
3rd-5th	Some Risk	50	26.7%
3rd-5th	High Risk	61	32.6%
6th-8th	Low Risk	19	35.2%
6th-8th	Some Risk	12	22.2%
6th-8th	High Risk	23	42.6%
9th-12th	Low Risk	22	32.4%
9th-12th	Some Risk	22	32.4%
9th-12th	High Risk	24	35.3%

27.1.2.2 ROAR-Word

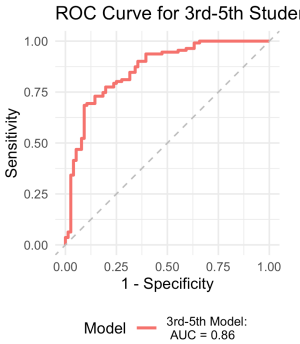
Figure 27.5 shows an ROC curve for all grades (grouped by K-2, 3-5, 6-8, 9-12) computed from a logistic regression model with ROAR-Word as a predictor of the Woodcock Johnson Basic Reading Skills "Some Risk" category. Figure 27.6 shows an ROC curve for all grades (grouped by K-2, 3-5, 6-8, 9-12) computed from a logistic regression model with ROAR-Word as a predictor of the Woodcock Johnson Basic Reading Skills "High Risk" category. The model of grades K-2 achieved exceptional accuracy with area under the curve (AUC) equal to or greater than 0.9 for both criterion measures. For older grades accuracy was lower, and this reflects the psychometric properties of the criterion measure in older students. Most middle

[2] https://www.brainandeducation.com/

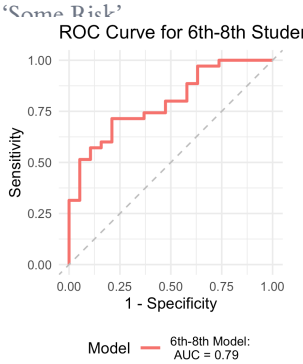
school and high school students are at the ceiling of the Woodcock Johnson Basic Reading Skills index (for example see Table 22.3 which shows the decline in reliability of WJ in older grades).



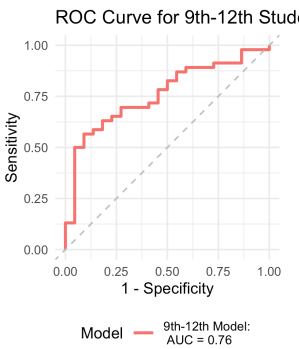
(a) Kindergarten and 1st Grade prediction of Woodcock Johnson BRS 'Some Risk'



(b) 3rd-5th Grade prediction of Woodcock Johnson BRS 'Some Risk'

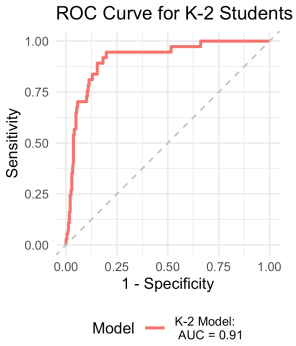


(c) 6th-8th Grade prediction of Woodcock Johnson BRS 'Some Risk'

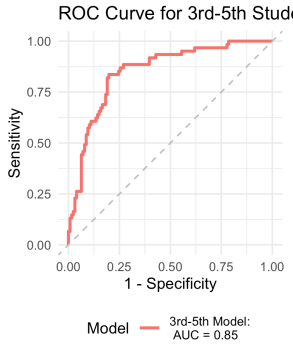


(d) 9th-12th Grade prediction of Woodcock Johnson BRS 'Some Risk'

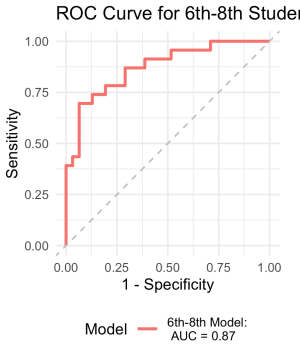
Figure 27.5: Prediction of Woodcock Johnson Basic Reading Skills 'Some Risk' category based on a logistic regression model with ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds.



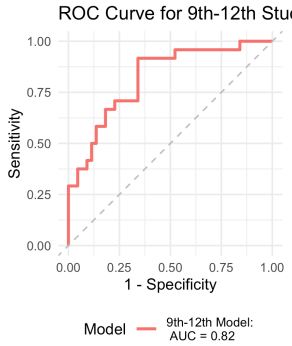
(a) Kindergarten and 1st Grade prediction of Woodcock Johnson BRS 'High Risk'



(b) 3rd–5th Grade prediction of Woodcock Johnson BRS 'High Risk'



(c) 6th–8th Grade prediction of Woodcock Johnson BRS 'High Risk'



(d) 9th–12th Grade prediction of Woodcock Johnson BRS 'High Risk'

Figure 27.6: Prediction of Woodcock Johnson Basic Reading Skills 'High Risk' category based on a logistic regression model with ROAR-Word. Receiver Operating Characteristic (ROC) curves display sensitivity and specificity at different thresholds.

Part VIII

Predictive Validity: Longitudinal evidence that ROAR predicts future reading development and dyslexia risk

References

Catts, Hugh W, Nicole Patton Terry, Christopher J Lonigan, Donald L Compton, Richard K Wagner, Laura M Steacy, Kelly Farquharson, and Yaacov Petscher. 2024. "Revisiting the Definition of Dyslexia." *Ann. Dyslexia*, January.

Elliott, Julian G, and Elena L Grigorenko. 2024. "Dyslexia in the Twenty-First Century: A Commentary on the IDA Definition of Dyslexia." *Ann. Dyslexia*, June.

Lyon, G Reid, Sally E Shaywitz, and Bennett A Shaywitz. 2003. "A Definition of Dyslexia." *Annals of Dyslexia* 53: 1–14.

Snowling, Maggie, and Charles Hulme. 2024. "Do We Really Need a New Definition of Dyslexia? A Commentary." *Ann. Dyslexia*, March.

Yeatman, Jason D. 2022. "The Neurobiology of Literacy." *The Science of Reading: A Handbook*, 533–55.

Yeatman, Jason D, and Alex L White. 2021. "Reading: The Confluence of Vision and Language." *Annual Review of Vision Science* 7 (1): 487–517.

28 Predictive Validity

28.1 Background: Published studies

Predictive validity of ROAR Foundational Reading Skills (see Section 9.1 for additional information on ROAR Foundational Reading Skills) was first reported by (Gijbels et al. 2024). Gijbels et al. (2024) examined the classification accuracy of ROAR Foundational Reading Skills administered in 1st grade for classifying students who were deemed "at risk" for reading difficulties based on the Fountas and Pinnell (F&P) Benchmark Assessment 8 months later in the fall of 2nd grade. This study included N=130 1st grade students from a public school in California. Students completed ROAR Foundational Reading Skills measures in their classroom and F&P Benchmark Assessments were administered by their classroom teachers. A Generalized Additive Model (GAM) (S. Wood and Wood 2015; S. N. Wood 2017) based on ROAR-Phoneme achieved an AUC=0.70, ROAR-Word achieved and AUC=0.83, and a GAM with ROAR-Phoneme and ROAR-Word achieved an AUC=0.84. The prediction accuracy of ROAR-Phoneme and ROAR-Word for reading skills assessed the following school year with individually-admininstered assessments demonstrated the promise of ROAR as a quick and automated screener.

28.2 Longitudinal studies of grades 1-3

We ran 2 additional studies to assess the predictive validity of ROAR Foundational Reading Skills

1. **2 year longitudinal study of predictive validity**: In a large California school district, all the 1st grade classrooms were administered ROAR Foundational Reading Skills measures three times per year and were followed longitudinally for 2 years. In the fall of 3rd grade, each student was individually administered the Woodcock Johnson Basic Reading Skills (WJ BRS) Composite Index. Based on this criterion measure, we assessed sensitivity and specificity of ROAR at each timepoint for predicting students who were classified as struggling readers with indications of dyslexia. Additionally we report prediction accuracy based on BRS as a continuous measure.
2. **Fall to spring prediction in 1st, 2nd, and 3rd grade**: In a second study we assessed predictive validity of ROAR Foundational Reading Skills measures admininstered in the Fall and Winter for predicting individually administered FAST™ earlyReading and FAST™ CBMreading in the Spring (for concurrent validity of ROAR Spring assessment see Chapter 27).

For each study we report Area Under the Curve (AUC), Sensitivity, and Specificity as measures of classification accuracy and Pearson's ρ as a measure of prediction accuracy for continuous criterion measures.

28.2.1 Study 1: 2 year longitudinal study with Woodcock Johnsons Basic Reading Skills (BRS) as the criterion

We implemented our ROAR measures beginning in the first grade. As shown in Table 28.1, ROAR-Word admininstered in first grade consistently predicts third-grade WJ-BRS outcomes.The correlation between ROAR-Word and WJ-BRS strengthens over time. Table 28.2 demonstrates that ROAR measures can predict reading fluency as early as the first grade, with ROAR-Sentence being the most relevant predictor.

Table 28.1: Predictive validity between ROAR measures and WJ BRS

ROAR Measure	ROAR Administration	N	Correlation
ROAR Word	Fall 2021	120	0.685
ROAR Word	Spring 2022	120	0.632
ROAR Word	Fall 2022	125	0.629
ROAR Word	Spring 2023	141	0.738
ROAR Word	Fall 2023	165	0.744
ROAR Phoneme	Spring 2022	74	0.408
ROAR Phoneme	Fall 2022	117	0.539
ROAR Phoneme	Spring 2023	133	0.557
ROAR Phoneme	Fall 2023	166	0.526
ROAR Sentence	Spring 2023	125	0.715
ROAR Sentence	Fall 2023	164	0.697

Table 28.2: Predictive validity between ROAR measures and WJ Sentence Fluency

ROAR Measure	ROAR Administration	N	Correlation
ROAR-Word	Fall 2021	120	0.703
ROAR-Word	Spring 2022	120	0.733
ROAR-Word	Fall 2022	125	0.672
ROAR-Word	Spring 2023	141	0.713
ROAR-Word	Fall 2023	165	0.737
ROAR-Phoneme	Spring 2022	74	0.336
ROAR-Phoneme	Fall 2022	117	0.486
ROAR-Phoneme	Spring 2023	133	0.574
ROAR-Phoneme	Fall 2023	166	0.513
ROAR-Sentence	Spring 2023	125	0.817

ROAR-Sentence	Fall 2023	164	0.831

Based on the WJ-BRS, 32 out of 170 students were identified as high-risk or at-risk struggling readers (scoring below the 50th percentile of the WJ-BRS norms). We treated this classification as the true score. Next, we examined the prediction accuracy of a logistic regression model using ROAR measures taken in the previous year. Figure 28.1 provides further evidence supporting the high sensitivity and specificity of ROAR-Word in predicting dyslexia classification with a lead time of two years.

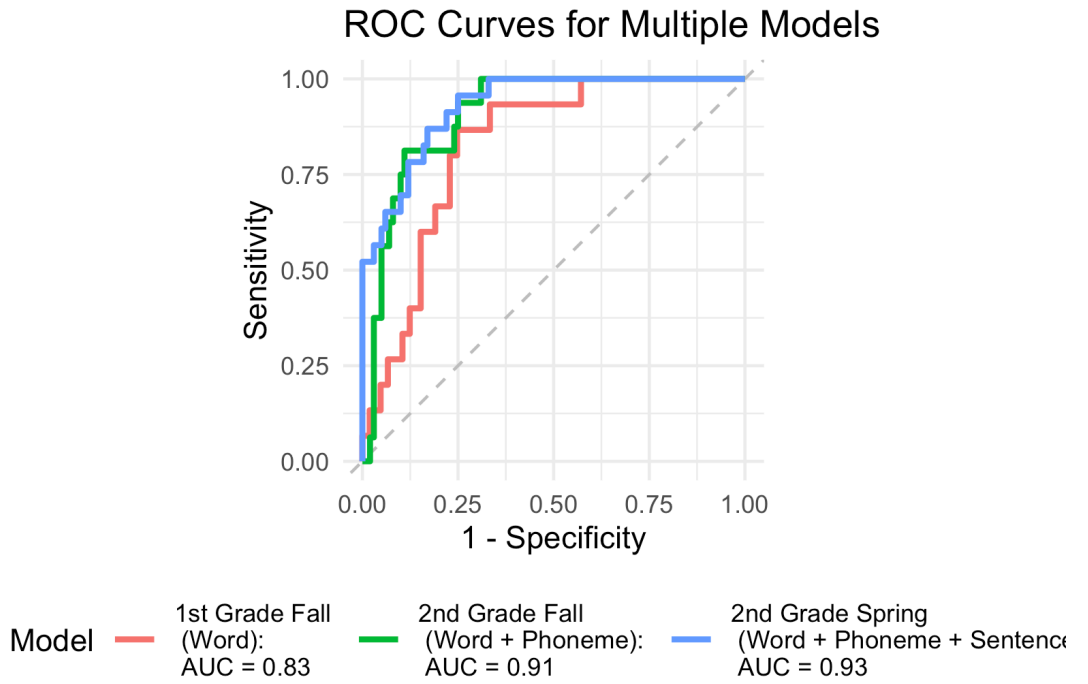


Figure 28.1: Prediction of Woodcock Johnsons Basic Reading Skills (BRS) risk categories based on a logistic regression model with ROAR measures in previous timepoints.

28.2.2 Study 2: Fall to Spring prediction of FAST™ earlyReading and FAST™ CBMreading

Table 28.3 demonstrates that ROAR-Word in the Fall, among ROAR measures, is the strongest predictor of FAST™ CBMreading performance in the Spring for 1st graders. For 2nd and 3rd graders, both ROAR-Word and ROAR-Sentence are strong predictors of FAST™ CBMreading outcomes. Additionally, Table 28.4 provides further evidence that ROAR-Word in the Fall is a robust predictor of FAST™ earlyReading performance in the Spring.

Table 28.3: Predictive validity between ROAR measures and FAST™ CBMreading

Grade	ROAR Measure	ROAR Administration	N	Correlation

1	ROAR-Word	Fall 2023	313	0.725
1	ROAR-Word	2024 Winter	336	0.782
1	ROAR-Word	Spring 2024	306	0.777
1	ROAR-Phoneme	Fall 2023	305	0.589
1	ROAR-Phoneme	2024 Winter	352	0.647
1	ROAR-Phoneme	Spring 2024	332	0.604
1	ROAR-Sentence	Fall 2023	263	0.647
1	ROAR-Sentence	2024 Winter	345	0.791
1	ROAR-Sentence	Spring 2024	307	0.796
2	ROAR-Word	Fall 2023	342	0.748
2	ROAR-Word	2024 Winter	338	0.705
2	ROAR-Word	Spring 2024	319	0.678
2	ROAR-Phoneme	Fall 2023	350	0.500
2	ROAR-Phoneme	2024 Winter	150	0.404
2	ROAR-Sentence	Fall 2023	333	0.765
2	ROAR-Sentence	2024 Winter	330	0.784
2	ROAR-Sentence	Spring 2024	322	0.780
3	ROAR-Word	Fall 2023	192	0.577
3	ROAR-Word	2024 Winter	163	0.583
3	ROAR-Word	Spring 2024	150	0.590
3	ROAR-Phoneme	Fall 2023	193	0.363
3	ROAR-Phoneme	2024 Winter	99	0.399
3	ROAR-Sentence	Fall 2023	190	0.587
3	ROAR-Sentence	2024 Winter	163	0.600
3	ROAR-Sentence	Spring 2024	149	0.594

Table 28.4: Predictive validity between ROAR measures and FAST™ earlyReading

Grade	ROAR Measure	ROAR Administration	N	Correlation
1	ROAR-Word	Fall 2023	313	0.725
1	ROAR-Word	2024 Winter	335	0.780
1	ROAR-Word	Spring 2024	306	0.777
1	ROAR-Phoneme	Fall 2023	305	0.589
1	ROAR-Phoneme	2024 Winter	351	0.645
1	ROAR-Phoneme	Spring 2024	331	0.601
1	ROAR-Sentence	Fall 2023	263	0.647
1	ROAR-Sentence	2024 Winter	345	0.791
1	ROAR-Sentence	Spring 2024	307	0.796

We examined the prediction accuracy of a logistic regression model using ROAR measures from Fall 2023 to predict the FAST™ classification (low risk vs. some risk and high risk) in

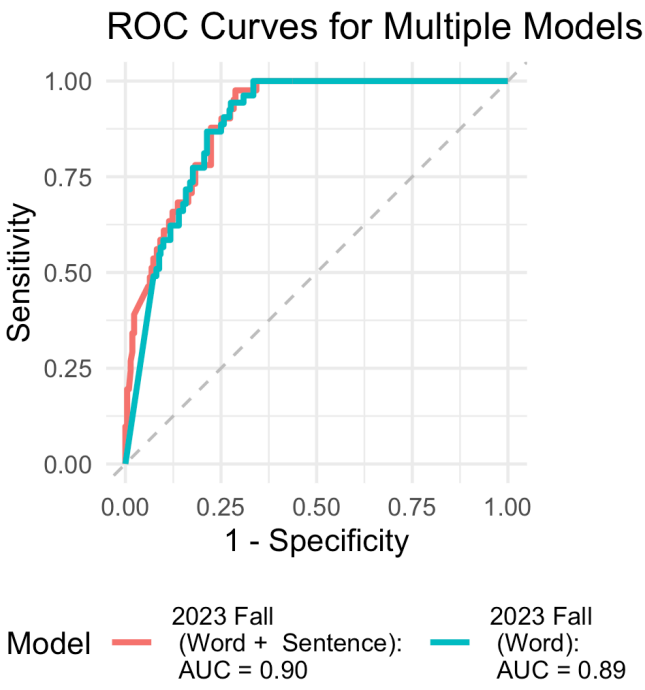
Spring 2024. Figure 28.2 provides evidence supporting the high sensitivity and specificity of ROAR-Word in predicting dyslexia classification in both 1st and 2nd grades. Additionally, ROAR-Phoneme is more useful in 1st grade and ROAR-Sentence proves to be more useful in 2nd grade.

References

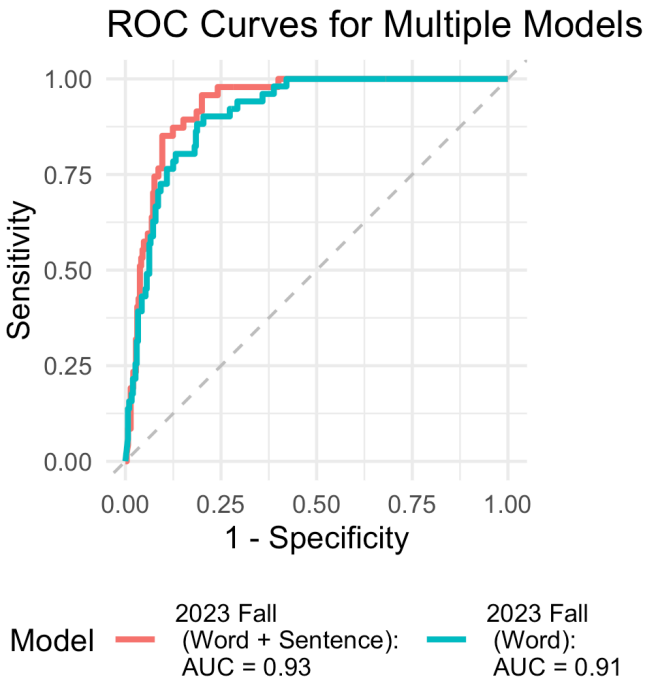
Gijbels, Liesbeth, Amy Burkhardt, Wanjing Anya Ma, and Jason D Yeatman. 2024. "Rapid Online Assessment of Reading and Phonological Awareness (ROAR-PA)." *Sci. Rep.* 14 (1): 1–16.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

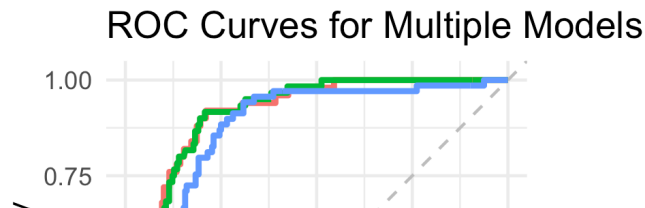
Wood, Simon, and Maintainer Simon Wood. 2015. "Package 'Mgcv'." *R Package Version* 1 (29): 729.



(a) 1st grade prediction of FAST™ CBMreading



(b) 2nd grade prediction of FAST™ CBMreading



29 Predictive validity of ROAR-RVP

29.1 Background: Published studies

The correlation between rapid visual processing and reading outcomes were first reported in a large cross-sectional sample by (Ramamurthy, White, and Yeatman 2023). Reading challenges are **not primarily an issue with visual processing**, though there is strong evidence that visual processing differences contribute to and/or exacerbate reading difficulties. Figure 29.1 shows published findings from Ramamurthy, White, and Yeatman (2023) demonstrating the relationship between RVP and reading ability is participants ages 6-18.

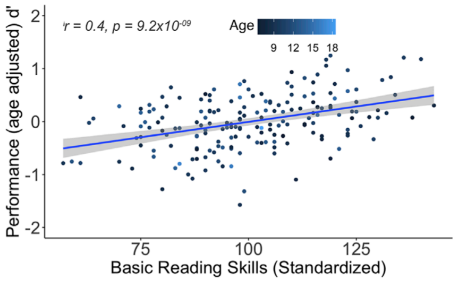


Figure 29.1: Differences in rapid visual processing predict reading ability across a broad aage range spanning kindergarten through adulthood. The x-axis show Woodcock Johnson Basic Reading Skills Standard Scores and the Y-axis shows age-standardized performance on the Rapid Visual Processing Task.

29.2 Correlating Reading Outcomes with Concurrent RVP Measures

In the calibration sample of n=175 kindergarten and first grade children (see Section 19.2.2.4) we replicated the correlation strength reported in the cross-sectional sample in the calibration cohort of kindergarten and first grade children (Ramamurthy et al., n.d.). Figure 29.2 shows the correlation between task performance in the Rapid Visual Processing and various reading outcomes.

29.3 Winter to spring predictions: RVP measured in the winter predicts end of year reading scores

In Winter 2023, 755 participants completed the RVP tasks and in Spring 2023 these students were administered Kaufman Test of Educational Achievement (KTEA) reading composite measure. The correlation strength between KTEA and task performance, in both the Letter (RVPL) [r = 0.42, p < 2.2x10-16; CI: 0.359 - 0.477] and pseudo-letter (RVPS) [r = 0.37,

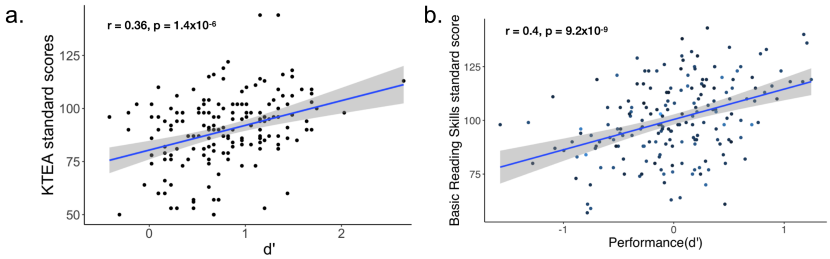


Figure 29.2: ROAR-RVP correlates with standardized reading outcome measures. Panel a shows the correlation with KTEA standard scores and panel b shows the correlation with Woodcock Johnson Basic Reading Skills/ Both correlations replicate previously published effect sizes (e.g., Section 29.1).

p < 2.2x10-16; CI: 0.307 - 0.431] versions were similar [□r: 0.0489; z = 1.9166, p = 0.0553; computed using the cocor package in R (Diedenhofen and Musch 2015) that uses Hittner, May, and Silver's (2003) modification of Dunn and Clark's z (1969) using a back transformed average Fisher's (1921) Z procedure]. Figure 29.3 shows the predictions of end of year reading scores. The correlation strengths were also comparable to our calibration data set and the previous cross-sectional study reported above in Figure 29.2.

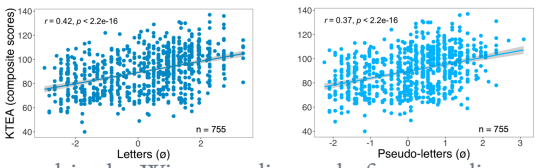


Figure 29.3: ROAR-RVP measured in the Winter predicts end of year reading assessments. The effect size is consistent with published studies (e.g., Section 29.1).

A regression model with both letter and pseudo-letter scores predicted KTEA better than either measure on its own [Model A. Performance in the RVPL task explained 17.5% of variance in KTEA composite scores; Model B. performance in the RVPS task explained 13.7% of variance in KTEA composite scores; and Model C. with both RVPL and RVPS (KTEA ~ letters + pseudo-letters) as predictors explained 18.6% of variance in KTEA composite scores. Model C is significantly better than either models: Models A vs C: F(1,752) = 10.609; p = 0.001; and Models B vs C: F(1,752) = 46.373, p = 2.002x10-11].

References

Diedenhofen, Birk, and Jochen Musch. 2015. "Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations." *PLoS One* 10 (3): e0121945.

Ramamurthy, Mahalakshmi, Klint Kanopka, Adam Richie-Halford, Benjamin W Domingue, Francesca Pei, Phaedra Bell, Lucy Yan, Andrea Hartsough, Maria Luisa Gorno-Tempini, and Jason D Yeatman. n.d. "Design and Validation of a Rapid Visual Processing Measure for Screening Reading Difficulties in Early Childhood."

Ramamurthy, Mahalakshmi, Alex White, and Jason D Yeatman. 2023. "Children with Dyslexia Show No Deficit in Exogenous Spatial Attention but Show Differences in Visual Encoding."

References