Exploring the concurrent validity of a linguistically responsive pre-reading assessment in bilingual children: the Urdu Phonological Tele-Assessment (U-PASS) tool

Insiya Bhalloo[1,2]* and Monika Molnar[1,2]

[1]Department of Speech-Language Pathology, University of Toronto, Toronto, ON, Canada, [2]Faculty of Medicine, Rehabilitation Sciences Institute, University of Toronto, Toronto, ON, Canada

Introduction: Childhood language and reading contribute to academic and socio-economic success. It is therefore important to identify school-aged children's language and reading abilities as early as possible to provide additional support. Oral-language assessment tools are used by speech-language pathologists (SLPs) and educators for early identification of children at risk for reading difficulties. Particularly, phonological processing is an oral-language and pre-reading skill commonly used to examine reading skills at the early grade levels. However, there is a clear English-language assessment bias when it comes to these pre-reading assessment tools, which impacts clinical and research practices. Bilingual children are commonly misidentified with reading disorders—partially due to the lack of appropriate assessment tools. To address this bias, we developed and evaluated an age- and linguistically-appropriate Urdu Phonological Tele-Assessment (U-PASS) tool across three phases. The U-PASS consists of 10 phonological awareness, phonological memory, and rapid automatized naming subtests.

Methods: We tested 115 typically-developing Urdu-English simultaneous bilinguals in Grades 1–2 across Canada and Pakistan on the U-PASS, vocabulary and reading accuracy measures, and background questionnaires.

Results: Item-level analysis, involving point-biserial correlations, accuracy rates, and Cronbach's alpha, and linear regression, involving subtest- and composite-level analysis, were used to collect evidence for U-PASS's internal reliability and criterion-based concurrent validity.

Discussion: Our linear regression analyses indicated significant phonological processing-reading associations after accounting for background variables, thereby demonstrating U-PASS's concurrent validity for assessing Urdu word/non-word reading. The open-access tool enables SLPs and educators to provide early pre-reading assessment and support for the successful development of oral-language and reading skills in bilingual children speaking Urdu—a common, yet under-investigated, language globally.

KEYWORDS

Urdu, bilingual, phonological processing, tool development, reading, school-aged children

Introduction

Most language assessment tools are developed and normed for monolingual and predominantly English-speaking children. We are lacking tools for a linguistically and culturally diverse population, despite over half of the global population being bilingual–many of whom attend schools in North American countries (Grosjean, 2010). As classrooms globally become linguistically diverse, it is crucial to provide linguistically responsive language assessments using appropriate heritage language tools. To facilitate comprehensive evaluation and to prevent unfair comparison to monolingual norms, we developed an oral-language assessment tool suitable for Urdu-English bilingual children. Specifically, in this tool, we focused on phonological processing skills. These skills are commonly used by educators and speech-language pathologists (SLPs) to predict early reading skills. It has been highlighted that there is a "pressing need" to develop a valid and reliable Urdu phonological measure (Ambreen and To, 2021), due to Urdu's global prevalence (Eberhard et al., 2025) and the broader lack of linguistically responsive assessment tools to support bilinguals in their dominant heritage language (Rose et al., 2022).

Phonological processing, a linguistic skill developing during the preschool age, predicts future word decoding[1], reading accuracy, and fluency (Landerl et al., 2019). These skills are directly related to general reading ability. Globally, reading difficulties are common in children, with 56% unable to meet age appropriate reading levels (UNESCO, 2017). Because delayed identification of reading difficulties has long-term academic, health, and socio-economic ramifications (Ritchie and Bates, 2013), early assessment of pre-reading skills–such as phonological processing–by educators, reading specialists, and SLPs is common at the kindergarten–Grade 2 levels to identify and support reading abilities.

Reading assessments and language development

Learning to read is an important part of language development. Before school age, spoken language is the main form of linguistic input. This changes by Grade 3, when children primarily learn to use language and acquire knowledge via textual information sources, including print and digital media. Through reading, children are exposed to increasingly complex language forms and structures, including vocabulary, syntax, and grammar–thereby enhancing their ability to use language for comprehension and production. As also demonstrated by brain imaging studies, reading directly interacts with language development (Buchweitz, 2016; Fletcher et al., 2000; Rüsseler et al., 2021). Children and adults with reading deficits demonstrate difficulties with language-associated skills, such as verbal fluency, attention, perception, verbal working memory, and spoken information processing (van Linden and Cremers, 2008). As such, developing reading skills at an early age is crucial for preventing future language deficits, particularly as reading difficulties may form a barrier to acquiring language

via print sources. As discussed below, phonological processing assessments play a major role in supporting successful development of reading and, in turn, language abilities in young children at the kindergarten–Grade 2 levels (Caravolas et al., 2012).

Phonological processing assessments

Phonological processing (an oral language skill) is a reading ability precursor, across diverse alphabetic orthographies including English and Urdu (Landerl et al., 2019; Mirza et al., 2017). Phonological processing involves identifying and using speech-sounds present in a language. The oral-language skill consists of three correlated, but distinct, components: phonological awareness, phonological memory, and rapid automatized naming (RAN; Landerl et al., 2019). Phonological awareness is a meta-linguistic skill that involves identification and manipulation of phonological grain units, such as whole words, onset-rimes, syllables, and phonemes (i.e., phonemic awareness), in spoken words (Landerl et al., 2019); it is commonly assessed via word blending and segmentation tasks. Emerging readers, at kindergarten and Grade 1–2 levels (i.e., 4–7 years), extend their knowledge of phonological units when acquiring and forming associations between corresponding graphemes. They identify and sequence learned symbol (i.e., grapheme)-speech sound (i.e., phoneme) correspondences to decode presented words in alphabetic languages. Phonological memory involves temporarily storing and recalling verbal information, such as novel words or digits, from working or short-term memory (STM). Phonological memory deficits limit children's ability to learn novel vocabulary or decode unfamiliar words (Caravolas et al., 2012). RAN is the ability to fluently process phonological information, associated with written or visual information, encoded in long-term memory. This includes efficiently retrieving phonemes associated with written letters during decoding (Caravolas et al., 2012). Together, the three components represent phonological processing abilities. They have been reliably used for early assessment/identification of word-level reading performance and for support of pre-reading and reading abilities in young bilingual children (Landerl et al., 2019).

Phonological processing skills enable children to recognize and manipulate spoken sound structures, which are later associated with written letters–thereby facilitating and predicting reading development in both monolingual and bilingual children including Urdu speakers/readers (Caravolas et al., 2012; Landerl et al., 2019; Mirza et al., 2017). In addition, weak childhood phonological skills are linked to future reading difficulties, while children with strong phonological processing skills demonstrate higher early word-reading scores (Caravolas et al., 2012). SLPs and educators commonly use various types of tests developed for assessment of phonological processing, including the Comprehensive Test of Phonological Processing (CTOPP-2; Wagner et al., 2013). These assessments are used with both English-speaking monolinguals and bilinguals to examine reading abilities at the kindergarten and Grade 1–2 levels, and identify children requiring additional support. Existing assessments, however, are primarily developed for English monolingual children. Evaluating bilingual children – including Urdu-English bilinguals–solely on direct translations of standardized English assessment tools and norms does not constitute evidence-based practice (Rose et al., 2022). This is because direct translations do not consider differences in

---

1 The ability to identify and blend sequences of grapheme (symbol) - phoneme (sound) associations to articulate/read a written word in alphabetic or semi-alphabetic orthographies.

phonological and orthographic properties and age appropriate word frequency.

Developing linguistically responsive pre-reading assessment tools

Bilingual children are disproportionately under-identified for oral-language and pre-reading deficits. This is, in part, due to delayed assessment because of limited assessment tools in the dominant heritage language. Additionally, pre-reading deficits may be misidentified as language proficiency differences, due to limited awareness of typical and atypical norms across linguistically diverse populations (Rose et al., 2022). Given the lack of heritage language pre-reading and reading assessment tools, bilinguals are often only assessed in English using tools standardized for English monolinguals or via direct translations of existing English measures (Arias and Friberg, 2016; Gillam et al., 2013; Hemsley et al., 2014; Rose et al., 2022; Teoh et al., 2018). This approach results in delayed assessments until the child gains sufficient English proficiency. As well, when bilingual children are only assessed in English, it is difficult to determine if reading deficits are due to broader domain-general cognitive or language-related disorders (i.e., if present in both languages), or due to language proficiency differences (i.e., if only present in less dominant language). The lack of appropriate heritage language tools also influences assessment practices in research and understandings of bilingual language and reading development. Such an English-language assessment bias limits our ability to examine performance and age-based milestones on oral language tasks in bilingual and non-English speaking children (Goldstein and Fabiano, 2007; Williams and McLeod, 2012).

Globally, Urdu is the 11th most spoken language, with over 246 million speakers—many living in English-speaking countries like Australia, Canada, the United States, and United Kingdom (Eberhard et al., 2025). Despite global prevalence, Urdu is under-represented in child development research and, in turn, clinical practices–which necessitated the Urdu Phonological Tele-Assessment (U-PASS) tool. As further discussed in the current paper, our developed open-access U-PASS tool is a comprehensive measure of Urdu phonological processing, with established concurrent validity for assessing Urdu word/non-word reading. While some prior studies have demonstrated the importance of Urdu phonological processing as an indicator of children's Urdu reading skills (Mirza and Gottardo, 2022, 2023; Mirza et al., 2017), they only assessed a component of phoneme-level phonological awareness, sound elision, via a 10-item task. These studies did not comprehensively examine children's Urdu phonological processing skills, including phonological awareness, phonological memory, and RAN components, in relation to their reading abilities (Mirza and Gottardo, 2022, 2023; Mirza et al., 2017). Further, existing Urdu phonological processing assessments tools are limited. They primarily focus on rhyming skill, as compared to syllable and phoneme-level phonological grain units. However, rhyming is not an effective reading predictor in emerging readers (Macmillan, 2002). Currently available Urdu measures do not outline their tool development processes and consideration of language-specific phonological features and lexical items. They have not been validated in terms of internal reliability and external criterion-based concurrent validity, and have only been used for research purposes (Mirza et al., 2017; Mumtaz and Humphreys, 2001).

Theoretical framework and structure of the U-PASS

Our work follows tool development phases outlined in prior literature (e.g., Flowers et al., 2015; Hueniken et al., 2020) to ensure that the U-PASS is culturally- and linguistically-appropriate for Urdu-learning children. In line with tool validation frameworks (Bachman and Palmer, 2010; Chapelle, 2020; Weir, 2005), we developed and evaluated the U-PASS in a tele-assessment[2] format, and ensured in-person and online assessment options, to facilitate equitable access. Based on Open Science, we report and share all the details of the three main tool development and validity phases (Phase 1: Item and Subtest Structure Development, Phase 2: Item Revisions and Problematic Item Reconciliation, and Phase 3: Tool Evaluation and Item Tryouts); details that go beyond standard reporting requirements can be found in supplementary files S1–S40 and supplementary figures 1, 2 can be accessed via the Borealis Dataverse Research Data Repository: https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/BRFIXM.

We relied on commonly-used tool validation frameworks used in language assessment. These include the Assessment Use Argument (AUA) Approach (Bachman and Palmer, 2010), Argument-based Validation Approach (Chapelle, 2020), and Socio-Cognitive Framework (Weir, 2005). The three language assessment frameworks emphasize the importance of considering various factors during tool development and validation, including the (i) type of assessed language construct and its linguistic properties, (ii) type of assessment measure and the level of tool development or adaptation required, (iii) assessed population and their characteristics, including language proficiency, (iv) involvement of knowledge stakeholders in tool development and use, and (v) overall consistency/reliability and validity of the developed language assessment measure (Bachman and Palmer, 2010; Chapelle, 2020; Weir, 2005). Our consideration of these factors across the three phases of tool development is highlighted in the Methods section.

In addition, like most phonological processing tools (e.g., Wagner et al., 2013), our novel U-PASS tool is based on a model of reading-related phonological processing–as shown in Figure 1 (derived from Mitchell, 2001; Wagner and McBride-Chang, 1996). For example, the Comprehensive Test of Phonological Processing (CTOPP-2; Wagner et al., 2013), a widely-used English standardized phonological processing measure, is significantly correlated ($r = 0.73$) with word/non-word reading accuracy tests such as the Woodcock Reading Mastery Test (WRMT-III; Woodcock, 2011)—independent of other reading precursors and language proficiency. In line with tool validation frameworks for language assessment (Bachman and Palmer, 2010; Weir, 2005), we also considered the type of assessed language construct (i.e., phonological processing) and its linguistic properties (including comprehensively assessing its three component skills and analysis of phonological grain units at the onset-rime, syllable, and phoneme levels).

The U-PASS includes the 10 subtests listed below, based on the three phonological processing constructs (see S1 for description):

---

2   The use of technology to facilitate service delivery (Wood et al., 2021).

1. Phonological Awareness: Sound Elision, Word Blending, Word-Initial and Final Sound Matching, and Non-Word Blending subtests.
2. Phonological Memory: Non-Word Repetition, and Recall for Digits subtests.
3. Rapid Automatized Naming (RAN): Rapid Digit Naming, Rapid Letter Naming, Rapid Colour Naming, and Rapid Object Naming subtests.

U-PASS: reliability and validity

We examined U-PASS's internal structure, via item-level analysis and its relationship with external outcome measures (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). We chose to examine evidence for U-PASS's internal reliability and external aspects of validity at the subtest and composite levels based on tool validation frameworks (Chapelle, 2020; Weir, 2005) and prior tool validation studies (Wagner et al., 2013). To evaluate U-PASS's internal reliability, we examined item discrimination power, accuracy rates, and reliability. We conducted point-biserial correlations (r_{pb}) to examine each item's discrimination power/reliability. Item discrimination is the degree to which an individual item differentiates amongst high and low performers on a tool. This analysis examines correlations between an item and the associated subtest, and the degree to which they assess the same ability (Anastasi and Urbina, 1997). We also examined item accuracy rates (i.e., percentage of children who responded correctly to a given item) within subtests 1, 2, 3, 4, 10. This enabled us to identify and exclude items that were not age appropriate (i.e., too easy, or difficult) and re-organize item ordering within subtests to ensure progressive difficulty. We calculated Cronbach's alpha per subtest to examine each subtest's internal reliability/consistency and the amount of shared variance amongst items comprising each subtest.

The aim of the study

We examined U-PASS's criterion-based concurrent validity to collect evidence for external validity at the subtest (i.e., subtests 1–10) and composite (i.e., aggregate across subtests 1–10) levels. Criterion-based concurrent validity is the degree to which a tool effectively predicts an intended criterion outcome, when both measures are assessed at the same timepoint (McIntire and Miller, 2005). We examined associations for the U-PASS tool, at subtest and composite levels, in relation to word/non-word reading accuracy outcomes and external background variables of Grade 1–2 Urdu-English bilingual children.

As a response to the English-centric bias in assessment tools, we developed and explored the initial validity of an age appropriate and linguistically responsive Urdu Phonological Tele-Assessment (U-PASS) tool for Urdu-speaking monolingual and bilingual children. Specifically, we evaluated and explored evidence for U-PASS's internal reliability and initial criterion-based concurrent validity. We assessed phonological processing, as measured by U-PASS, at the same timepoint as the Urdu reading criterion outcome.

Research questions

In this paper, we (i) describe the development and (ii) explore the validity of an age appropriate and linguistically responsive (i.e., based on phonological and orthographic features, and phonotactics) Urdu Phonological Tele-Assessment (U-PASS) tool, to support its use with Urdu-English bilingual children.

To collect evidence for U-PASS's internal reliability and external validity, we address the following research questions:

(i) Item-level evidence based on internal reliability

- Do our item-specific analysis and revisions, based on (a.) discrimination power and (b.) accuracy rates per subtest, support the U-PASS's internal reliability?

(ii) Subtest- and composite-level evidence based on criterion-based concurrent validity and relations to other external variables

- Do Urdu-English bilinguals who score higher on the individual 10 U-PASS subtests demonstrate higher word/non-word reading outcomes?
- Do Urdu-English bilinguals who score higher on the composite U-PASS tool demonstrate higher scores on the word/non-word reading external criterion outcome, after accounting for other background variables including age at assessment, grade-level at assessment, gender, Urdu language proficiency (i.e., language exposure and usage ratings), age of acquisition (AoA) and parental socio-economic status (SES)?

Materials and methods

Phase 1: item and subtest structure development

Figure 2 outlines the three development phases of U-PASS in line with prior tool adaptation and validation papers in speech-language pathology (SLP), education, and health sectors (Flowers et al., 2015; Hueniken et al., 2020; Mueller Gathercole et al., 2008) and tool validation frameworks (Bachman and Palmer, 2010; Chapelle, 2020; Weir, 2005). These three phases include: Item and Subtest Structure Development (Phase 1), Item Revisions and Problematic Item Reconciliation (Phase 2), and Tool Evaluation and Item Tryouts (Phase 3). Before constructing the tool, we reviewed common English phonological processing measures to select the 10 subtests included in U-PASS (see Figure 2 for reviewed measures). In determining subtests, we considered type of assessed phonological processing construct (i.e., phonological awareness, phonological memory, or RAN), phonological grain unit (i.e., word, syllable, onset-rime, or phoneme), and medium of assessment (i.e., auditory, visual, or processing speed) in relation to the intended age-group.

We selected age- and linguistically/culturally appropriate items for inclusion, in accordance with subtest goals, based on existing Urdu language resources and teacher experiences with Urdu language and reading instruction materials in language classes (see Figure 2 for examples). For the 8/10 subtests involving real words,
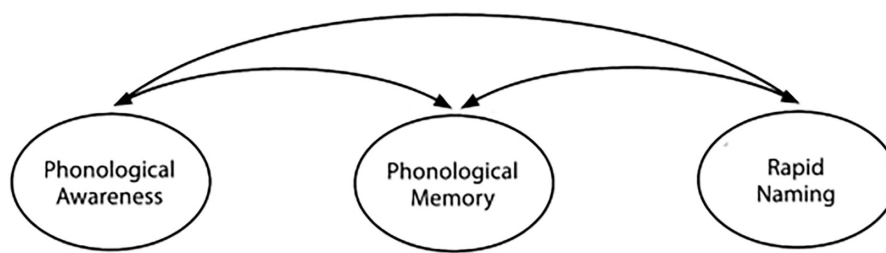
FIGURE 1
Model of reading-related phonological processing (Mitchell, 2001). Model illustrating the three interrelated, yet distinct, phonological processing constructs: phonological awareness, phonological memory, and rapid.

lexical items were selected from existing Urdu resources, including a MacArthur-Bates Communicative Development Inventory (Dar et al., 2015) and Receptive/Expressive Vocabulary Tools (Kandru-Pothineni et al., 2015; Warmington, 2015). For the 2/10 subtests with non-words, such as Non-Word Blending and Non-Word Repetition, existing words were selected from storybooks and adapted to form non-words by replacing phonemes or substituting their word-initial, medial, or final positions. We included enough items to allow for later item exclusion and replacement (N = 396).

To verify item appropriateness, two Urdu primary-school teachers from Child and Adolescent Development Programme (CADP)-affiliated community schools in Pakistan and three Urdu research MSc/PhD students from McGill University (Canada), Wilfred Laurier University (Canada), and Columbia University (United States) independently evaluated items for age- and linguistic-appropriateness and subtest relevance, and provided alternatives based on existing Urdu child language/reading materials and teaching experience. In line with tool validation frameworks (Bachman and Palmer, 2010; Chapelle, 2020), we consulted these knowledge stakeholders during our U-PASS tool development (see Phases 1 and 2) and usage/validation (see Phase 3) phases. We also considered the type of assessment measure, and the level of tool development or adaptation required to develop a linguistically responsive Urdu phonological processing measure (Bachman and Palmer, 2010; Chapelle, 2020). Due to the lack of Urdu measures, and linguistic differences between Urdu and English, we considered the subtest-level structure of existing standardized English phonological processing measures but could not directly adapt their content for the U-PASS. As such, our included items were not direct translations of existing English tools to ensure items are culturally appropriate and linguistically meaningful by taking under account Urdu's phonemic inventory and syllable structure. In addition to consulting these experts, we established progressive item difficulty within subtests based on type and number of phonemes, syllable count, and word length.

Phase 2: item revisions and problematic item reconciliation

2A. Item verification and problematic item identification

We informed seven evaluators of assessment goals for each subtest. These knowledge stakeholders/users were different from those consulted in Phase 1, to ensure comprehensive evaluation (Bachman and Palmer, 2010). Three Urdu linguists and four primary-level Urdu teachers, from universities and community schools affiliated with Aga Khan University and CADP in Pakistan, identified and replaced problematic items. They were instructed to review items and tasks, for age, linguistic and cultural appropriateness, and any additional identified criteria; and provide alternatives for problematic items (see S2 for items and reasons for revisions per subtest). Then, we categorized evaluator comments based on five common criteria in linguistic adaptation research (Flowers et al., 2015; Mueller Gathercole et al., 2008). These Urdu experts identified 151/396 problematic items, with 70% agreement between the seven evaluators, similar to other tool development (Flowers et al., 2015; Hueniken et al., 2020). See S3 for description of the five criteria:

1. Conceptual Equivalence
2. Age Appropriateness/Common Language Use
3. Clarity
4. Item and Subtest Difficulty
5. Linguistic Appropriateness

2B. Problematic item revision via expert consensus

All problematic items identified by the seven evaluators were replaced based on one or more of the five criteria. In case of disagreement regarding alternative item revisions/recommendations, we reconsulted these Urdu experts, and verified revisions with one Urdu translator and two early childhood education graduates in Pakistan. See S2 and S4 for total number and reasons for item revisions per subtest based on the five criteria. Below, we present some examples of replacements based on the 5 criteria:

1. **Conceptual Equivalence:** Sound Elision items 16A.2 and 16B.2, /qə.dəm/ "footsteps" and /qə.d/ "height," were replaced with /pa.ni/ "water" and /pan/ "betel leaves". This is because the formed Urdu real word required deletion of two word-final phonemes. In the Initial Sound Matching subtest, item 5B.1 /bʰa.lu/ "bear" was replaced with /ba.zu/ "arm", as it shared two final phoneme sounds with the target /ä.lu/ "potato".

2. **Age Appropriateness:** In the Sound Elision subtest, items 22A.2 and 22B.2, /d̄ʰɔl/ "drum" and /d̄ɔl/ "bucket", were
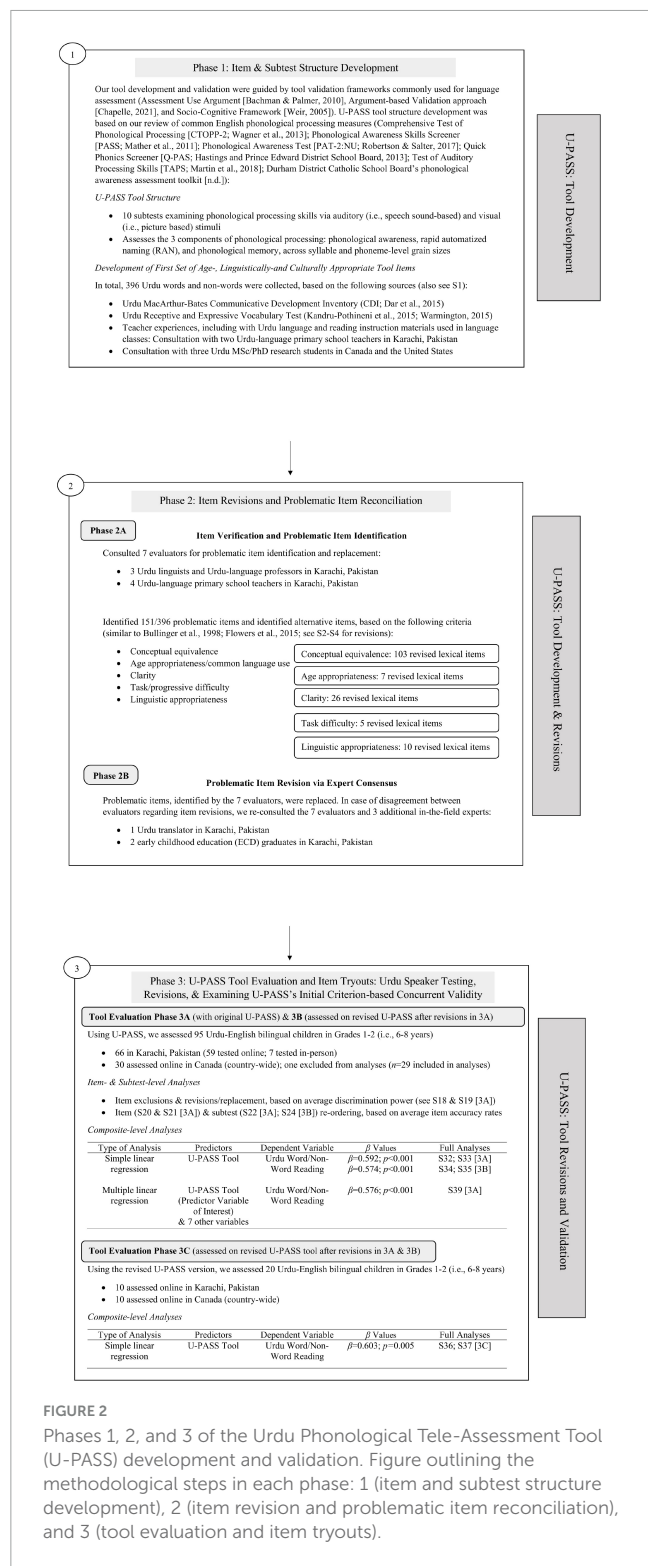
FIGURE 2
Phases 1, 2, and 3 of the Urdu Phonological Tele-Assessment Tool (U-PASS) development and validation. Figure outlining the methodological steps in each phase: 1 (item and subtest structure development), 2 (item revision and problematic item reconciliation), and 3 (tool evaluation and item tryouts).

replaced with /dada/ "paternal grandfather" and /dad/ "to clap/praise". In Final Sound Matching, item 15D.2 /kʰɑɪ.bu.zəʔ/ "cantaloupe" was replaced with / kʰətʷ/ "letter".

3. **Clarity:** In the Final Sound Matching subtest, item 15B.2 /ba.ɹɪʃ/ "rain" was replaced with /kɪʃ.mɪʃ/ "raisin" for pictorial clarity. In Word Blending, item 15.3 /ä.nkʰ/ "eye" was replaced with /hɔntˀ/ "lips" for auditory clarity, due to an

extra-long vocalic phoneme and phoneme digraph. In Recall for Digits and Rapid Digit Naming subtests, the Naqsh script ۴ "four"[t͡ʃaɪ] was replaced with the Nastaliq script ۴ for orthographic clarity.

4. **Item and Subtest Difficulty:** Initial Sound Matching item 10A.1 /dant/ "teeth" was replaced with /ɹɔ̄.ti/ "flatbread". In the Final Sound Matching subtest, item 10B.2 /seb/ "apple" was replaced with /t͡ʃi.ku/ "chikko".

5. **Linguistic Appropriateness:** Non-Word Repetition item 12.1 /wæ.nʌk/ was replaced with /ke.nʌk/. In the Non-Word Blending subtest, item 20.2, /zəʃ.mə/ was replaced with /ɹəʃ.mə/.

Phase 3: Tool evaluation and item tryouts — Urdu speaker testing, revisions, and examining U-PASS's initial criterion-based concurrent validity

Participants

Data collection was approved by University of Toronto's Research Ethics Board (Protocol 38608). Due to online testing because of COVID-19, we could only assess children with reliable internet access; these were typically from middle-high SES backgrounds. The non-random sample of participants in Canada and Pakistan were exposed to Urdu and English within the home prior to formal schooling (i.e., AoA before age 3), and therefore simultaneous bilinguals. This is similar to language profiles of Urdu-speaking bilinguals in diaspora communities globally. Due to language/demographic background similarities, we merged Urdu-English bilinguals across Canada and Pakistan in analyses. As the test-taker's profile has been shown to influence tool validation outcomes and performance scores (see Bachman and Palmer, 2010), it was important for us to collect information on our participants' language and demographic background and examine these factors in our analyses reported below.

The simultaneous bilinguals assessed in Pakistan and Canada had attended schools where English was the primary language of instruction and had been exposed to both Urdu and English before the age of three. However, in contrast to the Urdu-English bilinguals in Canada, those in Pakistan experienced greater exposure to formal Urdu language and reading instruction through their educational and broader societal environments. This is due to Urdu's official and societal-language status in Pakistan. In comparison, the Canadian bilinguals were typically exposed to Urdu at home as a heritage language or in supplementary language classes. Additionally, the educational systems in the two countries differ. Children in Pakistan are introduced to formal schooling and reading instruction at an earlier age and receive formal Urdu literacy instruction within the school system (Government of Pakistan, 2022; Government of Pakistan, 2021; Ontario Curriculum and Resources, 2023).

Inclusion and Exclusion Criteria. We assessed Urdu-English simultaneous bilinguals across Canada and Pakistan who: were in Grades 1–2 (i.e., 6–8 years), spoke Urdu as home/heritage language and English as school/societal language, and had no reported history of language, reading or cognitive difficulties.

Here, we define simultaneous bilinguals as children who were simultaneously exposed to both languages before the age of 3, prior to learning to read in kindergarten (Babatsouli, 2024; Patterson, 2002). Simultaneous bilingual status was verified via parent reported AoA and language exposure/usage ratings on the Child Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). By assessing simultaneous bilinguals, we ensured that each child received sufficient exposure to English and Urdu to ensure their responses to the assessment measures were meaningful. It was important to ensure this due to heterogeneity in bilingual language profiles (Babatsouli, 2024).

Demographic Information. A total of 115 simultaneous bilingual participants were tested and analyzed across the three phases (3A, 3B, 3C) of tool evaluation in Phase 3. In Tool Evaluation Phases 3A and 3B (i.e., assessment and analysis using in-progress U-PASS tool), 88/95 participants were tested online, due to research transitions during COVID-19. We did not separate these participants in analyses, as scatterplots demonstrated similar performance (see S5–S7 for scatterplots). In Phase 3C (i.e., assessment and analysis using revised U-PASS tool, after revisions in Phases 3A and 3B), all 20 participants were tested online. We merged participants across assessment country, due to similarities in language and SES background across Phases 3A, 3B (see S8 for reported background), and 3C (see S9 for reported background). While some participants could be described as multilingual, we use the term *bilingual* given that they were the most proficient in two languages (i.e., Urdu and English).

Our statistical analysis included data from 95 Urdu-English simultaneous bilinguals (48 females [Average age: 6.06 years; Range: 6.01–8.02 years of age; Standard Deviation: 0.51 years]; 47 males [Average age: 6.08 years; Range: 6.01–8.04 years of age; Standard Deviation: 0.52 years]) across Canada ($n = 29$) and Pakistan ($n = 66$) in Phases 3A and 3B. Phase 3C included 20 participants (10 females [Average age: 6.06 years; Range: 6.0–8.04 years of age; Standard Deviation: 0.50 years]; 10 males [Average age: 6.07 years; Range: 5.11–7.09 years of age; Standard Deviation: 0.47 years]) across Canada ($n = 10$) and Pakistan ($n = 10$). One Canadian child in Phase 3A was excluded from analysis due to incomplete measures. The participants in Phase 3C were different from those assessed in prior tool evaluation phases; we did so to verify criterion-based concurrent validity after item and subtest-level revisions in Tool Evaluation Phases 3A and 3B.

Data collection procedure

Parents who consented to child participation completed two questionnaires to ensure participants met inclusion criteria (S10 indicates Bilingual Language and Demographic Background Questionnaires). This ensured that participants met age (i.e., aged 6–8 years and in Grades 1–2), typical development (i.e., no reported language, reading, or cognitive difficulties), and Urdu and English simultaneous bilingual language status inclusion criteria. In addition to parental consent forms, we obtained verbal assent from children and verified sound quality and audibility prior to sessions. Testers also monitored the children's internet connection (as indicated by the Zoom network connection indicator) to ensure children could adequately access presented visual and auditory stimuli. We trained testers from related educational fields such as linguistics, psychology, education, and speech-language pathology. All our testers in Pakistan and Canada completed training and

testing simulations prior to conducting assessments, and could read and write Urdu with a high degree of proficiency. Our trained testers informed parents about testing procedures beforehand, including the importance of access to a quiet testing space with no distractions or additional children, using headphones, and parents being silent observers who help ensure a stable internet connection but refrain from prompting their child. We also followed established tele-assessment guidelines for oral-language and literacy assessments (Wood et al., 2021).

In line with tele-assessment recommendations, we adapted the Urdu assessment scoring and stimuli presentation documents for Zoom administration (Wood et al., 2021). Each child completed Urdu measures across three 1-hour sessions in the following manner: Session 1: expressive vocabulary (30 min) and word/non-word reading accuracy (30 min); Session 2: U-PASS part A (Sound Elision, Sound Matching, and Word Blending subtests; 60 min]; and Session 3: U-PASS part B (Phonological Memory, RAN, and Non-Word Blending subtests; 60 min). The three sessions were conducted on separate days within the same week. See S11 for Urdu Oral Language and Reading Measures.

Measures

Bilingual Language and Demographic Background Questionnaires.

Child Language Experience and Proficiency Questionnaire (Child LEAP-Q). The Child LEAP-Q (Marian et al., 2007) is a parental report of child language background used by researchers across multiple disciplines, including SLP and education. Our analysis included the following background variables: Age at assessment, grade-level at assessment, gender, Urdu language proficiency (i.e., language exposure and usage ratings), and AoA in years. For average language proficiency, parents indicated cumulative language exposure and usage rates across both languages (i.e., the total percentage of indicated Urdu and English exposure must be 100%, and the total percentage of indicated Urdu and English usage must be 100%). We examined language background as it influences reading development (Luft Baker et al., 2021).

Demographic Background Questionnaire. The adapted parental self-report includes demographic background questions such as birth-place, and/or age at immigration, and SES (see S10; Mirza et al., 2017). Our analysis included the parental SES (i.e., parental education and child school neighborhood/tuition rates) background variable, due to its influences on emergent language and reading outcomes beyond bilingual language experiences (Lundberg et al., 2012).

Urdu Oral Language and Reading Measures.

U-PASS Tool (Under Development/Investigation). See S1 for 10 subtests.

Lexical Expressive Vocabulary. We adapted an Urdu expressive vocabulary test (Mirza et al., 2017) to ensure linguistic and cultural-appropriateness across Canadian and Pakistani contexts. Children named 73 individually-presented pictures.

Word and Non-Word Reading. The adapted Urdu word and non-word reading measure consists of 72 Urdu words and 30 Urdu non-words, and follows progressive difficulty and stimuli presentation guidelines of Word Identification and Word Attack subtests from Woodcock Reading Mastery Test (WRMT; Woodcock, 2011). The Urdu reading measure was adapted from an

Urdu measure used in prior research with Urdu-English bilingual children (Mirza et al., 2017). Children were presented with 6–9 lexical items at a time, which they read aloud.

Methods of analysis

We consulted statistical specialists at University of Toronto for analyses. Our analyses are similar to those conducted in tool validation of existing phonological processing, oral-language, and reading assessment tools in other languages (e.g., CTOPP-2; Wagner et al., 2013; Chinese Character Acquisition Assessment: Chan et al., 2020; HABE C1 Basque Competence Test: Elosua, 2024; American English Sociopragmatic Comprehension Test: Timpe-Laughlin and Choi, 2017). In Phase 3, we evaluated the in-progress U-PASS tool with Urdu-English bilinguals to examine child performance, conduct further tool revisions based on performance, and examine the internal reliability and external aspects of validity across the 3 phases (3A, 3B, 3C) of U-PASS tool evaluation. To examine U-PASS's internal reliability and determine whether further tool improvements were needed, we conducted (i) *item-specific analysis*, examining average discrimination power and item accuracy rates across the 10 subtests. We calculated Cronbach's alpha per subtest as an indicator of U-PASS's internal consistency/reliability.

To examine U-PASS tool's initial criterion-based concurrent validity in Phase 3, we conducted: (i) *subtest-level analysis*, for U-PASS subtest 1–10 correlations in relation to Urdu word/non-word reading accuracy, and (ii) *composite-level analysis* by analyzing associations between Urdu composite phonological processing (i.e., aggregate score across subtests 1–10 as assessed by U-PASS) and word/non-word reading accuracy outcomes. Phonological processing abilities highly correlate with word and non-word reading performance, including with CTOPP-2 (Wagner et al., 2013), a widely-used English standardized phonological processing measure. Therefore, to examine U-PASS's initial criterion-based concurrent validity, we assessed participants' Urdu phonological processing in relation to their word/non-word reading skills, along with language and demographic background variables commonly associated with reading.

Results

Phase 3—Tool evaluation and investigating U-PASS's internal reliability and external aspects of validity

Tool Development Phase 3 consisted of Phases 3A, 3B, and 3C. In Phase 3A, we assessed and analyzed 95 Urdu-English bilinguals' performance on the in-progress U-PASS tool, before item and subtest-level analysis and revisions in Phases 3B and 3C. In Phase 3B, we examined these 95 bilinguals' performance on the revised U-PASS tool, based on subtest and item-level reordering and exclusions/replacements in Phase 3A. In Phase 3C, we assessed and analyzed an additional 20 bilinguals' performance on revised U-PASS, based on modifications in Phases 3A and 3B. Pearson's correlation coefficients ranging between $r = 0.4$–$r = 0.6$ were

categorized as moderate, and those ranging between $r = 0.7$–$r = 0.9$ were categorized as strong. The Results section discusses tool evaluation in Phase 3 and item, subtest, and composite-level analyses that we conducted for tool revision and to provide evidence for U-PASS's internal reliability and initial criterion-based concurrent validity.

Internal reliability: item-level analysis

Item Exclusions (based on Average r_{pb} Coefficients). In Phase 3A, we conducted corrected point-biserial correlation analyses to identify problematic items for exclusion based on average discrimination power (r_{pb} coefficients). This type of analysis examines correlations between an individual item and the associated subtest, thereby allowing us to investigate the U-PASS's internal reliability. In line with prior tool validation studies (e.g., Wagner et al., 2013; Nunnally and Bernstein, 1994), we considered an average r_{pb} coefficient (r_{pb}) ≥ 0.30 for subtests, and minimum item-specific r_{pb} coefficients ≥ 0.15–0.20[3]. Similar to standardized English phonological measures such as the CTOPP-2 (e.g., Wagner et al., 2013), U-PASS subtests 1, 2, 3, 4, 5 and 10 demonstrate an average $r_{pb} \geq 0.30$ in Phase 3A (see S12–S13 for average item r_{pb} distribution, Cronbach's alpha per subtest, and forest plot), with similar average $r_{pb} \geq 0.30$ in Phases 3B (see S14–S15) and 3C (see S16–S17). This indicates that the individual test item and associated subtest demonstrate positive correlations and measure the same phonological construct. Our results provide evidence for U-PASS's internal reliability at the item level in relation to the associated subtest. S12–S17 indicate percentage of items with problematic r_{pb} coefficients, including negative r_{pb}, $r_{pb} = 0.00$, and $r_{pb} < 0.20$.

As evident in S18 and S19, we identified 30/214 items (14%) with problematic r_{pb} values across U-PASS subtests 1, 2, 3, 4, 5 and 10, and excluded 3/214 items (1.40%) for psychometric reasons due to problematic r_{pb} values in Phase 3A. Excluded items include those with r_{pb} values that are negative[4], ceiling-level accuracy (0.00)[5], or < 0.20[6]. We retained 27/214 (12.62%) age- and linguistically-appropriate practice items with problematic r_{pb} values, due to potential item-order effects (see S18 for percentage of items retained per subtest). Items with r_{pb} values of 0.00 (i.e., ceiling-level performance) were retained as practice items, thereby reflecting subtest progressive difficulty. See S19 for item exclusion reasons across subtests based on r_{pb} coefficients and average discrimination power per item.

Item Ordering (based on Average Item Accuracy). To ensure U-PASS's progressive difficulty, we revised ordering of 185/214 items (86.45%) based on average accuracy rates (i.e., percentage of children who responded correctly to a given item on a subtest), across subtests 1, 2, 3, 4, 5, and 10 (see S20 for item ordering revisions and exclusions) in Phase 3A. Overall, 1/214 items (0.47%)

---

3 Tool items with $r_{pb} \geq 0.20$ have good discrimination power for differentiating between high and low performers on a given subtest. High subtest performers score correctly on these items, while low subtest performers score incorrectly.

4 Items with poor discrimination power. High subtest performers score incorrectly on these items, while low subtest performers score correctly.

5 Items with poor discrimination power, in which all tested participants either score correctly (ceiling effects) or incorrectly (flooring effects).

6 Items with poor discrimination power, which do not meet recommended r_{pb} values (i.e., between $r_{pb} = 0.20$–0.30).

were excluded, as average accuracy rates did not reflect item construct properties and intended difficulty levels. S20–S21 indicate reasons for item ordering revisions and exclusions based on average accuracy rates per item; S22–S23 indicate average item accuracy rates in Phase 3A across U-PASS subtests. The item accuracy rates per subtest in Phases 3B (see S24–S25 for accuracy rates) and 3C (see S26–S27 for accuracy rates) reflect item accuracy rates and proposed subtest ordering in Phase 3A (see S22–S23 for accuracy rates). Across the three phases, our Cronbach's alpha was ≥ 0.70, which indicates a high internal consistency/reliability (Nunnally, 1978; Lance et al., 2006; see S12, S14, S16).

U-PASS Administration: Proposed Subtest 1 – 10 Ordering (based on Average Item Accuracy). During U-PASS Tool Evaluation Phase 3A, we administered Urdu subtests in the same order as commonly-used English measures such as CTOPP-2, due to the lack of Urdu measures. Prior Urdu measures (e.g., Mirza et al., 2017) have not been designed based on subtest and item-level performance. S22 indicates proposed subtest ranking (from easiest to difficult), based on mean item accuracy rates; see S23 for average item accuracy rate distribution across subtests. We revised subtest ordering to ensure easy-difficulty progressive difficulty across U-PASS Tool Evaluation Phases 3B and 3C, based on average item accuracy rates and revisions in Phase 3A (see S22 for ordering). The U-PASS tool in the final Tool Evaluation Phase 3C reflects this revised subtest ordering. S28 further highlights subtest and item-level analyses, including revisions based on r_{pb} analyses and mean item accuracy rates.

Initial criterion-based concurrent validity: subtest-level analysis

Predictor Coefficients. The 10 predictor coefficients in the simple linear regression correspond to the 10 U-PASS subtests.

To investigate external aspects of validity at subtest level in Phase 3A, we conducted 10 simple linear regressions with pre-specified statistical significance of $p \leq 0.05$, to analyze performance on U-PASS subtests 1–10 in relation to Urdu word/non-word reading (see predictor coefficients listed above). Except for subtest 5 (non-word repetition; phonological STM), all subtests were significantly associated with word/non-word reading (S29 indicates analysis results). Our findings demonstrate initial criterion-based concurrent validity for the 10 individual U-PASS subtests in relation to word/non-word reading. The subtests' relative predictive contributions to word/non-word reading were: Subtests 6 (Digit RAN), 2 (Word Blending), 3 (Sound Matching), 8 (Color RAN), 10 (Non-Word Blending), 1 (Sound Elision), 9 (Object RAN), 7 (Letter RAN), and 4 (Recall for Digits) in Phase 3A (based on greatest to least contribution; see standardized betas in S29). Similar significant or near-significant subtest-level findings were evident in Phases 3B (S30 indicates analysis results) and 3C (S31 indicates analysis results), and reflect sample size ($n = 20$) influences on power.

Initial criterion-based concurrent validity: composite-level analysis

Predictor Coefficients. The eight predictor coefficients in the simple and multiple linear regression include the following background variables: Age at assessment, grade-level at assessment, gender, parental SES (i.e., parental education and child school neighborhood/tuition rates), Urdu language proficiency (i.e., language exposure and usage ratings), AoA, in addition to the expressive vocabulary measure and Urdu composite phonological processing scores (as measured by U-PASS tool under investigation).

To collect evidence for external aspects of validity at the composite-test level, we conducted simple and multiple linear regression across Phases 3A, 3B, 3C. In Phase 3A, the eight simple linear regressions identified four significant predictor coefficients with pre-specified statistical significance of $p \leq 0.05$ (S32 indicates analysis results). Our analyses highlighted that age, grade-level at assessment, Urdu expressive vocabulary, and Urdu composite phonological processing as assessed by U-PASS tool ($p < 0.001$; $\beta = 0.592$ for U-PASS tool in relation to Urdu reading; S33 indicates scatterplot for U-PASS tool and Urdu reading) were significantly associated with Urdu word/non-word reading. Similar significant correlations between U-PASS tool and word/non-word reading were evident in Phases 3B ($p < 0.001$; $\beta = 0.574$; see S34–S35 for analysis results and scatterplot) and 3C ($p = 0.005$; $\beta = 0.603$; see S36–S37 for analysis results and scatterplot).

In Phase 3A, we conducted a multiple linear regression to determine if demonstrated associations between Urdu phonological processing and word/non-word reading remain significant, with inclusion of oral-language and language proficiency/demographic variables shown to influence reading (Lundberg et al., 2012; Luft Baker et al., 2021). To do so, we used a purposeful model selection method with the eight predictor coefficients (listed in Predictor Coefficient subsection) and an adjusted statistical significance of $p \leq 0.006$ after Bonferroni's correction for multiple comparisons. We combined Urdu word and non-word reading as one dependent variable due to strong associations ($r = 0.883$; $p < 0.001$; S38 indicates scatterplot), and to meet model assumptions (i.e., linearity, multivariate normality, no multicollinearity, variable independence, and homoscedasticity). In the multiple regression, Urdu composite phonological processing (assessed by U-PASS) emerged as the only significant predictor for Urdu word/non-word reading (F [8, 86] = 6.859, $p < 0.001$: Urdu phonological processing [$p < 0.001$; $\beta = 0.576$]; S39 indicates analysis results), thereby demonstrating U-PASS's initial criterion-based concurrent validity for word/non-word reading at the composite level.

Discussion

Previous studies have established phonological processing as a reliable reading ability precursor (Caravolas et al., 2012; Wagner et al., 2013). Early pre-reading assessment and support by speech-language pathologists (SLPs) and educators contribute to grade-level reading performance and successful language learning abilities in young monolingual and bilingual children. While the Urdu Phonological Tele-Assessment tool (U-PASS) is an indicator of reading performance, it assesses speech-processing skills via receptive and expressive language-based tasks. Since reading facilitates higher-level language acquisition, the U-PASS can also be used to better understand and support a child's language abilities–via their reading skills.

Below, we discuss U-PASS's item-level internal reliability and tool revisions, subtest and composite-level level criterion-based

concurrent validity, and implications for assessment accessibility in school-aged children who are bilingual in Urdu–a commonly-spoken, yet under-represented language in oral-language and reading research and clinical practice.

Item-level evidence based on internal reliability

To investigate U-PASS's internal consistency/reliability, we calculated Cronbach's alpha for subtests across Phases 3A, 3B, and 3C. Overall, items in the U-PASS subtests demonstrated a high internal consistency/reliability, as indicated by our Cronbach's alpha of ≥ 0.70 (Lance et al., 2006; Nunnally, 1978). We also conducted item-specific revisions based on corrected point-biserial correlational (r_{pb}) analyses and mean item accuracy rates. These modifications increased the U-PASS tool's item discrimination power at composite and subtest levels. Standardized English assessments recommend item-specific r_{pb} values ≥ 0.15–0.20 (e.g., University of Washington, 2018; Wagner et al., 2013). U-PASS demonstrates good internal reliability, as indicated by our pre-determined average r_{pb} subtest values ≥ 0.30; the individual item and the associated subtest demonstrate positive correlations and measure the same phonological construct. Across U-PASS subtests, we included items with r_{pb} values ≥ 0.20 to ensure greater item discrimination power and average r_{pb} subtest values ≥ 0.30. The only exception is Word Blending where our r_{pb} cut-off was ≥ 0.15. For Word Blending subtests with low mean accuracy rates, r_{pb} values are not an effective sole indicator of item quality—due to item guessing or skipping and item-order effects. In Recall for Digits subtest, items with problematic r_{pb} values were not excluded due to its cognitive-load based processing difficulty. The subtest assesses WM and phonological STM based on rapidly-increasing sequences of commonly-known digits 1–9. Similarly, we did not conduct item-level analysis for RAN subtests 6–9, as their test design is based on fluent retrieval rather than item content difficulty or progressive item difficulty. For such subtests with extreme difficulty indexes (i.e., items with very low or high average accuracy rates; Loudon and Macias-Muoz, 2018), it is therefore important to evaluate overall item quality and clinical meaningfulness based on age-relevance, linguistic-appropriateness,assessment goals, and subtest design.

Subtest-and composite-level evidence based on criterion-based concurrent validity and relations to other external variables

Subtest-level performance

Urdu phonological awareness (Sound Elision, Sound Matching, Word and Non-Word Blending subtests) and RAN (subtests 6–9) constructs were consistently associated with reading in U-PASS Tool Evaluation Phases 3A and 3B. Similar significant or near-significant associations are evident in Phase 3C, thus reflecting sample size ($n = 20$) influences on power.

In comparison to English phonological processing tests such as CTOPP-2, U-PASS Sound Elision and Non-Word Repetition subtests demonstrated comparatively higher mean item accuracy. This may be due to Urdu's linguistic structure. Word-final vowel elision to form a real Urdu word is relatively more common, compared to consonant elision in word-initial or final position to form a real English word. As such, U-PASS's Sound Elision subtest involved identification and manipulation of long/consonantal vocalic phonemes in salient word-final positions. Similarly, U-PASS's Non-Word Repetition subtest's high accuracy rate and non-significant reading associations may be linked to Urdu speakers' increased conceptual familiarity with non-words. Mohmil words are the main form of Urdu compound word generation (Jabbar and Iqbal, 2016). A meaningful mohmil compound word, such as /bhiʀ-bhaʀ/ "crowd," results when a bound morpheme is combined with a rhyming non-word typically differing in rhyme or coda components. Comparatively, English compound words consist of two free morphemes.

Specific to Urdu, the Non-Word Repetition subtest was not a sensitive discriminator of age appropriate and struggling readers due to its limited difficulty range and lack of cognitive manipulation of phonological components. The subtest's non-significant associations reflect prior research on phonological STM constructs. In CTOPP-2 validation studies (Wagner et al., 2013), the English phonological STM construct demonstrates weak associations ($r = 0.34$) with English word and non-word reading tests (Wagner et al., 2013). This is in comparison to moderate-level associations for English phonological awareness ($r = 0.73$) and RAN ($r = 0.53$; see Wagner et al., 2013). As such, the Urdu Non-Word Repetition subtest's non-significant associations may be explained by its limited difficulty range and cognitive manipulation of phonological components. We could not conduct further comparative analysis between Urdu and English subtest performance, as English phonological measures report limited subtest and item-level psychometrics. We retained the Non-Word Repetition subtest in the current U-PASS tool battery due to significant composite-level associations. This will enable follow-up analysis in a larger sample across bilingual language background and SES profiles.

Investigating criterion-based concurrent validity at subtest and composite levels

To investigate external aspects of validity at subtest and composite levels, we examined U-PASS tool in relation to word/non-word reading skills as the criterion outcome, along with background variables. Phonological processing is a metalinguistic skill that influences future reading development in monolingual and bilingual children. Emerging readers with strong phonological processing skills extend their speech-sound awareness when identifying and combining corresponding graphemes (i.e., speech-sound symbols) to decode word-level text. Our examination of U-PASS's initial criterion-based concurrent validity, after item-level revisions for discrimination power and accuracy rates in earlier tool evaluation phases 3A and 3B, demonstrates that U-PASS is significantly associated with word/non-word reading outcomes at the individual subtest and composite levels. Children with higher U-PASS scores also scored higher on the Urdu reading measure. Our findings are similar to commonly-used English phonological tools including CTOPP-2 (Wagner et al., 2013). Similar to other biliteracy studies (e.g., Mirza et al., 2017; O'Brien et al., 2019),

we analyzed and found moderate-level significant associations for Urdu phonological processing composite scores in relation to Urdu word/non-word reading, while accounting for external language and demographic background variables. The novel tool is well-functioning at subtest and composite levels, as evident by associations between phonological processing and word/non-word reading measures.

Clinical implications

It is important to develop culturally-and linguistically-responsive (i.e., based on language-specific phonological and orthographic structure) tools such as the U-PASS to assess and support language development across languages in school-aged children. Such tools enable identification of reading and language learning abilities at an earlier age via the dominant, but comparatively less assessed, heritage language in bilingual children. Dual language and biliteracy assessment allow educators and SLPs to determine if demonstrated reading deficits are due to *language-general* cognitive or linguistic disorders, in the case that these deficits are present in both heritage and societal languages regardless of language balance. Dual language assessment also enables clinicians to determine if demonstrated reading deficits are due to *language-specific* proficiency differences, in the case that these deficits are only present in the less dominant language.

Heritage language assessment tools, such as the U-PASS, facilitate timely identification of bilingual children, across English proficiency levels, who require additional pre-reading and reading skill support. These bilingual children face an increased risk of under-identification, for future reading support, when solely assessed on traditional English phonological measures used in the school system (Rose et al., 2022). In addition, access to standardized heritage language assessment tools allows SLPs and educators to comprehensively assess and compare bilinguals to age-matched peers with similar language and proficiency backgrounds (Mueller Gathercole et al., 2008). This prevents misidentification of pre-reading and language deficits in bilingual speakers of non-European languages, due to limited accessible tools and associated performance norms and language development milestones. From both a research and clinical perspective, U-PASS promotes linguistic diversity in child research by generating data from under-represented languages. SLPs and researchers can utilize and adapt the tool to collect data on language and reading development milestones and age-based norms for monolingual and bilingual readers of Urdu and related languages like Arabic and Persian.

Overall, the results of this study demonstrate that our Urdu Phonological Tele-Assessment (U-PASS) tool is associated with and can provide meaningful information about Urdu-English bilingual children's Urdu reading performance at the early grade levels. Such pre-reading assessment tools also provide insight into a child's language learning abilities, which develop alongside and as a result of their reading skills as children *read to learn* novel information (including about language forms and structures) at higher grade levels. Pre-reading assessment tools, such as the U-PASS, facilitate early identification and support of reading abilities and, in turn, help prevent future oral-language difficulties. Oral-language deficits may occur due to the inability to access textual sources in the classroom, which often contain language-related information (Mol and Bus, 2011).

The open-access tool also helps reduce assessment delays across bilingual communities, including speakers of Urdu and related languages, facing resource and economic barriers to reading acquisition—thereby mitigating long-term societal effects of child illiteracy. While U-PASS has not been validated for in-person assessment, our open-access tool is available for educators and SLPs to administer in both in-person and tele-assessment formats (see S40 for open access). We did so to address equity barriers relating to resource (e.g., technological or internet) limitations, and based on prior research demonstrating reliable use of pre-reading assessment tools for both in-person and online settings (Wood et al., 2021).

Limitations and future directions

In its current format, U-PASS is ideal for assessment purposes in Urdu speakers/readers. However, additional research is needed to determine its utility as a screening-based assessment, or as a diagnostic assessment of reading disorders/difficulties. In addition to the currently reported psychometric properties of the tool, a future longitudinal study will examine the U-PASS's predictive validity along with its clinical/diagnostic relevance for children with reading disorders. Given sample size differences and background similarities, we could not conduct cross-country comparisons between Pakistan and Canada in the current study. As well, due to online testing as a result of the COVID-19 pandemic, we could only assess children, from lower-upper middle SES backgrounds, with reliable internet access. This may affect our results' generalizability to lower SES backgrounds. As such, to enhance the U-PASS's generalizability, our future study will involve assessing children across diverse language and SES backgrounds.

In addition, while the current study incorporated various aspects of tool development highlighted by tool validation frameworks for language assessment (Bachman and Palmer, 2010; Chapelle, 2020; Weir, 2005), we could not consider influences of the type of testing environment and broader child physiological, psychological, and experiential characteristics at the time of assessment. This is due to our study design and voluntary sampling method, which involved tele-assessment of children by trained testers within their homes due to COVID-19 social distancing restrictions – rather than a uniform testing environment. While we collected and examined child language and demographic characteristics in our analyses, these were based on commonly-used parental reports and did not include a formal assessment of child characteristics—hence motivating our examination of these factors in a future study.

Author note

Correspondence may be addressed to Insiya Bhalloo, Department of Speech-Language Pathology, Rehabilitation Sciences Institute, University of Toronto (500 University Avenue, Toronto, ON, M5G 1V7). Phone: ++1 416-946-8638. Email: insiya.bhalloo@mail.utoronto.ca

Data availability statement

Supplementary files S1–S40 and supplementary figures 1, 2 can be accessed via the Borealis Dataverse Research Data Repository: https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/BRFIXM.

Ethics statement

The studies involving humans were approved by University of Toronto's Research Ethics Board (Protocol 38608). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

IB: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. MM: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review and editing.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ambreen, S., and To, C. K. (2021). Phonological development in Urdu-speaking children: A systematic review. *J. Speech Lang. Hear. Res.* 64, 4213–4234. doi: 10.1044/2021_JSLHR-21-00148

American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association [APA].

Anastasi, A., and Urbina, S. (1997). *Psychological testing*, 7th Edn. Hoboken, NJ: Prentice-Hall.

Arias, G., and Friberg, J. (2016). Bilingual language assessment: A contemporary versus recommended practice in American schools. *Lang. Speech Hear. Serv. Sch.* 48, 1–15. doi: 10.1044/2016_LSHSS-15-0090

Babatsouli, E. (2024). "Phonological disorders in child bilingualism," in *The Cambridge handbook of bilingual phonetics and phonology*, ed. M. Amengual (Cambridge: Cambridge University Press), 309–337.

Bachman, L. F., and Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Buchweitz, A. (2016). Language and reading development in the brain today: Neuromarkers and the case for prediction. *J. Pediatria* 92, 8–13. doi: 10.1016/j.jped.2016.01.005

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavskı, M., Onochie-Quintanilla, E., et al. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychol. Sci.* 23, 678–686. doi: 10.1177/0956797611434536

Chan, S. W., Cheung, W. M., Huang, Y., Lam, W. I., and Lin, C. H. (2020). Development and validation of a Chinese character acquisition assessment for second-language kindergarteners. *Lang. Test.* 37, 215–234. doi: 10.1177/0265532219876527

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: SAGE.

Dar, M., Anwaar, H., Vihman, M., and Keren-Portnoy, T. (2015). Developing an Urdu CDI for early language acquisition. *York Pap. Ling.* 14, 1–4.

Durham District Catholic School Board, (2020). *Kindergarten tools: Phonological awareness*. Oshawa, ON: Durham District Catholic School Board.

Eberhard, D., Simons, G., and Fennig, C. (2025). *Ethnologue: Languages of the world*. Available online at: https://www.ethnologue.com/ (accessed May 20, 2025).

Elosua, P. (2024). A three-step DIF analysis of a reading comprehension test across regional dialects to improve test score validity. *Lang. Assess. Q.* 85, 1–18. doi: 10.1080/15434303.2024.2307621

Fletcher, J. M., Simos, P. G., Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., and Papanicolaou, A. C. (2000). "Neuroimaging, language, and reading: The interface of

brain and environment," in *Proceedings of a research symposium on high standards in reading for students from diverse language groups: Research, practice & policy*, (Washington, DC: The Office of Bilingual Education and Minority Languages Affairs, U. S. Department of Education), 41–58.

Flowers, H. L., Flamand-Roze, C., Denier, C., Roze, E., Silver, F. L., Rochon, E., et al. (2015). English adaptation, international harmonisation, and normative validation of the Language Screening Test (LAST). *Aphasiology* 29, 214–236. doi: 10.1080/02687038. 2014.965058

Gillam, R. B., Pea, E. D., Bedore, L. M., Bohman, T. M., and Mendez-Perez, A. (2013). Identification of specific language impairment in bilingual children: I. Assessment in English. *J. Speech Lang. Hear. Res.* 56, 1813–1823. doi: 10.1044/1092-4388(2013/12-0056)

Goldstein, B. A., and Fabiano, L. (2007). Assessment and intervention for bilingual children with phonological disorders. *ASHA Lead.* 12, 6–31. doi: 10.1044/leader.FTR2. 12022007.6

Government of Pakistan, (2021). *Single national curriculum*. Available online at: https://pide.org.pk/research/analyzing-the-pros-and-cons-of-single-national-curriculum/

Government of Pakistan, (2022). *New curriculum*. Available online at: https://www.mofept.gov.pk/Detail/NWJmMmM2YTQtM2YzYi00NjJkLTgzNDEtYzMxMT I4MTllY2Qw

Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press.

Hastings/Prince Edward School Board (2013). *Quick phonics awareness screener (Q-PAS)*. Belleville, ON: Hastings/Prince Edward School Board.

Hemsley, G., Holm, A., and Dodd, B. (2014). Identifying language difference versus disorder in bilingual children. *Speech Lang. Hear.* 17, 101–115. doi: 10.1179/2050572813Y.0000000027

Hueniken, K., Douglas, C. M., Jethwa, A. R., Mirshams, M., Goldstein, D. P., Martino, R., et al. (2020). Measuring financial toxicity incurred after treatment of head and neck cancer: Development and validation of the Financial Index of Toxicity questionnaire. *Cancer* 126, 4042–4050. doi: 10.1002/cncr.33032

Jabbar, A., and Iqbal, S. (2016). *Urdu compound words. Linguistics Pakistan*. Chandigarh: Punjab University.

Kandru-Pothineni, S., Warmington, M., Clarke, A., Hitch, G., Babayiðit, S., and Mishra, R. K. (2015). "Oral language assessment in bilingual children: Methodological issues," in *Proceedings of the society for the scientific study of reading: 22nd annual meeting (Hawaii)*, (Honolulu, HI).

Lance, C. E., Butts, M. M., and Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. *Organ. Res. Methods* 9, 202–220. doi: 10.1177/1094428105284919

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., et al. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Sci. Stud. Read.* 23, 220–234. doi: 10.1080/10888438.2018. 1510936

Loudon, C., and Macias-Muoz, A. (2018). Item statistics: Implications for exam design. *Adv. Physiol. Educ.* 42, 565–575. doi: 10.1152/advan.00186. 2016

Luft Baker, D., Park, Y., and Andress, T. T. (2021). Longitudinal predictors of bilingual language proficiency, decoding, and oral reading fluency on reading comprehension in Spanish and in English. *Sch. Psychol. Rev.* 52, 1–14. doi: 10.1080/2372966X.2021.2021447

Lundberg, I., Larsman, P., and Strid, A. (2012). Development of phonological awareness during the preschool year: The influence of gender and socio-economic status. *Read. Writ.* 25, 305–320. doi: 10.1007/s11145-010-9269-4

Macmillan, B. M. (2002). Rhyme and reading: A critical review of the research methodology. *J. Res. Read.* 25, 4–42. doi: 10.1111/1467-9817.00156

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Martin, N., and Hamaguchi, P. (2018). *Test of auditory processing skills (TAPS-4)*. Austin, TX: Pro-Ed.

Mather, N., Podhajski, B., Rhein, D., and Babur, N. (2011). *Phonological awareness skills screener (PASS). Special education network and inclusion association (SENIA)*.

McIntire, S. A., and Miller, L. A. (2005). *Foundations of psychological testing*, 2nd Edn. Thousand Oaks, CA: Sage.

Mirza, A., and Gottardo, A. (2022). The effects of second language literacy instruction on first language literacy: A comparison between Hindi-English and Urdu-English Canadian bilinguals. *J. Cult. Cogn. Sci.* 6, 229–248. doi: 10.1007/s41809-022-00100-4

Mirza, A., and Gottardo, A. (2023). The role of context in learning to read languages that use different writing systems and scripts: Urdu and English. *Languages* 8:86. doi: 10.3390/languages8010086

Mirza, A., Gottardo, A., and Chen, X. (2017). Reading in multilingual learners of Urdu (L1), English (L2) and Arabic (L3). *Read. Writ.* 30, 187–207. doi: 10.1007/s11145-016-9669-1

Mitchell, J. J. (2001). Comprehensive test of phonological processing. *Assess. Effect. Interv.* 26, 57–63. doi: 10.1177/073724770102600305

Mol, S. E., and Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychol. Bull.* 137, 267–296. doi: 10.1037/a0021890

Mueller Gathercole, V. C., Mon Thomas, E., and Hughes, E. (2008). Designing a normed receptive vocabulary test for bilingual populations: A model from Welsh. *Int. J. Biling. Educ.* 11, 678–720. doi: 10.2167/beb481.0 .

Mumtaz, S., and Humphreys, G. (2001). The effects of bilingualism on learning to read English: Evidence from the contrast between Urdu-English bilingual and English monolingual children. *J. Res. Read.* 24, 113–134. doi: 10.1111/1467-9817.t01-1-00136

Nunnally, J. C. (1978). *Psychometric theory*, 2nd Edn. New York, NY: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric theory*, 3rd Edn. New York, NY: McGraw-Hill.

O'Brien, B. A., Mohamed, M. B. H., Yussof, N. T., and Ng, S. C. (2019). The phonological awareness relation to early reading in English for three groups of simultaneous bilingual children. *Read. Writ. Interdiscip. J.* 32, 909–937. doi: 10.1007/s11145-018-9890-1

Ontario Curriculum and Resources, (2023). *Elementary curriculum (K-8)*. Available online at: https://www.dcp.edu.gov.on.ca/en/curriculum#elementary (accessed January 19, 2025).

Patterson, J. (2002). Relationships of expressive vocabulary to frequency of reading and television experience among bilingual toddlers. *Appl. Psycholing.* 23, 493–508. doi: 10.1017/S0142716402004010

Ritchie, S., and Bates, T. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychol. Sci.* 24, 1301–1308. doi: 10.1177/0956797612466268

Robertson, C., and Salter, W. (2017). *Phonological awareness test (PAT-2: NU)*. Lutz, FL: PAR Inc.

Rose, K., Armon-Lotem, S., and Altman, C. (2022). Profiling bilingual children: Using monolingual assessment to inform diagnosis. *Lang. Speech Hear. Serv. Sch.* 53, 494–510. doi: 10.1044/2021_LSHSS-21-00099

Rüsseler, J., Arendt, D., Münte, T. F., Mohammadi, B., and Boltzmann, M. (2021). Literacy affects brain structure – what can we learn for language assessment in low literates? *Lang. Assess. Q.* 18, 492–507. doi: 10.1080/15434303.2021.1931231

Teoh, W. Q., Brebner, C., and McAllister, S. (2018). Bilingual assessment practices: Challenges faced by speech-language pathologists working with a predominantly bilingual population. *Speech Lang. Hear.* 21, 10–21. doi: 10.1080/2050571X.2017. 1309788

Timpe-Laughlin, V., and Choi, I. (2017). Exploring the validity of a second language intercultural pragmatics assessment tool. *Lang. Assess. Q.* 14, 19–35. doi: 10.1080/15434303.2016.1256406

UNESCO (2017). *More than one half of children and adolescents are not learning worldwide*. Paris: UNESCO.

University of Washington (2018). *Understanding item analyses. Educational assessment*. Seattle, WA: University of Washington.

van Linden, S., and Cremers, A. H. (2008). *Cognitive abilities of functionally illiterate persons relevant to ICT use*. Berlin: Springer Berlin.

Wagner, R. K., and McBride-Chang, C. (1996). The development of reading-related phonological processes. *Ann. Child Dev.* 12, 177–206.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. A. (2013). *Comprehensive test of phonological processing (CTOPP-2)*. Austin, TX: Pro-Ed.

Warmington, M. (2015). *Receptive and expressive vocabulary test*. York: University of York.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Macmillan.

Williams, C. J., and McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *Int. J. Speech Lang. Pathol.* 14, 292–305. doi: 10.3109/17549507.2011.636071

Wood, E., Bhalloo, I., McCaig, B., Feraru, C., and Molnar, M. (2021). Towards development of guidelines for virtual administration of paediatric standardized language and literacy assessments: Considerations for clinicians and researchers. *SAGE Open Med.* 9:510. doi: 10.1177/20503121211050510

Woodcock, R. W. (2011). *Woodcock reading mastery tests (WRMT)*. London: Pearson Assessments.