# A Fair Lexical Decision Task for Monolingual and Multilingual Spanish-speakers

Julian M. Siebert[1], Mia Fuentes-Jimenez[1], Wanying Anya Ma[1], Carrie Townley-Flores[1], Ana Saavedra[1], The ROAR Developer Consortium[1], and Jason D. Yeatman[1,2]

[1]Graduate School of Education, Stanford University

[2]Division of Developmental-Behavioral Pediatrics, Stanford University School of Medicine

## Author Note

Julian M. Siebert https://orcid.org/0000-0002-0472-4677

Mia Fuentes-Jimenez https://orcid.org/0009-0006-2682-2549

Wanying Anya Ma https://orcid.org/0000-0001-5761-8707

Carrie Townley-Flores https://orcid.org/0000-0002-7464-4840

Ana Saavedra https://orcid.org/0000-0001-6442-4919

Jason D. Yeatman https://orcid.org/0000-0002-2686-1293

We would like to thank the school districts, including Redwood City School District, families, and students that made this research possible through a research practice partnership model. We would also like to thank Sendy Caffarra for help developing and reviewing items. This work was funded by NICHD R01HD095861, the Stanford-Sequoia K-12 Research Collaborative, the Advanced Educational Research and Development Fund, Stanford Impact Labs, and Neuroscience:Translate grants to JDY.

Correspondence concerning this article should be addressed to Julian M. Siebert, Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305, Email: jms312@stanford.edu

## Abstract

This study describes the development and validation of ROAR Palabra, a novel Spanish lexical decision task designed for use with both Spanish-speaking children and Spanish-English bilinguals. This self-administered task requires students to decide whether a string of letters presented on the screen is a real word in Spanish. While there is evidence that scores on English lexical decision tasks are highly predictive of performance on conventional (time- and resource-intensive) word reading assessments in English (Yeatman et al., 2021), we explore whether this holds in Spanish, which has a much more transparent orthography. The specific goals are (i) to create a linguistically fair task and an item-response theory model for it and (ii) to evaluate whether such task can serve as a reliable proxy for conventional word reading measures, offering a quick and easy-to-administer tool for assessing reading skills across linguistic and cultural contexts. Results demonstrated strong correlations between performance on ROAR Palabra and standardized word reading assessments such as the Woodcock-Muñoz Batería IV, suggesting its effectiveness as a substitute measure. Notably, the task was sensitive to differences in language proficiency across both monolingual and multilingual groups, reflecting expected developmental and environmental influences. While not designed for the comparisons between monolingual and multilingual populations, the findings underscore the potential of this task as a versatile and culturally adaptable tool for reading assessments in different Spanish-speaking and bilingual contexts.


*Keywords:* multilingualism, Spanish, lexical decision task, reading assessment

## A Fair Lexical Decision Task for Monolingual and Multilingual Spanish-speakers

Reading proficiency is a cornerstone of academic achievement and cognitive development, influencing individuals' ability to engage with and comprehend written information across various contexts. Therefore, learning to read is one of the main goals in early elementary school education (Catts, 2021). Efficient and equitable assessment of reading skills is essential for the early identification of struggling students so that instruction can be tailored to each student's unique needs. However, traditional reading assessments are often time-consuming, resource-intensive, and may lack cultural and linguistic adaptability, particularly when applied to diverse populations such as multilingual individuals (Solano-Flores, 2016)—a population that is understudied and often misconceptualized (Bialystok, 2017; Cummins, 2000; Grosjean, 2008). In response to these challenges, there is a growing need for innovative assessment tools that are both reliable and scalable, capable of functioning effectively across different linguistic settings.

In English, lexical decision tasks (LDTs), which have a long tradition in cognitive science research (Balota et al., 2007), have been found to be efficient and reliable predictors of reading performance, correlating strongly with traditional assessments of, for example, word reading (Yeatman et al., 2021). In this study, we extend this approach to Spanish: We describe the development of ROAR Palabra, a novel self-administered Spanish LDT and investigate its relationship to traditional proctored assessment of Spanish word reading. Importantly, the task is designed around the linguistic diversity of both monolingual Spanish speakers in Latin America and Spanish-English bilinguals in the United States (US).

## Reading in Transparent Versus Opaque Orthographies

Orthographic transparency refers to the level of consistency in the graphemes-phoneme correspondence in a language's writing system. A language's orthographic transparency influences the cognitive processes involved in word recognition and reading fluency and, therefore, is a crucial consideration when developing any reading assessment. Spanish is characterized by a highly transparent (shallow) orthography, where most letters or combinations of letters reliably represents specific sounds. In contrast, opaque orthographies like English exhibit numerous

irregular spellings and inconsistent phoneme-grapheme mappings, which makes the process of reading—and learning to read— more complicated and thus extends the length and amount of instruction required to achieve mastery (Seymour et al., 2003; Ziegler & Goswami, 2005).

The transparency of the Spanish language facilitates the ease of acquisition of foundational reading skills. Research indicates that Spanish-speaking children typically develop phonological skills more rapidly than same-aged peers learning to read in less transparent languages (Ziegler et al., 2009). This accelerated acquisition of letter-sound correspondence and phonological awareness, in turn, allows for an earlier focal shift toward decoding skills and reading fluency (Aguasvivas et al., 2020).

Decoding skills allow readers to translate written text into spoken language based on acquired letter-sound correspondence. Over time, decoding becomes automatized allowing for rapid and accurate word recognition. Automated word-level decoding skills facilitate the reading of individual words and are precursors to sentence-level reading efficiency and comprehension (Ehri, 2005; Perfetti, 1985). The relatively straightforward syllable structure of Roman languages, characterized by predominantly open syllables (CV-CV) and limited consonant clusters, facilitates more efficient grapheme-to-phoneme mapping and thus enhances the ease of decoding for readers (Seymour et al., 2003). Therefore, in transparent orthographies where decoding is relatively straightforward, word recognition (alongside reading fluency) becomes one of the primary early indicators of reading proficiency.

## Lexical Decision Tasks

LDTs require participants to determine whether a string of letters constitutes a real word or a pseudoword, a process that necessitates both decoding skills and lexical retrieval. LDTs are particularly effective in assessing word decoding and are widely used in psycholinguistic research to study word recognition and lexical access (Balota et al., 2006; Katz et al., 2012; Keuleers & Brysbaert, 2011). The literature largely agrees on the assumption that the underlying visual word recognition processes of a two-alternative forced-choice (2AFC) design in LDTs mirror the cognitive processes at play during other word recognition tasks, such as single-word reading out

loud (Balota et al., 2006; Seidenberg & McClelland, 1989).

By measuring the accuracy and/or speed of responses on LDTs, such tasks can provide valuable insights into a student's reading development. The simplicity and efficiency of the task (easy and short administration) offers a quick and cost-effective means of gauging reading proficiency that can serve as a proxy for more comprehensive, individually-administered assessments. LDTs' utility in predicting performance on traditional reading measures has been well-documented, particularly in languages with complex orthographies like English. Yeatman et al. (2021) show that students' scores on an English LDT are highly correlated ($r$ = .94) with their scores on the Woodcock-Johnson Letter-word Identification subtask.

LDTs also offer a number of practical advantages over conventional word reading tasks, such as the Woodcock-Muñoz Letter-word Identification subtest (Woodcock et al., 2019). For one, their easy 2AFC design allows for objective, automatic, and immediate scoring without the need for verbalization. Because each item takes less than a second, the task can be completed within a few minutes and is amenable to computer adaptive testing (Ma et al., 2023). Last, the silent nature of a LDT allows for completion in a large group setting (e.g., classroom), which translates to less loss of instructional time and lower demands on resources.

The application of LDTs in languages with transparent orthographies presents unique opportunities and challenges. In Spanish, the ease of grapheme-phoneme correspondence may enhance the utility of LDTs in assessing word recognition and reading fluency, due to the relatively low decoding demand (Seymour et al., 2003). However, the higher transparency also means that LDTs must be carefully designed to differentiate between varying levels of lexical access and processing speed among different proficiency levels (Vega-Mendoza et al., 2015). Generally, research is sparse on the use of LDTs in Spanish (Aguasvivas et al., 2020), particularly for assessment of multilingual learners.

## Multilingualism

More than half of the global population is believed to be multilingual (Grosjean, 2010). In the US, about 10% of K-12 students speak a language other than English as their first language; in

California this holds true for about 20% of the population (California Department of Education, 2023; National Center on Improving Literacy, 2023). The development of most psychological and educational assessments, however, continues to largely operate from within a monolingual English mindset. Decisions based on inappropriate assessment choice or interpretation of results can have drastic consequences and may result in multilinguals' educational needs going unmet (Umansky, 2016). In this paper, we aim to shift this paradigm toward a more careful consideration of multilingual individuals in assessment development to ensure fairly comparable outcomes (Faulkner-Bond & Soland, 2020; Solano-Flores, 2023).

## Reading Development in Multlingual Individuals

Multilingual individuals are a heterogeneous population with different levels of language proficiency and reading skills across their languages. They exhibit unique developmental trajectories that reflect differences in the amount and sequence of language acquistion, exposure, formal and informal learning environments, as well as sociocultural context (Solano-Flores, 2016; Surrain & Luk, 2019). Durán et al. (2024), for example, report that multilingual Spanish-English bilinguals with different levels of English proficiency show different levels of growth on various foundational reading skills in kindergarten and first grade, when assessed in English. Especially students with high levels of proficiency in their different languages (balanced multilinguals) are able to tap into all of their languages' linguistic resources, which benefits their metalinguistic skills (e.g., phonological awareness) cross-linguistic knowledge transfer (Barac et al., 2014).

The effect of concurrent or sequential exposure to multiple languages and linguistic environments can allow for cross-linguistic transfer, where skills developed in one language influence the acquisition and proficiency of another (Cummins, 2000). In the context of reading, individuals may also transfer phonological awareness and decoding strategies from their dominant language to their second language, enhancing their reading development in both. The nature and extent of this transfer can vary depending on factors such as language similarity, proficiency levels, and the context of language use. For Spanish-English bilinguals, the transparent orthography of Spanish may facilitate the transfer of decoding skills to English, while the less transparent English

orthography may, in turn, influence reading strategies employed in Spanish (Bialystok, 2017).

## Assessing Multilingual Individuals

Understanding the dynamics of reading development is essential for developing assessments that accurately reflect the reading abilities of multilingual individuals. The variability in developmental trajectories of multilingual readers mean there is a need for assessment tools that are sensitive to these learning differences and can provide equitable measures of reading proficiency across both monolingual and multilingual populations. Unfortunately, most readily available reading assessments were designed with monolingual English populations in mind—as well as in the calibration and norming samples. These assessments do not adequately account for the heterogeneity and complexities of the multilingual experience, therefore often underestimating multilingual individuals' true linguistic abilities (Bialystok, 2001; Luk & Bialystok, 2013; Solano-Flores et al., 2009; Solano-Flores & Hakuta, 2017).

Linguistically fair assessment means for a measure to produce equally valid and accurate results for test-takers with different linguistic backgrounds (e.g., first languages, different levels of proficiency of the same language, etc.), but equal levels of the latent trait of interest. In other words, a linguistically fair measures for use with mono- and multilingual individuals must not be biased in favor or against those test-takers that are multilingual. This becomes a difficult endeavor when the trait to be assessed is a language-related construct, such as in the case of an LDT.

For LDTs, when used with multilingual populations, this means that they must account for cross-linguistic influences, cultural influences, and varying degrees of language dominance to ensure accurate assessment. Thus far, Spanish LDTs were successfully used with Spanish-English bilinguals in a sample of tertiary students with high and low English proficiency (Fairclough, 2011). Moreover, Aguasvivas et al. (2020), using LDTs as a measure of Spanish vocabulary, also found no statistically significant changes between monolingual and bilingual tertiary students. We are not aware of any study that examined this in younger populations of multilinguals.

## Resarch Questions and Aims

1. The first goal is to develop a reliable Spanish lexical decision task. Specifically, we aim to build the ROAR Palabra, a self-administered Spanish lexical decision task use with both monolingual children in Colombia and multilingual children in the United States.
2. Second, we investigate the efficacy of such a task for use as a proxy for Spanish single word reading skills, as measured by conventional, resource-intensive, proctored assessments.

## Methods

### Participants

Our sample ($N$ = 6448) comprises children from two locations: We recruited a mostly monolingual Spanish-speaking sample from Bogotá, Colombia, ($n$ = 5602), as well as a sample of Spanish-English multilingual children from across the United States (US), mostly from California ($n$ = 846). The Colombian sample comprises students in grades 1 to 11 at two public schools and one concession schools located in Bogotá, Colombia. Concession schools (colegios en concesión) were first launched in Bogotá in 2000; the government contracts private operators to run these schools. Concession school students, on overage, receive higher scores on national standardized tests (pruebas Saber) relative to students in other public schools. Our sample comes from a schools located in two different low-income neighborhood of Bogotá. Additional demographic information was limited.

Table 1 provides an overview of the US sub-sample's demographics. The majority of this sub-sample comes from a Northern Californian school district with a large proportion of students classified as English learners and an intentional focus on multilingual learning, manifesting in, for example, the provision of dual-language immersion programs. At other US sites, selection of students happened at the school's discretion, though they mostly also selected students classified as English learners or used teacher judgement.

**Table 1**

*United States Sub-sample's Demographic Characteristics.*

| **Characteristic** | **Gr 1** N = 327[1] | **Gr 2** N = 330[1] | **Gr 3+** N = 189[1] |
|---|---|---|---|
| Location (within US) | | | |
| California | 327 (100%) | 330 (100%) | 51 (27%) |
| Other | 0 (0%) | 0 (0%) | 138 (73%) |
| Gender | | | |
| Female | 159 (51%) | 129 (40%) | 1 (50%) |
| Male | 150 (49%) | 194 (60%) | 1 (50%) |
| Unknown | 18 | 7 | 187 |
| English proficiency designation | | | |
| English Learner | 202 (65%) | 211 (66%) | 0 (0%) |
| English-only | 73 (24%) | 75 (23%) | 1 (50%) |
| English-proficient | 34 (11%) | 36 (11%) | 1 (50%) |
| Unknown | 18 | 8 | 187 |
| Free or reduced-price lunch eligibility | | | |
| Eligible | 218 (70%) | 235 (72%) | 1 (25%) |
| Not eligible | 94 (30%) | 91 (28%) | 3 (75%) |
| Unknown | 15 | 4 | 185 |

[1] n (%)

## Measures

### *ROAR Palabra*

ROAR Palabra is a silent Spanish lexical decision task, which requires test-takers to decide whether an item flashing on the screen for 350 ms is a real Spanish word versus a made up word (pseudoword) and to respond via pressing a button on the keyboard/touchscreen. There is no limit

on the response time but the item is only presented for 350ms. It is an online assessment instrument, developed to accurately measure students' word reading ability in a time- and cost-efficient manner, doing away with the necessity for one-on-one assessments by trained assessment experts. It is modeled on ROAR-Word (Yeatman et al., 2021), which has shown to have high internal consistency reliability ($r$ = .95) and scores on which highly correlate with Woodcock-Johnson letter-word identification scores ($r$ = .94).

ROAR Palabra is explicitly *not* a translation of ROAR-Word—as a simple translation does not create equivalent versions of the same test (Solano-Flores et al., 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 Spanish words that are (i) frequent, (ii) known to pre- and middle-schoolers, (iii) known to Spanish-speakers across the Americas, (iv) and occurring in *all* the varieties of Spanish spoken there.

We then then used the Wuggy algorithm (Keuleers & 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. Five speakers of various versions and dialects of Spanish (from Colombia, Ecuador, Mexico, Spain, and United States) then 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 versions of Spanish were removed. With the Spanish-English bilingual context in the US in mind, we also removed generated pseudowords that were real words in English.

This process resulted in an initial item bank with 378 items (that is 189 real words and 189 matched pseudowords). To keep administration time reasonable, we selected 70 core items that were hypothesized to span a broad difficulty range (35 real and 35 pseudowords). We refer to the remaining 308 items as the extended corpus. Every test-taker responded to all core items, as well as random sample of additional (extended-corpus) items selected from the larger item pool.

### *Woodcock-Muñoz Batería IV*

To assess the degree to which performance on the silent, self-administered ROAR Palabra can function as a proxy for conventional, individually administered word- and nonword reading, we used two subtests of the WM (Woodcock et al., 2019)—a Spanish parallel of the Woodcock-Johnson IV (Schrank et al., 2014):

- The *identificación de letras y palabras* (letter-word identification; WM-LWID) test, measuring children's oral reading ability by having them read out aloud increasingly difficult words.
- The *análisis de palabras* (word attack; WM-WA) test, requiring children to read increasingly complex nonsense word out aloud, thereby tapping into their phonics and decoding skills.

Both tasks are scored for pronunciation accuracy by trained test-administrators following the guidelines of the scoring manual. The age-standardized scores of both the WM-LWID and WM-WA can be aggregated to provide a basic reading skills (WM-BRS) composite score.

## Procedures

In both settings, we worked closely with school partners in conducting this study. School partners provided information about the study to parents or guardians, who then had the opportunity to opt out if they preferred for their children to not participate in the study. Students saw assent forms on the screen before beginning ROAR-Palabra and the researchers ascertained verbal assent before completing WM testing. All protocols were approved by the Institutional Review Board at Stanford University.

In Colombia, we trained a group of twenty field assistants and a field coordinator for the administration of both ROAR Palabra and as well as both WM subtests. All students at participating schools took ROAR Palabra in the computer rooms of the school, unless their parents had opted out. Then, we randomly selected approximately 25% of those who had completed ROAR Palabra (balanced across grade levels) to also complete the WM-LWID and

WM-WA. For this, students were taken to a dedicated space in the school library or a multi-purpose room and completed the task on a laptop with a proctor–either in-person or on a laptop with headphones, connected to a proctor via a video-conferencing software. Scores obtained this way were double-scored by experienced WM administrators.

In the US, exact study procedures varied by school district. In most instances, schools made laptops or tablets available to students and tested whole classrooms at a time. Research coordinators provided training and (on-site) support before and during the administration periods.

## Analysis

In addressing the first aim of the study–building the first version of ROAR Palabra–we undertook several steps to obtain a final item-response theory (IRT) model. Prior to doing model building, we (i) filtered responses based on children's median response times (< 450 ms) and correctness rate (< 65%) to exclude random guessers, (ii) excluded items with low (< .10) point-biserial correlations to ROAR Palabra (core corpus only) totals and WM-LWID totals, and (iii) excluded items with suboptimal item in- and out-fit (< .60 or > 1.40) within the core corpus. We applied these criteria separately to both the US and Colombian sub-sample, so that characteristics of both populations are represented in the final item selection.

Using the responses retained after excluding rapid guessers and with those core items surviving the point-biserial correlation exclusion criteria, we iteratively fit a 1PL model and excluded all items with poor fit until we obtained a stable model. We fit a 1PL model of the form

$$P(X_i = 1|\theta) = 0.5 + 0.5\frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}, \tag{1}$$

where $P(X_i = 1|\theta)$ is the probability of a correct response given item $i$'s difficulty level, $b_i$, which measured on the same scale as the respondent's ability, $\theta$. Given the 2AFC task design, we imposed a .50 lower bound on the probability of a correct response (guessing parameter). We used the `mirt` package (Chalmers, 2012) for `R` (R Core Team, 2018) for all IRT analyses and calculated theta scores using the default EAP estimator.

We then fit a two-parameter (2PL) model, in order to be able to evaluate item

discrimination parameters. While this model is not used for final theta estimation, items' discrimination parameters indicate how effectively an item differentiates between respondents with varying levels of the latent trait being measured. This provides important information about the item's quality and usefulness in the final measure.

Next, we fit another (final) 1PL model using Equation 1. This time, we included all those items in the *entire* (core and extended) item bank that survived the exclusion criteria outlined above. We fixed the scale using the item parameters obtained in the 1PL model for the core-corpus items and only estimated item parameters for the extended-corpus items. Again, we fit a 2PL model for the purpose of obtaining item discrimination parameters.

Following this, we evaluated the reliability of the final (1PL) model. Given that ROAR Palabra is a fixed-length task scored using a 1PL model, the appropriate reliability metric is empirical reliability ($\rho_{xx'}$), estimated using Equation 2.

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

Then, we assessed the final model's parameter invariance—that is, we checked whether item difficulty and discrimination parameters are significantly different in the two sub-samples. To do so, we compared item parameters from a jointly calibrated 1PL model to parameters obtained from 1PL models fit separately for each sub-sample, as well as parameters from the two separately calibrated models. Because the two sub-samples cover very different grade ranges, we conducted the parameter invariance analysis on a separate set of models using only data from respondents in the overlapping grade range.

In a next step, we assessed ROAR Palabra's criterion validity. For this we used the ROAR Palabra theta scores obtained using the final 1PL model (and expected a-posteriori [EAP] estimation) the Colombian students' raw scores on the WM-LWID, WM-WA,and WM-BRS. We correlated students' observed WM scores with predicted WM scores obtained from a generalized additive model with a smooth function on ROAR Palabra theta scores. Finally, given the sub-samples different grade ranges, we repeated all analysis steps using only the overlapping grades as a sensitivity analysis in the appendix.

# Results

## Item Responses

For the 70 core items, we have 5602 observations per item for the Colombia sub-sample and 846 per item for the US sub-sample. For the extended corpus items, while we observe sufficiently large numbers for the purpose of item calibration for the Colombia sub-sample, the response counts from the US are too small to reliably calibrate an IRT model to the US sub-sample. Therefore, we refrain from comparing performance on the extended corpus and restrict our detailed analyses and item parameter estimation to the core corpus. Afterwards, we refit the model with the extended-corpus items while holding the core items' parameters fixed, so that the extended-corpus corpus items are calibrated to the same measurement scale that was defined based on the detailed analysis of the core item bank.

## Sample Performance and Median Response Times

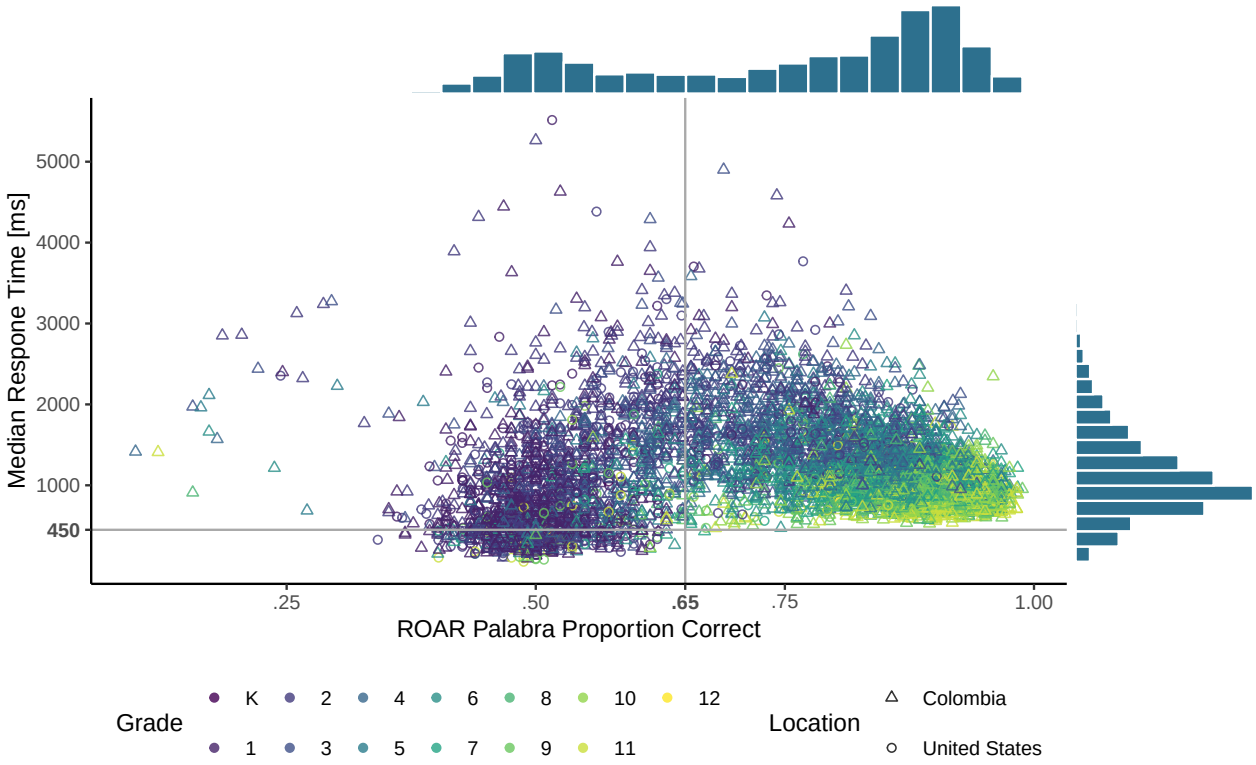
Figure 1 shows median response times as a function of raw scores (calculated as the percentage of correct responses), disagreggated by grade. Barely any of the students performing above chance exhibit median response times < 450 ms. At the same time, students with extremely fast response times (<450ms) perform around the chance level, which is likely indicative of rapid guessing. The bimodality of the raw score distribution can be explained by the large grade range, which includes both children still learning how to read words, as well as high schoolers who largely mastered that skill. Indeed, Figure B1 shows the same analysis for only those grades (1 and 2) that are represented in both samples; the overall patterns are very similar and no difference based on study location is observed.

## Item Properties

For the initial descriptive analysis, item difficulty is calculated as the proportion of all respondents that responded correctly to a given item. Panel A in Figure 2 shows that, for both sub-samples, item difficulty follows a bimodal distribution with real words (right peak) being easier than pseudowords (left peak). The shift in positions of the distributions indicates that, on

**Figure 1**

*Median Response Time as a Function of Raw (Proportion Correct) Score on ROAR Palabra.*




average, items are easier for the Colombian students (which cover a much larger grade range). The very high correlation between item difficulty in the two sub-samples ($r$ = 0.93) indicates that the relative positions of items on item difficulty distribution are very similar in both sub-samples. This finding also holds in the separate analysis for grades 1 and 2 only (Panel A of Figure B2).

Point-biserial correlations between student responses to a given ROAR Palabra item and their raw ROAR Palabra score (proportion correct) are indicators of the degree to which an individual item taps into the same construct that is measured by the overall scale. Panel B of Figure 2 shows that ROAR Palabra forms a mostly very coherent scale; within the Colombian sub-sample, all correlations are > .20, within the US sub-sample, the majority is, too.

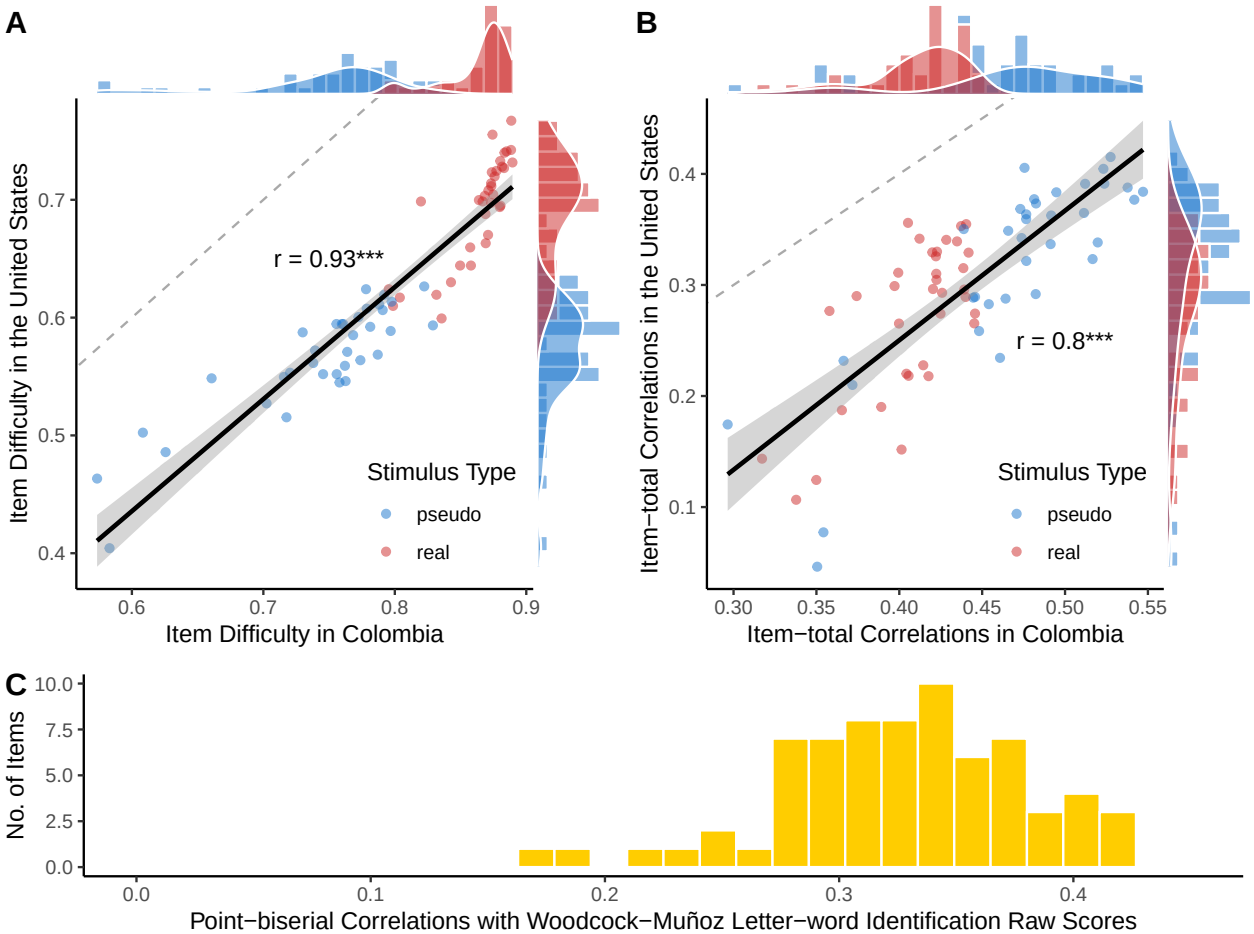
Further, point-biserial correlations between students' responses to individual ROAR Palabra items and their WM-LWID raw score indicate to what extent each item is related to the construct of word reading. Panel C in Figure 2 shows the distribution of these point-biserial

correlations. While some real words show very low (< .10) point-biserial correlations, the majority of items fall into an acceptable range.

Overall correlations of item parameters between the two study locations show similar trends when separately analysed for the lower grades only. Panels A to C of Figure B2 report repetitions of the analyses described here for the subs-sample of students in grades 1 and 2 only.

**Figure 2**

*ROAR Palabra Item Properties with Item Difficulty (Proportion Correct) Distribution in Panel A, Item-total (Point-biserial) Correlations in Panel B, and Item-WM-LWID (Point-biserial) Correlations in Panel C (Colombian Subsample Only), for Core Items.*



## IRT Model Building

We started the model building process with the 70 items in the core corpus (35 real words and 35 pseudowords), because we had sufficiently large numbers of observations in both contexts. For the extended-corpus items, of which test-takers only saw random selection of 30 out of the 308 items, response counts in the US sub-sample were low. Therefore, we decided to add those items at a later stage, after the calibration of a measurement model based on the core corpus. Prior to the estimation of an IRT model, we carried out four item selection steps as follows:

### *Participant Exclusion Criteria: Median Response Time*

As a first data cleaning step prior to calibrating an IRT model, we excluded data from participants whose response behaviour was indicative of random guessing or clicking through the task without a serious attempt at the task. This was operationalized as a median response time < 450 ms and a score of < 65 % correct. This resulted in an exclusion of 6.95 % ($n = 448$) of participants in total, 5.73 % ($n = 321$) of Colombian participants and 15.01 % ($n = 127$) of participants from the US. Item counts are not affected by this exclusion criterion.

### *Item Exclusion Criteria*

**Criterion 1: Item-total Correlations.** As a first step toward ensuring we are measuring a single and coherent construct, we proceeded to eliminate all those items that exhibited point-biserial correlations with the total task score (proportion of correct responses) of less than .10—a very lenient threshold. In other words, this means removing those items whose response patterns are unrelated to the overall proportion correct scores. To account for both contexts, we applied this criterion separately to the Colombian and the US sub-sample. This resulted in the exclusion of 3% of items (2 items; 0 real words and 2 pseudowords) based on data from the US. The pseudowords excluded were *cumpleapos, estudionte*. No items had to be excluded based on data from Colombia.

**Criterion 2: Item-WM Correlations (Colombia only).** Next, we computed correlations to WM-LWID raw scores and had planned to exclude all items that exhibited point-biserial correlations < .10. No items were flagged in this stage.

**Criterion 3: Item-fit.** Next, we iteratively fit a 1PL model and assessed item fit. A lenient range for good item in-fit and out-fit parameters is .60–1.40. Only 2 items fell outside this range.

### *Core Model*

After applying the four above criteria, the resultant core item pool contains a total of 66 items (33 real words and 33 pseudowords). While implemented using a 1PL model, we also fit a 2PL model to obtain item discrimination information. The item discrimination value ($\alpha$) indicates the steepness of the slope of the item characteristic curve and is an indicator of how well that item discriminates between two respondents whose ability levels are around the item's difficulty. Typically, the range of the $\alpha$ parameter is from 0 to 3, with an $\alpha < .50$ indicating less productive measurement. None of the items fall below that threshold. Moreover, as was the case with the raw score, the distributions of theta scores show differences between the US and Colombian sub-samples, though these disappear when filter to only draw on the earlier grades.

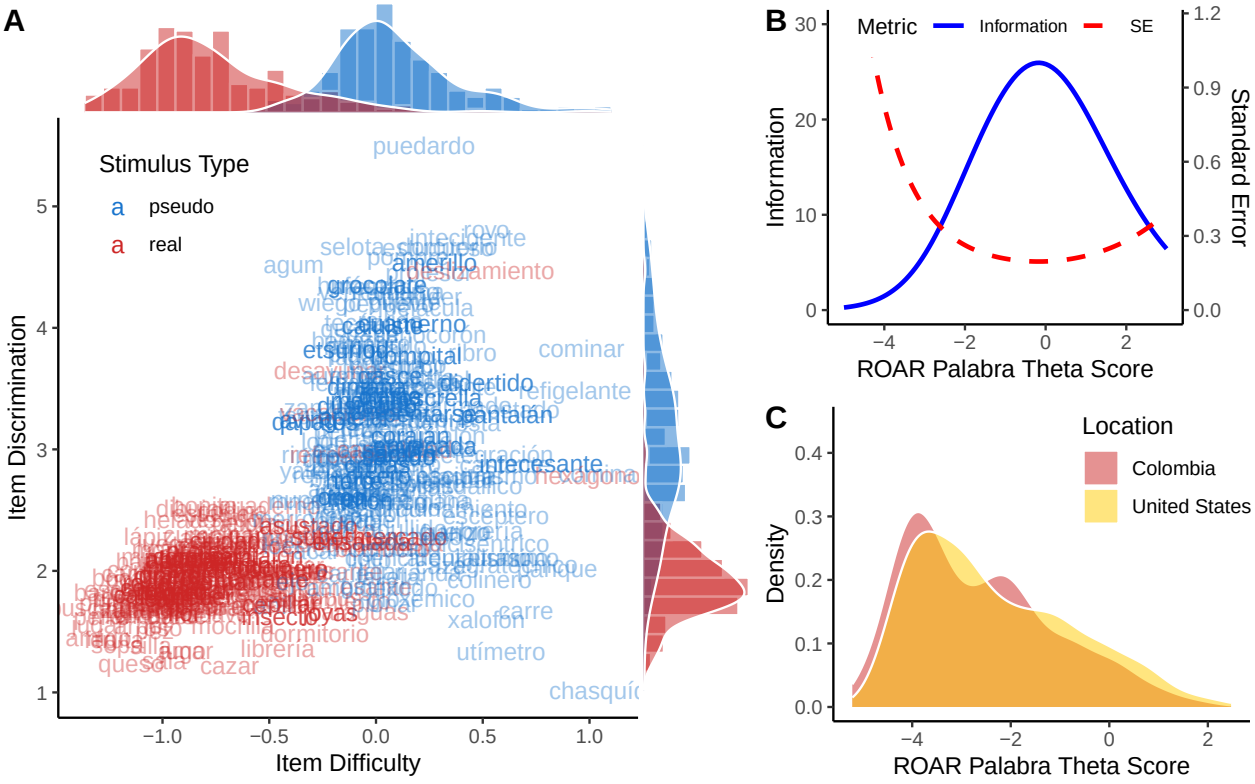
### *Final Model*

Before adding the items from the extended corpus to the core measurement model, we subjected them to the same four criteria we used to prune the core item pool. We then fixed the core items' parameters and refit a 1PL model with guessing parameter to estimate the new (extended-corpus) items' difficulty parameters. We also reran the 2PL model to obtain item discrimination estimates, this time only estimating the new items' discrimination parameters. Figure 3 summarizes the final models' characteristics and performance: Panel A shows the resultant distribution of item difficulty and discrimination parameters; Panel B shows the test information curve and the associated standard error; and panel C shows the distribution of theta scores by location for grades 1 and 2 (which can be compared fairly, as they are covered in both sub-samples), which closely mirror the distributions obtained using the core model. Figure A1 shows the theta score distribution for the entire grade distribution.

**Reliability.** We estimated empirical reliability for the final model ($n$ = 6000), comprising both core and extended-corpus items, using Equation 2. Overall reliability is high, with $\rho_{xx'}$ = 0.938, as are reliability estimates for Colombia ($\rho_{xx'}$ = 0.936, $n$ = 5281) and the United

**Figure 3**

*Summary of Final ROAR Palabra Item-response Theory Model, Showing the Bivariate Distribution of Item Difficulty and Discrimination (Panel A), the Test Information Curve With the Associated Standard Error (SE; Panel B), and the Distribution of Theta Scores for the Overlapping Grade Range (Grades 1 and 2) by Location (Panel C).*




States ($\rho_{xx'}$ = 0.887, $n$ = 719), separately. Table 2 shows empirical reliability estimates by grade, drawing on both the Colombian and US data.

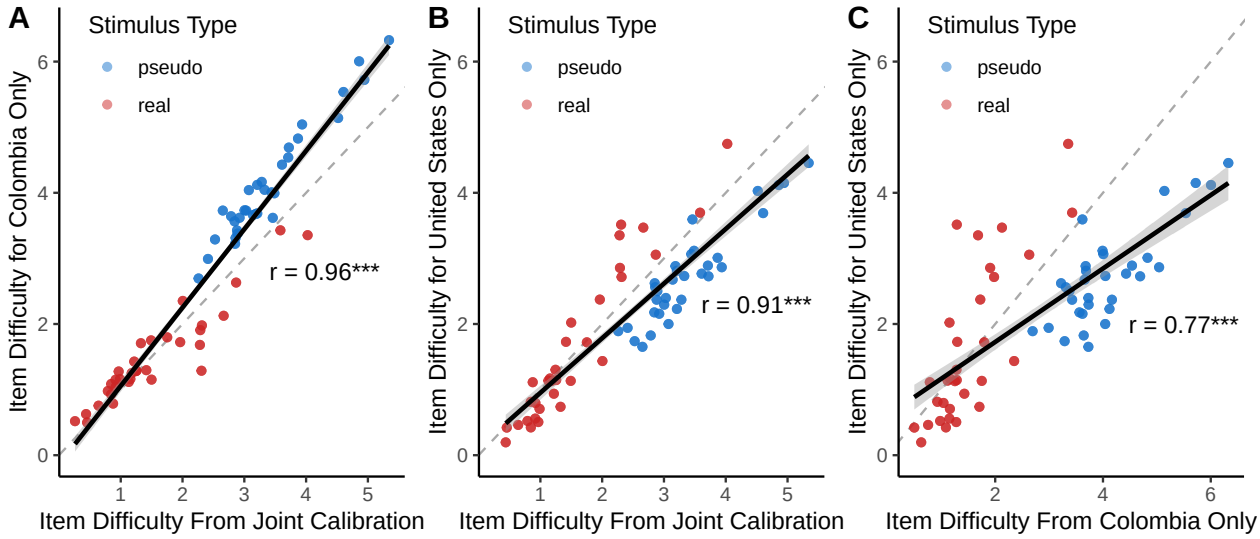
**Table 2**

*ROAR Palabra Empirical Reliability by Grade (Colombia and US).*

| | All | Gr 1 | Gr 2 | Gr 3 | Gr 4 | Gr 5 | Gr 6 | Gr 7 | Gr 8 | Gr 9 | Gr 10 | Gr 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 5996 | 645 | 800 | 532 | 520 | 580 | 616 | 544 | 436 | 467 | 431 | 425 |
| r | 0.938 | 0.705 | 0.877 | 0.918 | 0.919 | 0.910 | 0.890 | 0.860 | 0.847 | 0.866 | 0.797 | 0.839 |

**Figure 4**

*Parameter Invariance Analysis for 1-Parameter Logistic Model (Grades 1 and 2 Only) in the Form of Correlations Between Jointly and Separately Calibrated Item Parameters for Colombian Sub-sample (Panel A) and United States Sub-sample (Panel B), as Well as Between the Two Separately Calibrated Models (Panel C).*




**Parameter Invariance.** Next, we assessed parameter invariance. To account for the difference in grade ranges, we fit another set of 1PL models using only data from respondents in those grades represented in both samples (grades 1 and 2). We compared item parameters of a jointly calibrated IRT model to parameters for separately calibrated models, as well as correlations. between parameters of the two separately calibrated models. Figure 4 shows the resultant correlations. Both sub-samples' item parameters are very highly correlated with those obtained from a joint calibration. The correlation between separately calibrated US and Colombian parameters, though somewhat lower, still suggests that parameters are similar in both contexts. Here, the lexicality effect, with pseudowords being easier for the Colombian subsample warrants further investigation.

## Validity Evidence

These analyses draw on a sub-set of the Colombian sub-sample. To assess whether ROAR Palabra scores can be used of indicators of word reading performance, we correlated students'

ROAR Palabra theta scores (obtained from the final model) with their scores on the Woodcock-Muñoz Letter-word Identification and Word Attack scores (panels A and B of Figure 5, respectively).

Cross-sectional growth patterns can provide additional validity evidence. As children progress through the grades, their score on lexical decision tasks is reasonably expected to increase, given the additional reading instruction and vocabulary expansion. Therefore, a good lexical decision task should produce higher scores for students in higher grades. Panel D in Figure 5 shows that mean ROAR Palabra scores increase monotonically across grade levels.

## Discussion

This study investigated the feasibility of fairly using the same Spanish lexical decision task (LDT) with both monolingual Spanish-speaking and Spanish-English bilingual students, as well as such a task's utility as a proxy for traditional, proctored word reading assessments. Specifically, we (i) successfully developed ROAR Palabra as a linguistically fair Spanish LDT with very similar item parameters and score distributions among US and Colombian first- and second-graders and (ii) showed its moderate to high correlations with the Woodcock-Muñoz Batería IV Letter-word Identification and Word Attack subtasks. Additionally, we found that–for both mono- and multilinguals–item difficulty is affected by lexicality.

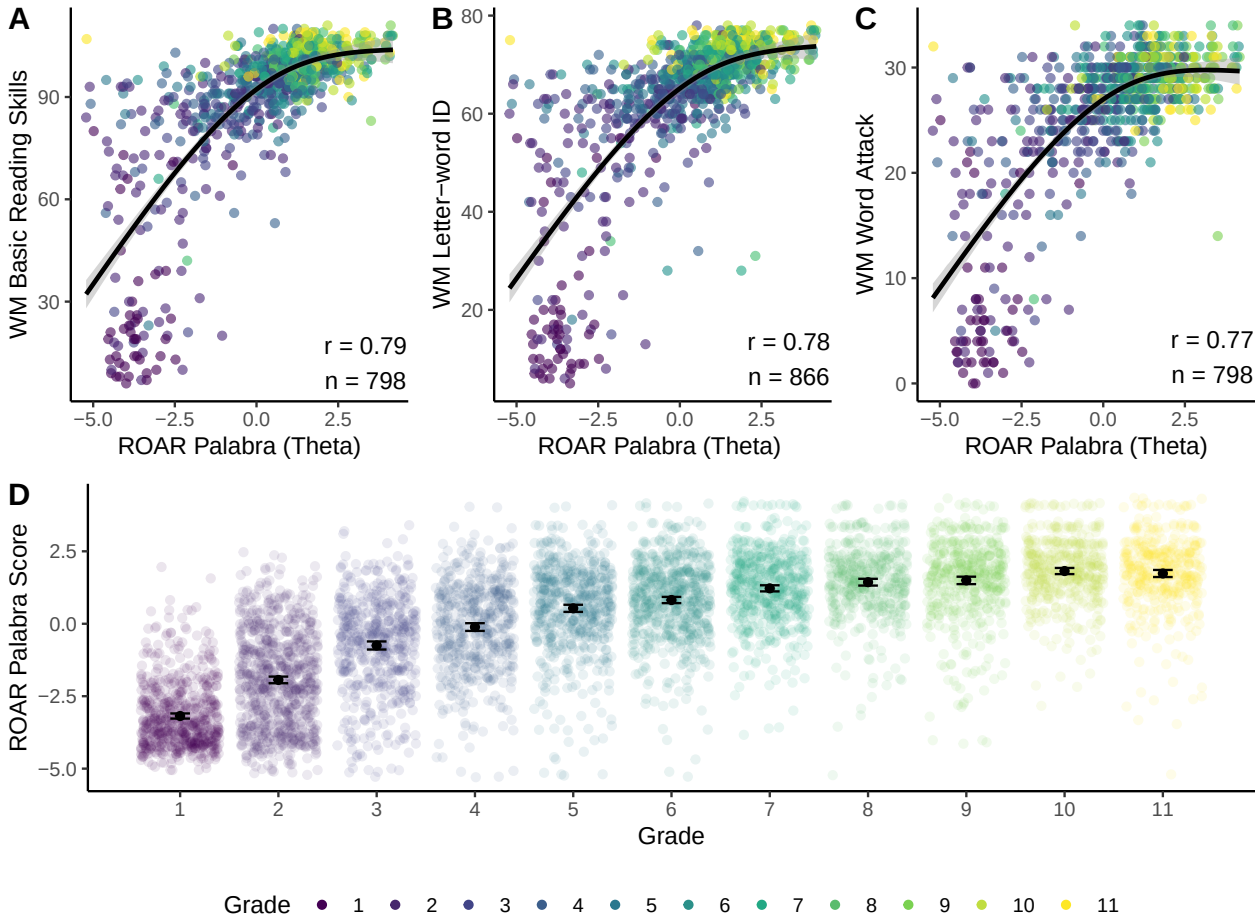
### Lexicality Affects Item Difficulty

The bimodal distributions of item difficulty and item discrimination (panel A of Figure 3) suggest the presence of a lexicality effect. Items cluster closely together and the cluster of real-word items is less difficult and less discriminating than the cluster of pseudoword items. This means that correctly responding to real-word items (i.e., recognising known words) is easier than correctly identifying pseudowords as such. Moreover, this tells us that pseudowords are more useful in telling apart high- from low-performers.

We are confident that this difference is largely based on lexicality. Given that we effectively controlled for length, phonotactic constraints, and orthographic neighbourhood size when constructing the pseudoword-items by using the Wuggy algorithm (Keuleers & Brysbaert,

**Figure 5**

*Validity Evidence for ROAR Palabra By Means of Correlations Between ROAR Palabra Theta Scores and Woodcock-Muñoz Basic Reading Skills (Panel A), Letter-word Identification (Panel B), and Word Attack (Panel C) Raw Scores, as well as Cross-sectional Growth Across Grades on ROAR Palabra (Panel D) For the Colombian Sub-sample.*




2010), these stimuli characteristics are similarly distributed within the real-word and pseudoword item groups. This lexicality effect holds true for both the mono- and the multilingual sample and is consistent across grades.

This clustering of items is not observed in other (less transparent) languages. In a very similar English task, Yeatman et al. (2021), for example, did not observe this pronounced difference in item difficulty and discrimination between real words and pseudowords.

This lexicality effect likely generalizes to other languages with similarly transparent

orthographies, but further research is needed to confirm this. In addition to that, this finding could be corroborated if this pattern holds even with a larger set of real-word items that are less common, longer, or more difficult due to some other item features.

### Decoding vs. Vocabulary

One possible explanation for the presence of a lexicality effect in this Spanish LDT, but not in a comparable English task, is the lower decoding demand in Spanish (Ziegler & Goswami, 2005). Given that the LDT task design sees the stimuli only appear briefly, efficient decoding is necessary in order to, in a second step, make a lexicality decision. In Spanish, due to its transparent orthography, the ability to correctly decode both words and pseudowords develops much faster than other reading skills, so that Spanish readers can successfully decode text before they can comprehend it (López-Escribano et al., 2013).

This would suggest that Spanish LDTs load less on efficient decoding skills, but are more directly affected by vocabulary size–or the ability to recognise a known word. Indeed, others have also used Spanish LDTs to assess vocabulary size (Aguasvivas et al., 2020). Our findings of lower correlations between ROAR Palabra scores and WM-LWID and WM-WA scores (both of which assess decoding) compared to parallel English analyses (Yeatman et al., 2021) supports this hypothesis.

### Linguistically Fair Assessment?

A linguistically fair assessment instrument produces equally accurate results for test-takers who have different linguistic backgrounds but equal levels of mastery of the target construct. In the case of LDTs, this begs the question as to whether we would expect all multilinguals in a certain developmental stage–regardless of amount and context of language experience–to have achieved the same level of mastery of Spanish lexical decisions. Here, developmental stage is operationalized as grade level. Thus, this study suggests that, when developed carefully, the same Spanish LDT may be used with both monolingual and multilingual Spanish-speaking first- and second-graders–at least in Colombia and the US. We presented evidence showing that ROAR Palabra item parameters are not statistically significantly different in the two contexts and that

final model theta scores for grades 1 and 2 are almost identically distributed.

This is in line with Aguasvivas et al. (2020), who found no statistically significant differences in vocabulary size assessed via LDT between monolingual and bilingual Spanish-speaking adults. In contrast, Izura et al. (2014) showed large difference between first- and second-language speakers of Spanish in favour of the former on the LexTALE-Esp, a similar task requiring the correct identification of presented stimuli as words or pseudowords. One possible explanation for this difference might be the different age groups for the sample; our sample of students in grades 1 and 2 is likely to largely comprise beginning decoders and that differences might only start manifesting later.

Finally, while we establish parameter invariance and obtain very similar score distributions for the grade range compared, these findings ought to be corroborated by additional bias analyses, especially for larger grade ranges. Nonetheless, the present findings strongly suggest that ROAR Palabra is suitable for use in both populations represented in the sample. Caution is to be exerted, however, when comparing students from different backgrounds, especially when extrapolating to grades 3 and higher.

### Next Steps and Limitations

Our data is currently imbalanced, in favour of the Colombian context. Therefore, we are currently collecting both ROAR Palabra and WM-LWID data from a more sizeable sample of Spanish-English bilinguals in higher grades in the US to further corroborate our claims. This will allow us to also provide criterion validity evidence for the US context. Even though the parameter invariance analysis suggested that item parameters are sufficiently similar in both the US and Colombian sub-samples, this will also need to be verified with the larger sample. Particularly for the items in the extended corpus, our observation counts in the US sub-sample are too low to inform item pruning. Additionally, the suggested lexicality effect warrantss further investigation.

Furthermore, we plan to create new items—particularly more difficult real words—in order to increase the size of the item bank. A larger item bank with more well-performing items is necessary for an efficient use of a CAT algorithm. Though unlikely, once we will have collected

more (US) data on the extended-corpus items, as well as on a set of additional items, we will have to revisit the IRT model and decide whether a re-calibration is warranted.

Additional analyses of interest relate to the features of individual items. Which features (length, cognates, age of acquisition, etc.), in addition to stimulus type, make items easier or more difficult? Analysing the characteristics of well-performing items will also help facilitate longer-term item bank development and inform other scholars developing Spanish reading measures.

Other important questions that need answering pertain to the generalizability of these findings. Do these findings hold true across other Spanish-speaking populations, or other school types or socioeconomic strata? It would also be particularly insightful to disaggregate the US data by instructional model, in order to check for effect of children' language of instruction on ROAR Palabra scores and to reflect the notion that multilinguals are a heterogeneous population (Solano-Flores et al., 2009; Solano-Flores & Hakuta, 2017).

## Conclusion

Linguistically fair behavioral assessment for the growing multilingual population is a pressing global need. ROAR-Palabra is a reliable Spanish lexical decision task poised to respond to this. It was specifically developed for use with both first- and second-graders in monolingual Spanish-speaking settings, as well as in multilingual settings, such as with Spanish-speaking bilinguals in the US. Additionally, sensitivity to cross-sectional growth across grades and moderate to strong correlations with the Woodcock-Muñoz Basic Reading Skills cluster underpin its potential as an efficient proxy for otherwise time- and cost-intensive proctored reading assessments.

Beyond presenting psychometric evidence for this new task, however, this paper also functions as a blueprint for the development of linguistically fair assessment instruments. We argue that the development of a task intended for use in multilingual populations requires careful subgroup-specific analyses and, most importantly, the consideration of this population and its linguistic and cultural context throught the *entire* development process. The former consideration

manifests in subjecting the item pool to subgroup-specific item pruning and establishing parameter invariance between mono- and multilingual speakers. The latter substantiates in the representation of speakers of multiple varieties of Spanish, including bilingual Spanish-English speakers, in the developer team to ensure an unbiased item pool, as well as in the inclusion of multilingual speakers in the model calibration (and later norming) sample.

# Declarations

## Funding

This work was funded by NICHD R01HD095861, the Stanford-Sequoia K-12 Research Collaborative, the Advanced Educational Research and Development Fund, Stanford Impact Labs, and Neuroscience:Translate grants to JDY.

## Conflicts of Interest

None declared.

## Ethics Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Stanford Institutional Review Board (IRB).

## Consent

All individual participants or their parents/guardians were provided with consent/assent information and were given sufficient opportunity to opt out of participation in the studies.

## Consent for Publication

Not applicable.

## Open Practices Statement/Availability of Data and Materials

Anonymised data and code are available in this OSF Project: https://osf.io/rhu3w/.

## References


Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. *Behavior Research Methods*, *52*(5), 1867–1882. https://doi.org/10.3758/s13428-020-01357-9

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual Word Recognition. In *Handbook of psycholinguistics* (pp. 285–375). Elsevier.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Barac, R., Bialystok, E., Castro, D. C., & Sanchez, M. (2014). The cognitive development of young dual language learners: A critical review. *Early Childhood Research Quarterly*, *29*(4), 699–714. https://doi.org/10.1016/j.ecresq.2014.02.003

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge University Press.

Bialystok, E. (2017). The Bilingual Adaptation: How Minds Accommodate Experience. *Psychological Bulletin*, *143*(3), 233–262. https://doi.org/10.1037/bul0000099

California Department of Education. (2023). *English Learners by Grade & Language*. https://www.cde.ca.gov/ds/ad/fileselsch.asp

Catts, H. W. (2021). Commentary: The critical role of oral language deficits in reading disorders: reflections on Snowling and Hulme (2021). *Journal of Child Psychology and Psychiatry*, *62*(5), 654–656. https://doi.org/10.1111/jcpp.13389

Chalmers, R. P. (2012). mirt : A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire*. Multilingual Matters.

Durán, L., Siebert, J. M., Zegers, M., Gutiérrez, N., Pei, F., Catts, H., Petscher, Y., &

Gorno-Tempini, M. L. (2024). Comparing the Performance and Growth of Linguistically Diverse and English-only Students on Commonly Used Early Literacy Measures. *Under Review*.

Ehri, L. C. (2005). Learning to Read Words: Theory, Findings, and Issues. *Scientific Studies of Reading*, *9*(2), 167–188. https://doi.org/10.1207/s1532799xssr0902/_4

Fairclough, M. (2011). Testing the lexical recognition task with Spanish/English bilinguals in the United States. *Language Testing*, *28*(2), 273–297. https://doi.org/10.1177/0265532210393151

Faulkner-Bond, M., & Soland, J. (2020). Comparability When Assessing English Learner Students. In J. W. P. Amy I Berman Edward H Haertel, *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 149–176). National Academy of Education.

Grosjean, F. (2008). *Studying Bilinguals*. Oxford University Press.

Grosjean, F. (2010). The bilingual as a competent but specific speaker-hearer. *Journal of Multilingual and Multicultural Development*, *6*(6), 467–477. https://doi.org/10.1080/01434632.1985.9994221

Izura, C., Vuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*(1), 49–66.

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2012). What lexical decision and naming tell us about reading. *Reading and Writing*, *25*(6), 1259–1282. https://doi.org/10.1007/s11145-011-9316-9

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/brm.42.3.627

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon*, *6*(1), 34–52. https://doi.org/10.1075/ml.6.1.02keu

López-Escribano, C., Juan, M. R. E. de, Gómez-Veiga, I., & García-Madruga, J. A. (2013). A predictive study of reading comprehension in third-grade Spanish students. *Psicothema*,

*25*(2), 199–205. https://doi.org/10.7334/psicothema2012.175

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

Ma, W. A., Richie-Halford, A., Burkhardt, A. K., Kanopka, K., Chou, C., Domingue, B. W., & Yeatman, J. D. (2023). ROAR-CAT: Rapid Online Assessment of Reading ability with Computerized Adaptive Testing. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/7tpx2

National Center on Improving Literacy. (2023). *State of dyslexia*. https://improvingliteracy.org/state-of-dyslexia

Perfetti, C. (1985). *Reading Ability*. Oxford University Press.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV*. Riverside.

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295x.96.4.523

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143–174. https://doi.org/10.1348/000712603321661859

Solano-Flores, G. (2016). *Assessing English language learners: Theory and practice*. Routledge.

Solano-Flores, G. (2023). How Serious are We About Fairness in Testing and How Far are We Willing to Go? A Response to Randall and Bennett with Reflections About the Standards for Educational and Psychological Testing. *Educational Assessment*, *28*(2), 105–117. https://doi.org/10.1080/10627197.2023.2226388

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of Test Translation Error. *International Journal of Testing*, *9*(2), 78–91. https://doi.org/10.1080/15305050902880835

Solano-Flores, G., & Hakuta, K. (2017). *Assessing Students in Their Home Language*.

Surrain, S., & Luk, G. (2019). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition*, *22*(2), 401–415. https://doi.org/10.1017/s1366728917000682

Umansky, I. M. (2016). To Be or Not to Be EL. *Educational Evaluation and Policy Analysis*, *38*(4), 714–737. https://doi.org/10.3102/0162373716664802

Vega-Mendoza, M., West, H., Sorace, A., & Bak, T. H. (2015). The impact of late, non-balanced bilingualism on cognitive performance. *Early Childhood Research Quarterly*, *137*, 40–46. https://doi.org/10.1016/j.cognition.2014.12.008

Woodcock, R. W., Alvarado, C. G., Schrank, F. A., McGrew, K. S., Mather, N., & Muñoz-Sandoval, A. F. (2019). *Batería IV Woodcock-Muñoz*. Riverside.

Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., Caffarra, S., Takada, M. E., Kanopka, K., Ben-Shachar, M., & Domingue, B. W. (2021). Rapid online assessment of reading ability. *Scientific Reports*, *11*(1), 6396. https://doi.org/10.1038/s41598-021-85907-x

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2009). Orthographic Depth and Its Impact on Universal Predictors of Reading. *Psychological Science*, *21*(4), 551–559. https://doi.org/10.1177/0956797610363406

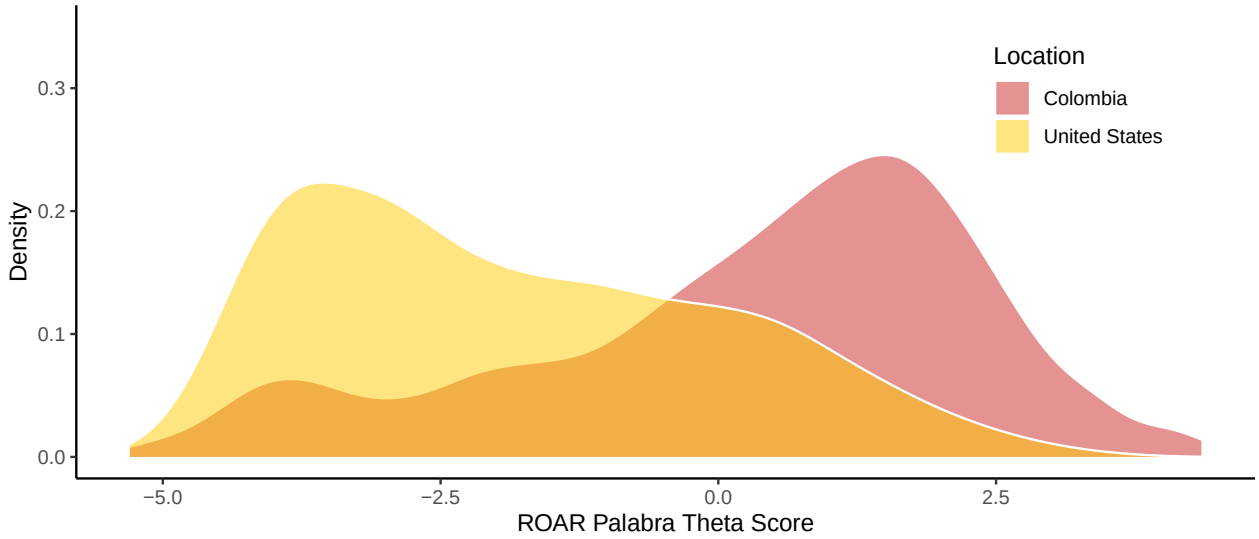
Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

# Appendix A

## Theta Distribution (All Grades)

**Figure A1**

*Distribution of Theta Scores Obtained from Final Model (All Grades).*



# Appendix B

## Re-analysis for Grades 1 and 2 Only

**Figure B1**

*Median Response Time as a Function of Raw (Proportion Correct) Score on ROAR Palabra (Grades 1 and 2 Only).*

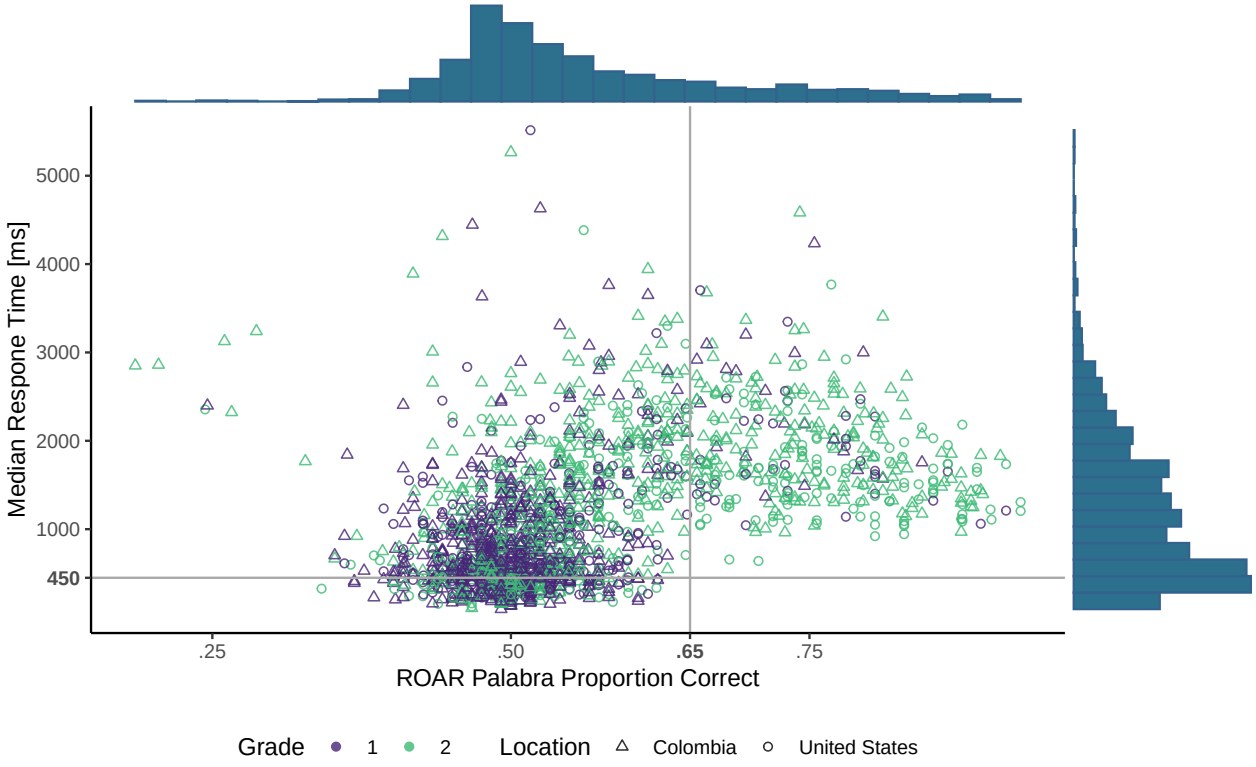


**Figure B2**

*ROAR Palabra Item Properties with Item Difficulty (Proportion Correct) Distribution in Panel A, Item-total (Point-biserial) Correlations in Panel B, and Item-WM-LWID (Point-biserial) Correlations in Panel C (Colombian Subsample Only), for Core Items (Grades 1 and 2 Only).*

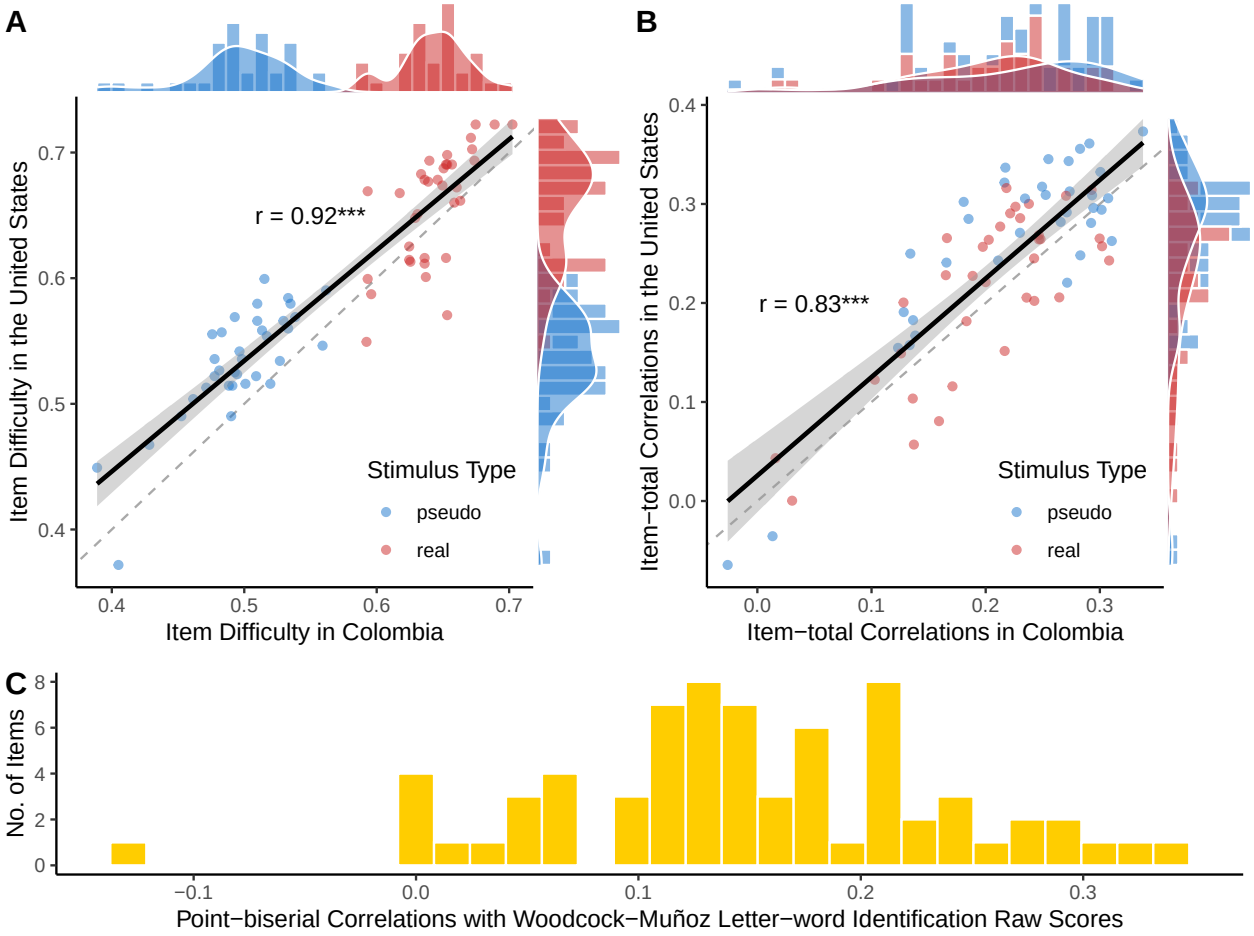


**Figure B3**

*Validity Evidence for ROAR Palabra By Means of Correlations Between ROAR Palabra Theta Scores and Woodcock-Muñoz Basic Reading Skills (Panel A), Letter-word Identification (Panel B), and Word Attack (Panel C) Raw Scores (Grades 1 and 2 Only).*