Journal of English Laguage, Literature and Linguistics (JELLL)
Volume 4, Issue 1, 2026, 84–116

Urdu-English Vocabulary Development Assessment in Bilingual Young Children: Focusing on Domain-Specific Vocabulary Growth and Conceptual Scoring

Uzma Ghumman*

Abstract

This study examined vocabulary acquisition in Urdu-English bilingual toddlers across five domain-specific topic areas: animals, vehicles, kitchenware, food, and body parts, over six months. The final analytic sample comprised 110 children from Urdu-dominant households in North Nazimabad, Karachi, assessed at two time points using structured picture-card vocabulary tasks. Vocabulary knowledge was scored in three ways: Urdu-only, English-only, and conceptual scoring, which credited knowledge of a concept in either language.

Results showed measurable vocabulary growth across domains from T1 to T2, with receptive knowledge consistently exceeding expressive production. Conceptual scoring produced the most complete estimate of bilingual lexical knowledge, whereas English-only scoring underestimated children's vocabulary and increased the risk of classifying children as having low vocabulary when they knew the relevant concepts in Urdu. Baseline-adjusted models further indicated that domain-matched caregiver input, child interest, interactive English exposure, and SES were associated with vocabulary development in theoretically meaningful ways.

Overall, the findings support the use of bilingual and conceptual scoring in early vocabulary assessment and show why domain-specific interactional experience must be considered when interpreting lexical development in multilingual Pakistani contexts.

Keywords: bilingualism, toddlers, receptive vocabulary, expressive vocabulary, domain-specific learning, conceptual scoring, lexical development, Pakistan

Introduction

Pakistan is a multilingual country with Urdu as the national language, English as a constitutionally entrenched official language and a language of high institutional value, provincial languages such as Sindhi, Punjabi, Pashto, and Balochi, and many regional languages and varieties used in everyday community life (Rahman, 2002, 2005; Ahmad & Guijun, 2022; Ashraf, 2023). These languages do not carry equal social, educational, or economic value. English is strongly associated with government, higher education, elite schooling, and professional employment, whereas Urdu and provincial/regional languages organize family, neighbourhood, and cultural interaction (Ashraf, 2023). For this reason, a child's early bilingual vocabulary in Pakistan is shaped not only by exposure to two codes but also by the unequal social functions attached to those codes.

Given the social stratification of English access in Pakistan, family socioeconomic status, caregiver education, school aspirations, and parents' English-medium schooling histories may affect the English children hear before they start school (Mansoor, 2003; Rahman, 2005; Ashraf, 2023). This is true even for preschool children: children may hear English in the form of parents' aspirations for English-medium schooling, older siblings, media, books, and caregiver-child play. The toddler years are particularly revealing because vocabulary growth is fast and can be tracked over short time frames (Fenson et al., 2007; Côté et al., 2022). However, caregiver input cannot be reduced to a word count, as the strength of associations between input and vocabulary varies according to definitions of input (Coffey & Snedeker, 2025).

In this article, "back-and-forth" refers to turn-taking: reciprocal interactions in which the child vocalizes, gestures, names, or responds, and the caregiver responds with contingent language. This turn-taking provides the child with practice associating words with their referents, feedback, and opportunities to reuse labels in context; as such, it is more likely to predict vocabulary development than simply counting the words adults use (Donnelly & Kidd, 2021; Clark, 2018). The phrase "daily recordings" refers to naturalistic audio recordings in children's natural home environments, typically over the course of a day using small wearable or room-based

recorders. They are made at home because the goal is to capture the true quantity, composition, speakers, and activities that children experience beyond the test session; they are used to demonstrate the variability of children's language environments across families and communities (Bergelson et al., 2023; Bunce et al., 2025).

In addition to interactional input, recent research focuses on child-internal factors, particularly interest and attention. Children under five years old are not passive language learners; they focus on objects, activities and categories that are interesting to them. Experimental studies demonstrate that children learn labels more readily for categories in which they are interested, and longitudinal studies show that maintaining interest in a category predicts later vocabulary in that domain (Ackermann et al., 2020; Madhavan & Mani, 2025). In a stratified society like Pakistan, interest should not be thought of as purely intrinsic: it is mediated by material resources, caregiver practices, availability of toys/books, family practices, and the social value of semantic domains (Dailey & Bergelson, 2022; Rowe & Weisleder, 2020).

A key challenge in the study of bilingualism is quantifying vocabulary across languages. In multilingual settings, children's vocabulary may be spread across languages: some concepts are known only in one language, some have translation equivalents, and some are linked to borrowed or shared words. So, single-language tests may underestimate vocabulary size in bilinguals, while total vocabulary scores may overestimate vocabulary size by counting translation equivalents twice. Conceptual scoring overcomes this issue by defining the known concept as the unit of analysis and scoring the child once if they know the concept in either language (Bedore et al., 2005; Gross et al., 2014; Byers-Heinlein et al., 2024; Fitton et al., 2023).

Parent reports are commonly used to assess early vocabulary because parents can report words the child knows and uses in everyday contexts that may not be captured in the lab (Fenson et al., 2007; Weisleder et al., 2024). But in bilingual research, the validity of parent report hinges on clear choices about language, scoring, and caregiver language background. The parent may complete the checklist in one

language or in both languages, and the meaning of the results depends on whether the responses are scored as Urdu-only, English-only, total vocabulary, or conceptual vocabulary (Byers-Heinlein et al., 2024; Weisleder et al., 2024).

Aim of the Study

This research seeks to understand vocabulary development in Urdu-English bilingual toddlers in a multilingual Pakistani environment with unequal access to English. The research examines children around three years old from Urdu-speaking families in North Nazimabad, Karachi, who were not yet in school. Children's receptive and expressive vocabulary of 35 concepts from five semantic categories (animals, vehicles and parts, kitchenware, body parts, and food) was assessed at two time points six months apart. The study compares Urdu-only, English-only, and conceptual scoring to understand how scoring affects estimates of bilingual vocabulary and risk of misclassification. It also explores whether domain-specific caregiver input, child interest, interactive English exposure, and SES predict vocabulary growth. This study provides evidence from an understudied sociolinguistic context and offers a practical solution to the problem of how to fairly assess bilingual children's knowledge.

Literature Review

Sociolinguistic Context of Pakistan

Current scholarship increasingly treats bilingual development as a dynamic system shaped by the child, the interactional environment, and the broader social context. Variation among bilingual children is not caused by bilingualism alone; it reflects differences in age, timing of exposure, language balance, domain of use, interactional quality, and the social value attached to each language (Byers-Heinlein & Lew-Williams, 2013; Côté et al., 2022; Byers-Heinlein et al., 2025). This point is especially relevant in Pakistan because Urdu and English are not simply two interchangeable languages in the child’s environment. They are associated with different settings, expectations, and social opportunities.

Bilingual Development as a System

Bilingual vocabulary is difficult to test because a concept may be lexicalized in one language, both languages, or neither language. If a child produces an Urdu label but not an English label, they may not be delayed; the difference may be due to language exposure, communicative need, word frequency in routines, phonological salience, cross-linguistic similarity, ease of retrieval, or the context in which the word is used (Lam & Sheng, 2020; Côté et al., 2022; Siow et al., 2025). The problem is that standard tests typically measure only one language, and these tests may conflate limited access to a particular label with limited knowledge of the concept. Thus, scoring approaches impact what can be inferred about bilingual children (Byers-Heinlein et al., 2024; Fitton et al., 2023).

Measurement of Bilingual Vocabulary

Of the three scoring methods, conceptual scoring is most suitable for bilingual assessment because it counts a concept as known if the child knows it in one language. This prevents children from being penalised for unequal language distribution, which is common in bilingual children. For instance, a child who knows kutta in Urdu but not dog in English has shown knowledge of the concept. Research comparing single-language, total, and conceptual scoring demonstrates that conceptual scoring can prevent misclassification and provide a more accurate measure of bilingual vocabulary than single-language scoring (Bedore et al., 2005; Gross et al., 2014; Fitton et al., 2023; Byers-Heinlein et al., 2024).

Table 01. Measurement of Bilingual Vocabulary

Scoring Approach	Definition	Limitation	Analytical Implication
Single-language	Words counted in one language only	Underestimates bilingual competence	May falsely indicate deficits
Total vocabulary	Combined words across languages	Double-counts translation equivalents	Inflates vocabulary size

Scoring Approach	Definition	Limitation	Analytical Implication
Conceptual scoring	Counts unique concepts across languages	Requires careful mapping	Best reflects underlying knowledge

Evidence therefore indicates that apparent bilingual "deficits" may arise partly from scoring artifacts rather than from developmental limitations (Bedore et al., 2005; Gross et al., 2014; Byers-Heinlein et al., 2024; Fitton et al., 2023; Bylund et al., 2023).

Caregiver Input and Interaction

The relation between caregiver speech and vocabulary development is complex but well established. It is complex because studies differ in how they quantify input—adult word counts, child-directed speech, conversational turns, responsiveness, diversity of vocabulary, and domain-specific talk—and these measurement choices change the size of observed effects (Coffey & Snedeker, 2025). Evidence increasingly suggests that reciprocal, contingent interaction is especially important: when caregivers respond to the child's focus of attention, elaborate labels, and sustain turn-taking, the child receives usable opportunities to connect words with meanings (Donnelly & Kidd, 2021; Clark, 2018). Naturalistic home recordings also show substantial variation in daily language exposure across families; therefore, input should be measured directly rather than inferred only from social background variables such as SES, parental education, or school type (Bergelson et al., 2023; Bunce et al., 2025).

Parent-Report Measures

Parent-report instruments are vocabulary checklists completed by caregivers, not reading tests for children. Caregivers indicate whether their child understands or says each listed word, usually by marking items such as animals, foods, body parts, action words, and routines. The MacArthur-Bates Communicative Development Inventories are a widely used example: they were designed as caregiver-completed forms for assessing early

comprehension and production in infants and toddlers and have been adapted across many languages (Fenson et al., 2007; Weisleder et al., 2024). These instruments are useful because caregivers observe the child across repeated daily contexts, but bilingual use requires explicit reporting of checklist language, caregiver proficiency, scoring rules, and whether conceptual scoring is used (Weisleder et al., 2024).

Table 02. Parent-Report Measures

Aspect	Requirement for Validity	Risk if Unspecified
Scoring method	Clear definition (single vs conceptual)	Misclassification of ability
Outcome type	Separation of receptive vs expressive	Ambiguous findings
Language exposure	Caregiver proficiency and usage	Measurement bias

Domain-Specific Vocabulary Development

Children do not learn all word types at the same rate. Domain-specific vocabulary grows through the convergence of at least three mechanisms: the child's interest in a category, the frequency and quality of caregiver talk about that category, and the practical salience of the domain in daily routines. Food and body-part words, for instance, may recur during meals, dressing, bathing, and health routines, whereas vehicle or kitchenware words may depend more on household practices, mobility, toys, books, and safety rules. Thus, domain growth is not simply a general vocabulary process; it reflects the distribution of meaningful experience across semantic fields (Ackermann et al., 2020; Madhavan & Mani, 2025; Rowe & Weisleder, 2020).

Table 03. Domain-Specific Vocabulary Development

Factor	Mechanism	Empirical Insight
Child interest	Directs attention to specific domains	Facilitates faster word learning
Caregiver input	Provides domain-relevant exposure	Predicts domain-specific growth
Interaction alignment	Matches interest with input	Strengthens vocabulary acquisition

Domain-based designs therefore allow researchers to test why vocabulary growth differs across semantic categories by examining whether domain-specific interest and domain-matched input jointly predict vocabulary growth (Ackermann et al., 2020; Madhavan & Mani, 2025).

Socioeconomic Status (SES)

SES is relevant to language development because it indexes a cluster of resources and routines, including caregiver education, time, books and toys, interactional style, and access to English-valued institutions; it should not be interpreted as a simple causal variable by itself (Dailey & Bergelson, 2022; Rowe & Weisleder, 2020). In this study, the children were not yet in formal schooling, so SES is interpreted through pre-school pathways: parents' own educational histories, aspirations for English-medium schooling, older siblings' schooling, access to English media/books, and the likelihood that caregivers use English during play and naming routines. This connection is especially important in Pakistan, where English-medium education and English proficiency are closely tied to social mobility and classed opportunity structures (Mansoor, 2003; Rahman, 2005; Ashraf, 2023).

Cross-Language Mapping

Bilingual children develop interconnected lexical systems through translation equivalents, singlets, loanwords, and cognates. Vocabulary growth therefore involves mapping

concepts across languages rather than acquiring two fully independent vocabulary lists.

Table 04. Cross-Language Mapping

Concept	Description	Developmental Role
Translation equivalents	Same concept, different languages	Reflect conceptual overlap
Singlets	Words without equivalents	Indicate language-specific knowledge
Cognates	Phonologically similar equivalents	Facilitate faster acquisition

Translation equivalents, singlets, and cognates provide evidence that bilingual vocabulary is distributed rather than duplicated across languages. Early bilingual children may know a concept in only one language because that language is more useful in the relevant routine or because the word form is easier to retrieve. As vocabulary grows, children gradually add translation equivalents and build more connected cross-language mappings. Cognates and phonologically similar words can be learned earlier or retrieved more easily because similarity provides converging cues across the two languages (Tsui et al., 2022; Mitchell et al., 2024; Siow et al., 2025).

Research Gap and Significance

Although bilingual development is widely studied, young Urdu-English bilingual children in Pakistan remain underrepresented, especially in research that connects language development with social stratification. Much of the existing evidence comes from North American or European bilingual contexts and cannot be assumed to transfer directly to Pakistani families, where Urdu, English, provincial languages, and regional varieties carry distinct social meanings and institutional value.

The present study addresses this gap by combining domain-specific vocabulary tracking, conceptual scoring,

child interest, caregiver input, interactive English exposure, and SES in a single longitudinal design. This approach links sociolinguistic structure with measurement practice and developmental process, showing not only what Urdu-English bilingual toddlers know but also why different assessment methods can produce different conclusions about their vocabulary.

Research Objectives

The study aimed to examine vocabulary growth, focusing on both receptive and expressive changes across domains between T1 and T2. It also compared different measurement approaches, including Urdu-only, English-only, and conceptual scoring methods. Additionally, the research investigated predictors of vocabulary development, specifically children's interest, caregiver input, English exposure, and socioeconomic status (SES).

Research Questions

- RQ1: How do receptive and expressive domain-specific vocabulary scores change between T1 and T2 in Urdu-English bilingual toddlers from Urdu-speaking households in Karachi, and do growth trajectories differ across semantic domains (animals, vehicles, kitchenware, body parts, and food)?
- RQ2: To what extent do different vocabulary scoring methods (Urdu-only, English-only, and conceptual scoring) produce different estimates of bilingual vocabulary growth, and how does the choice of scoring method affect the rate of misclassification of children as having low vocabulary?
- RQ3: To what degree do child-level factors (domain-specific interest), caregiver-level factors (interaction-based input and interactive English exposure), and family-level factors (socioeconomic status) independently and jointly predict domain-specific vocabulary growth in Urdu-English bilingual toddlers between T1 and T2?

Method

Study design

This study used a quantitative longitudinal design to examine vocabulary development among Urdu-English bilingual toddlers. The same children were assessed at two points separated by six months (T1 and T2), allowing developmental change to be estimated within participants. No intervention or experimental manipulation was introduced; the design was therefore appropriate for the research questions because the aim was to describe naturally occurring growth, compare scoring methods, and identify predictors of vocabulary gains rather than test the effect of a treatment.

At each visit, children completed two counterbalanced structured picture-card vocabulary blocks: Urdu mode and English mode. Receptive and expressive vocabulary were assessed for domain-specific nouns and scored using Urdu-only, English-only, and conceptual scoring schemes. Caregivers completed questionnaires and participated in a standardized recorded play interaction with the child, providing observational and report-based data on home language use.

The design directly addressed three linked aims: change in domain-specific vocabulary over time, variation introduced by scoring method, and the predictive value of child interest, caregiver input, interactive English exposure, and SES.

Sampling and Participants

Our final sample included 110 Urdu-English bilingual toddlers (70 girls, 40 boys) from Urdu-dominant families in North Nazimabad, Karachi, Pakistan. Children were purposively recruited to be around three years of age, exposed to both Urdu

and English, and not yet in formal education. Parents gave written consent for voluntary participation. Children were assessed at T1 (33.6–36.0 months; M = 34.8 months) and again at T2 (39.6–42.0 months; M = 40.8 months). Participants needed to be exposed to Urdu and English and have no caregiver-reported developmental, hearing or visual disabilities. A total of 200 children were recruited; 110 were retained for analysis as they completed both T1 and T2. We conservatively limited the sample to children with complete data at both time points; available comparisons showed no systematic differences between retained and non-retained children.

Target Domains

Vocabulary was assessed across five culturally relevant domains, each containing seven concepts (35 total).

- **Animals:** dog, cat, cow, crow, parrot, lizard, rat
- **Vehicles & Parts:** car, wheel, keys, door, window, firetruck, garbage truck
- **Body Parts:** leg, ear, lips, eyes, nose, finger, head
- **Kitchenware:** spoon, plate, bowl, knife, stove, rolling pin, bottle
- **Food:** apple, bread, egg, grapes, banana, water, milk

Two different picture cards were used to illustrate each of the concepts to reduce repetition effects within language blocks. The order of presentation was randomized and alternation of domain was done to maintain attention. Pre-specified Urdu and English sets of labels and rules against mixed responses and loanwords were predefined against each concept before the data collection.

Vocabulary Measures

Expressive Vocabulary

Expressive vocabulary was measured through structured picture-naming tasks in Urdu-mode and English-mode blocks. A brief warm-up established the target language context before the task. Elicitation involved two stages:

Step 1 involved spontaneous naming using prompts such as "Yeh kya hai?" or "What's this?".

Step 2 probed language-specific ability when needed, using prompts like "Do you know it in English too?" or "Kya ye apko Urdu me be aata hai?".

No translations or corrective feedback were provided. Two expressive outcomes were defined: spontaneous production (primary) and ability-based production (sensitivity).

Receptive Vocabulary

Receptive knowledge was assessed through picture identification tasks in both language modes using prompts such as "Mujhe ___ dikhāo." and "Show me the ___." Children responded by pointing or touching.

Scoring and Coding

Both receptive and expressive measures were scored using language-specific and conceptual approaches. Coding recorded correctness, language of label, and stage of production.

Table 05. Scoring and Coding

Scoring Type	Definition	Range per Domain
Urdu-only	Correct responses in Urdu mode	0–7
English-only	Correct responses in English mode	0–7
Conceptual	Concept known in either language	0–7
Expressive (Spontaneous)	Correct at Step 1 only	0–7
Expressive (Ability)	Correct at Step 1 or Step 2	0–7

Mixed utterances were coded according to the lexical item used for the target concept rather than the language frame of the sentence. For example, bottle in an Urdu sentence was coded as

an English lexical item unless the child also produced the Urdu/Urdu-adapted form botal. When both labels were produced (e.g., darwaza … door), both language-specific scores were credited, while the conceptual score counted the underlying concept once.

Potential loanwords and shared forms were handled through pre-specified rules. Sensitivity analyses reclassified ambiguous borrowed items and, in a separate check, excluded them to determine whether the main findings depended on borrowing decisions.

Stimuli and Label Sets

All stimuli consisted of picture cards representing target concepts. Accepted labels included both Urdu and English variants, for example:

Dog: English – dog, doggy | Urdu – kutta

Car: English – car | Urdu – gaari

Bottle: English – bottle | Urdu – botal

Water: English – water | Urdu – pani

Phonological approximations were accepted if clearly identifiable. Overgeneral labels such as “animal” or “thing” were coded as incorrect but recorded for descriptive purposes.

Predictor Measures

Child domain interest was measured through caregiver ratings on a Likert scale for each of the five semantic domains. Caregiver input was measured using both observed domain-relevant utterances during play and caregiver-reported frequency of domain-related engagement.

Interactive English exposure was calculated from caregiver utterances during a standardized 15-minute play session:

Interactive English exposure = English utterances / (English + Urdu + Mixed utterances).

A robustness index excluded mixed utterances from the denominator. These observational measures were triangulated with caregiver-reported estimates of daily language use.

SES was measured at T1 as a standardized composite based on caregiver education, caregiver occupation, and car ownership.

Procedure

At each visit, caregivers first completed questionnaires. Children then completed vocabulary tasks in counterbalanced Urdu-mode and English-mode blocks. A 15-minute caregiver-child play session was recorded to capture interaction-based language use. The same procedures were repeated at T2 to maintain comparability across time.

Coding and Reliability

All sessions were audio-recorded and transcribed. Responses were coded for correctness, language classification, and production stage. A randomly selected 25% of sessions was independently double-coded. Cohen's kappa was used for categorical agreement, and intraclass correlation coefficients (ICCs) were used for continuous exposure and input measures. Disagreements were resolved through discussion and adjudication.

Analysis

Domain-level analyses used baseline-adjusted models in which T2 scores were predicted from T1 scores. This approach estimated change while controlling for initial differences in vocabulary knowledge. Sensitivity analyses compared Urdu-only, English-only, and conceptual scoring.

Misclassification analyses defined low vocabulary using the lowest 10% of conceptual scores as the reference classification. Bootstrap confidence intervals were used to estimate false-positive and false-negative rates produced by language-specific scoring.

Predictor models included child interest, caregiver domain input, interactive English exposure, SES, and theoretically relevant interaction terms. Primary analyses used spontaneous expressive production; sensitivity analyses used ability-based expressive scoring.

This analytic strategy aligned with the research objectives by separately testing developmental change, scoring validity, and predictors of bilingual vocabulary growth.

Comprehensive methodology available in the Appendix.

Results

Growth trajectories (RQ1)

Children demonstrated vocabulary gains between T1 and T2 in all five semantic categories. The baseline-adjusted analyses considered each child's vocabulary level at T1, so the gains reported reflect change over the six months, not initial vocabulary level. As anticipated, receptive scores were higher than expressive scores across all scoring methods, and conceptual scores were higher than Urdu-only or English-only scores. Across the 35 target concepts, conceptual receptive scores increased from M = 24.48 at T1 to M = 27.78 at T2 (gain = 3.30), while conceptual expressive scores increased from M = 14.01 to M = 17.44 (gain = 3.43). Growth was evident across domains, but the size of gains differed by topic area, suggesting that lexical development reflected domain-specific everyday experience rather than a uniform increase across all words

outcome	scoring	model	domain	T1 mean	T2 mean	Raw gain (T2â^'T1)	Baseline-adjusted gain	95% CI (adj gain)
Expressive	Conceptual (spontaneous)	MixedLM (powell)	Animals	2.736364	3.572727	0.836363636	0.836363636	[0.57, 1.10]
Expressive	Conceptual (spontaneous)	MixedLM (powell)	BodyParts	2.872727	3.536364	0.663636364	0.663636364	[0.40, 0.93]
Expressive	Conceptual (spontaneous)	MixedLM (powell)	Food	2.809091	3.672727	0.863636364	0.863636364	[0.60, 1.13]
Expressive	Conceptual (spontaneous)	MixedLM (powell)	Kitchenware	3.036364	3.263636	0.227272727	0.227272727	[-0.04, 0.49]
Expressive	Conceptual (spontaneous)	MixedLM (powell)	VehiclesParts	2.554545	3.390909	0.836363636	0.836363636	[0.57, 1.10]
Expressive	English-only (spontaneous)	MixedLM (powell)	Animals	1.354545	1.8	0.445454545	0.445454545	[0.21, 0.68]
Expressive	English-only (spontaneous)	MixedLM (powell)	BodyParts	1.618182	1.909091	0.290909091	0.290909091	[0.05, 0.53]
Expressive	English-only (spontaneous)	MixedLM (powell)	Food	1.436364	1.863636	0.427272727	0.427272727	[0.19, 0.67]
Expressive	English-only (spontaneous)	MixedLM (powell)	Kitchenware	1.636364	1.8	0.163636364	0.163636364	[-0.07, 0.40]
Expressive	English-only (spontaneous)	MixedLM (powell)	VehiclesParts	1.318182	1.836364	0.518181818	0.518181818	[0.28, 0.76]
Expressive	Urdu-only (spontaneous)	MixedLM (powell)	Animals	1.645455	2.136364	0.490909091	0.490909091	[0.27, 0.71]
Expressive	Urdu-only (spontaneous)	MixedLM (powell)	BodyParts	1.718182	2.036364	0.318181818	0.318181818	[0.10, 0.54]
Expressive	Urdu-only (spontaneous)	MixedLM (powell)	Food	1.681818	2.227273	0.545454545	0.545454545	[0.32, 0.77]
Expressive	Urdu-only (spontaneous)	MixedLM (powell)	Kitchenware	1.527273	1.636364	0.109090909	0.109090909	[-0.11, 0.33]
Expressive	Urdu-only (spontaneous)	MixedLM (powell)	VehiclesParts	1.409091	1.736364	0.327272727	0.327272727	[0.11, 0.55]
Receptive	Conceptual	MixedLM (powell)	Animals	4.881818	5.618182	0.736363636	0.736363636	[0.52, 0.95]
Receptive	Conceptual	MixedLM (powell)	BodyParts	5.109091	5.609091	0.5	0.5	[0.28, 0.72]
Receptive	Conceptual	MixedLM (powell)	Food	5.1	5.536364	0.436363636	0.436363636	[0.22, 0.65]
Receptive	Conceptual	MixedLM (powell)	Kitchenware	4.690909	5.6	0.909090909	0.909090909	[0.69, 1.12]
Receptive	Conceptual	MixedLM (powell)	VehiclesParts	4.7	5.418182	0.718181818	0.718181818	[0.50, 0.93]
Receptive	English-only	MixedLM (powell)	Animals	2.918182	3.663636	0.745454545	0.745454545	[0.48, 1.01]
Receptive	English-only	MixedLM (powell)	BodyParts	3.127273	3.518182	0.390909091	0.390909091	[0.12, 0.66]
Receptive	English-only	MixedLM (powell)	Food	3.218182	3.363636	0.145454545	0.145454545	[-0.12, 0.41]
Receptive	English-only	MixedLM (powell)	Kitchenware	3.045455	3.863636	0.818181818	0.818181818	[0.55, 1.08]
Receptive	English-only	MixedLM (powell)	VehiclesParts	3.045455	3.527273	0.481818182	0.481818182	[0.22, 0.75]
Receptive	Urdu-only	MixedLM (lbfgs)	Animals	3.581818	3.9	0.318181818	0.318181818	[0.06, 0.57]
Receptive	Urdu-only	MixedLM (lbfgs)	BodyParts	3.672727	4.072727	0.4	0.4	[0.14, 0.66]
Receptive	Urdu-only	MixedLM (lbfgs)	Food	3.681818	3.972727	0.290909091	0.290909091	[0.04, 0.55]
Receptive	Urdu-only	MixedLM (lbfgs)	Kitchenware	3.336364	4.018182	0.681818182	0.681818182	[0.43, 0.94]
Receptive	Urdu-only	MixedLM (lbfgs)	VehiclesParts	3.472727	3.872727	0.4	0.4	[0.14, 0.66]

Table 06. Baseline-adjusted gains by domain (T2 controlling for T1). Values are baseline-adjusted gain with 95% CI.

Scoring framework (receptive)	T1, M (SD)	T2, M (SD)	Mean gain
Urdu-only	17.75 (4.24)	19.84 (3.70)	2.09
English-only	15.35 (4.77)	17.94 (4.75)	2.58
Conceptual	24.48 (4.66)	27.78 (3.89)	3.30

Table 07. Total scores across the 35 target concepts by scoring framework and time. Values are mean (SD); gain = T2 - T1.

Scoring framework (expressive)	T1, M (SD)	T2, M (SD)	Mean gain
Urdu-only	7.98 (3.31)	9.77 (3.26)	1.79
English-only	7.36 (3.90)	9.21 (4.41)	1.85
Conceptual	14.01 (4.27)	17.44 (4.26)	3.43

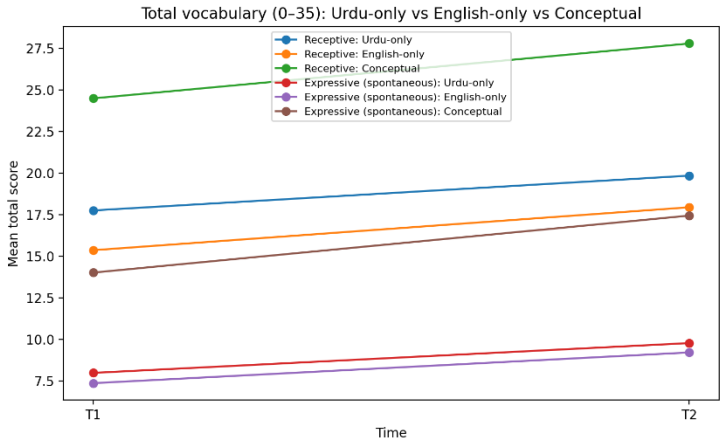

Figure 1. Mean total vocabulary scores (0-35) at T1 and T2 under Urdu-only, English-only, and conceptual scoring for receptive and expressive (spontaneous) outcomes. Conceptual scoring yields the highest estimates at both time points and shows the clearest growth over time.

Scoring framework, discrepancy, and misclassification (RQ2)

Conceptual scoring produced higher vocabulary estimates and larger growth than Urdu-only or English-only scoring, indicating that single-language totals underestimated children's lexical knowledge when knowledge could be expressed in either language. Urdu-English discrepancy scores (Urdu minus English) were positive on average (receptive T1: M = 2.39, SD = 4.36; expressive T1: M = 0.62, SD = 5.19; Table 08; Figure 2), but the size and direction of discrepancies varied across children and domains. When low vocabulary was defined as the lowest 10% of conceptual scores, English-only scoring produced more false-positive low-vocabulary classifications than conceptual scoring. For expressive vocabulary at T1, the English-only false-positive rate was 0.10 (95% CI [0.05, 0.15]), compared with 0.07 for Urdu-only scoring (Table 09; Figure 3). Robustness checks using a lenient borrowing rule and excluding potentially borrowed items produced the same overall pattern (Table 10).

Table 08. Urdu–English discrepancy (totals): Urdu total minus English total. Discrepancy is computed as Urdu-only total minus English-only total (range −35 to +35).

Outcome	Time	Mean	SD
Expressive (spontaneous)	T1	0.618	5.188
Expressive (spontaneous)	T2	0.563	6.066
Receptive	T1	2.39	4.358
Receptive	T2	1.9	4.635

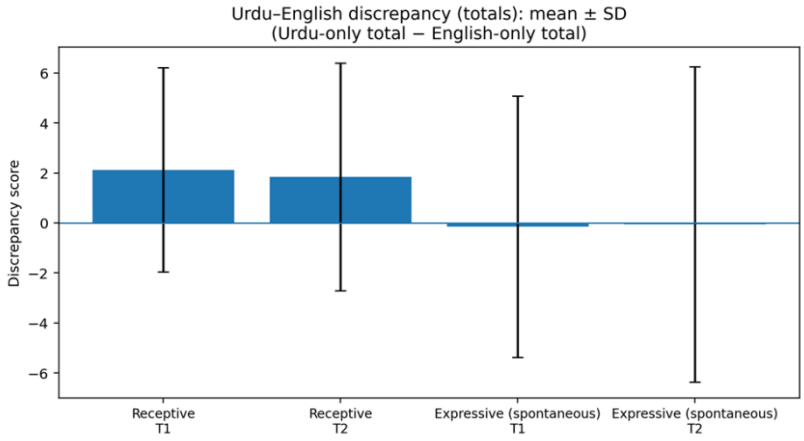

Figure 2 presents mean Urdu-English discrepancy scores, calculated as Urdu-only total minus English-only total, for receptive vocabulary and spontaneous expressive vocabulary at T1 and T2. Error bars represent ±1 SD. Positive values indicate higher Urdu-only totals, negative values indicate higher English-only totals, and the horizontal line at 0 indicates no Urdu-English difference.

outcome	time	method	Conceptual 10th pct threshold	Method 10th pct threshold	False positive	FP 95% CI	False negative	FN 95% CI
Receptive	T1	Urdu-only	19	12.9	0.02727273	[0.00, 0.10]	0.072727273	[0.00, 0.09]
Receptive	T1	English-only	19	10	0.10909091	[0.03, 0.15]	0.1	[0.03, 0.14]
Receptive	T2	Urdu-only	22.9	16	0.10909091	[0.00, 0.15]	0.009090909	[0.00, 0.08]
Receptive	T2	English-only	22.9	12	0.1	[0.05, 0.15]	0.072727273	[0.05, 0.14]
Expressive (spontaneous)	T1	Urdu-only	9	4	0.07272727	[0.04, 0.14]	0.081818182	[0.04, 0.15]
Expressive (spontaneous)	T1	English-only	9	3	0.1	[0.05, 0.15]	0.081818182	[0.03, 0.14]
Expressive (spontaneous)	T2	Urdu-only	11	6	0.11818182	[0.05, 0.15]	0.072727273	[0.04, 0.13]
Expressive (spontaneous)	T2	English-only	11	4	0.06363636	[0.03, 0.12]	0.045454545	[0.00, 0.12]

Table 09 presents misclassification against the conceptual reference classification (≤10th percentile). False positives indicate children classified as low by single-language scoring but not by conceptual

scoring; false negatives indicate children not classified as low by single-language scoring but classified as low under conceptual scoring. Bootstrap confidence intervals are percentile-based.

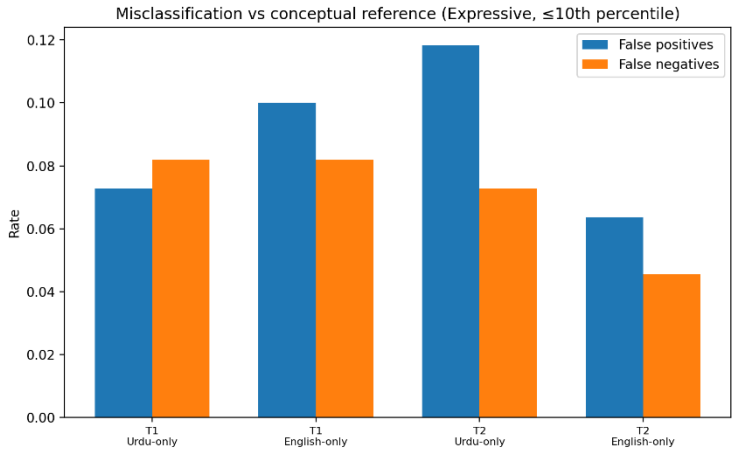

Figure 3 presents misclassification rates when single-language expressive classifications (Urdu-only and English-only) at the ≤10th percentile cutoff are compared with the conceptual expressive reference classification. False positives indicate children identified as ≤10th percentile by the single-language score but not by the conceptual reference; false negatives indicate children not identified as ≤10th percentile by the single-language score but classified as ≤10th percentile by the conceptual reference.

Table 10. Borrowing-robustness misclassification. The table shows false-positive (FP) and false-negative (FN) rates for single-language classification (Urdu-only vs English-only) relative to the conceptual reference, by outcome, time point, and conceptual scoring variant.

outcome	time	coding	method	FP rate	FP 95% CI	FN rate	FN 95% CI
Receptive	T1	Primary	Urdu-only	0.027273	[0.00, 0.10]	0.072727	[0.00, 0.09]
Receptive	T1	Primary	English-only	0.109091	[0.03, 0.15]	0.1	[0.03, 0.14]
Receptive	T1	Shared-status tolerant	Urdu-only	0.027273	[0.00, 0.07]	0.063636	[0.00, 0.09]
Receptive	T1	Shared-status tolerant	English-only	0.072727	[0.03, 0.15]	0.109091	[0.02, 0.14]
Receptive	T1	Excluding borrow-risk items (language totals only)	Urdu-only	0.045455	[0.02, 0.10]	0.081818	[0.01, 0.11]
Receptive	T1	Excluding borrow-risk items (language totals only)	English-only	0.109091	[0.04, 0.16]	0.1	[0.03, 0.14]
Receptive	T2	Primary	Urdu-only	0.109091	[0.00, 0.15]	0.009091	[0.00, 0.09]
Receptive	T2	Primary	English-only	0.1	[0.05, 0.15]	0.072727	[0.05, 0.14]
Receptive	T2	Shared-status tolerant	Urdu-only	0.081818	[0.01, 0.11]	0	[0.00, 0.08]
Receptive	T2	Shared-status tolerant	English-only	0.063636	[0.04, 0.12]	0.063636	[0.03, 0.11]
Receptive	T2	Excluding borrow-risk items (language totals only)	Urdu-only	0.090909	[0.00, 0.12]	0	[0.00, 0.08]
Receptive	T2	Excluding borrow-risk items (language totals only)	English-only	0.127273	[0.05, 0.18]	0.063636	[0.04, 0.13]
Expressive (spontaneous)	T1	Primary	Urdu-only	0.072727	[0.04, 0.15]	0.081818	[0.05, 0.15]
Expressive (spontaneous)	T1	Primary	English-only	0.1	[0.05, 0.15]	0.081818	[0.04, 0.15]
Expressive (spontaneous)	T1	Shared-status tolerant	Urdu-only	0.063636	[0.04, 0.17]	0.072727	[0.02, 0.15]
Expressive (spontaneous)	T1	Shared-status tolerant	English-only	0.036364	[0.02, 0.14]	0.063636	[0.01, 0.11]
Expressive (spontaneous)	T1	Excluding borrow-risk items (language totals only)	Urdu-only	0.090909	[0.05, 0.15]	0.1	[0.04, 0.16]
Expressive (spontaneous)	T1	Excluding borrow-risk items (language totals only)	English-only	0.1	[0.05, 0.16]	0.072727	[0.02, 0.14]
Expressive (spontaneous)	T2	Primary	Urdu-only	0.118182	[0.05, 0.15]	0.072727	[0.05, 0.13]
Expressive (spontaneous)	T2	Primary	English-only	0.063636	[0.03, 0.11]	0.045455	[0.00, 0.12]
Expressive (spontaneous)	T2	Shared-status tolerant	Urdu-only	0.090909	[0.03, 0.12]	0.063636	[0.02, 0.12]
Expressive (spontaneous)	T2	Shared-status tolerant	English-only	0.054545	[0.02, 0.09]	0.027273	[0.00, 0.09]
Expressive (spontaneous)	T2	Excluding borrow-risk items (language totals only)	Urdu-only	0.072727	[0.05, 0.14]	0.090909	[0.05, 0.13]
Expressive (spontaneous)	T2	Excluding borrow-risk items (language totals only)	English-only	0.090909	[0.05, 0.20]	0.063636	[0.02, 0.12]

Predictors and moderation (RQ3)

Baseline-adjusted predictor models examined whether T2 vocabulary scores were explained by T1 vocabulary, domain, child interest, caregiver input, English exposure, and SES. Caregiver input and domain-specific child interest accounted for additional variation in growth, particularly for expressive outcomes (Table 11). In the English-only spontaneous expressive model, higher interaction-based caregiver input and greater interactive English exposure were associated with larger gains (input: b = 0.12, 95% CI [0.00, 0.24], p = 0.047; English exposure: b = 0.42, 95% CI [0.25, 0.58], p < .001). The Input × English Exposure interaction was not statistically reliable (b = -0.09, 95% CI [-0.24, 0.07], p = .264), so the data did not provide strong evidence that the input effect differed by level of English exposure.

outcome	scoring	term	b	SE	95% CI	p
Receptive	Conceptual	T1_score_z	0.307061511	0.069555	[0.17, 0.44]	2.39E-05
Receptive	Conceptual	interest_T1_z	0.139285366	0.064548	[0.01, 0.27]	0.0331343
Receptive	Conceptual	input_T1_z	-0.016393827	0.063991	[-0.14, 0.11]	0.7982868
Receptive	Conceptual	eng_exposure_T1_z	0.124523472	0.077616	[-0.03, 0.28]	0.1115306
Receptive	Conceptual	SES_z	0.090258691	0.078916	[-0.07, 0.25]	0.2552427
Receptive	Conceptual	interest_T1_z:eng_exposure_T1_z	-0.154026588	0.094979	[-0.34, 0.03]	0.1077589
Receptive	Conceptual	input_T1_z:eng_exposure_T1_z	0.015425619	0.075783	[-0.13, 0.17]	0.8390851
Receptive	Conceptual	interest_T1_z:SES_z	0.14867766	0.073646	[0.00, 0.29]	0.0459619
Receptive	Conceptual	input_T1_z:SES_z	0.040315575	0.0777	[-0.11, 0.19]	0.6049102
Receptive	English-only	T1_score_z	0.072252017	0.065948	[-0.06, 0.20]	0.2756711
Receptive	English-only	interest_T1_z	0.029235303	0.067872	[-0.11, 0.16]	0.6675049
Receptive	English-only	input_T1_z	0.031172344	0.063908	[-0.10, 0.16]	0.6266913
Receptive	English-only	eng_exposure_T1_z	0.467382187	0.070064	[0.33, 0.61]	1.09E-09
Receptive	English-only	SES_z	0.265239011	0.085618	[0.10, 0.43]	0.0024789
Receptive	English-only	interest_T1_z:eng_exposure_T1_z	-0.004431457	0.092592	[-0.19, 0.18]	0.9619154
Receptive	English-only	input_T1_z:eng_exposure_T1_z	0.008776286	0.075688	[-0.14, 0.16]	0.9079026
Receptive	English-only	interest_T1_z:SES_z	0.068761938	0.081174	[-0.09, 0.23]	0.3988007
Receptive	English-only	input_T1_z:SES_z	0.013356369	0.088983	[-0.16, 0.19]	0.8809633
Receptive	Urdu-only	T1_score_z	0.145844859	0.071177	[0.00, 0.29]	0.0428595
Receptive	Urdu-only	interest_T1_z	0.111339059	0.074751	[-0.04, 0.26]	0.1392517
Receptive	Urdu-only	input_T1_z	0.092582996	0.073623	[-0.05, 0.24]	0.211251
Receptive	Urdu-only	eng_exposure_T1_z	0.025068925	0.068604	[-0.11, 0.16]	0.71551
Receptive	Urdu-only	SES_z	0.034849848	0.083335	[-0.13, 0.20]	0.6766326
Receptive	Urdu-only	interest_T1_z:eng_exposure_T1_z	-0.09704028	0.105557	[-0.31, 0.11]	0.3599612
Receptive	Urdu-only	input_T1_z:eng_exposure_T1_z	-0.056355439	0.0956	[-0.25, 0.13]	0.5567529
Receptive	Urdu-only	interest_T1_z:SES_z	0.096239609	0.088178	[-0.08, 0.27]	0.2774927
Receptive	Urdu-only	input_T1_z:SES_z	0.12789065	0.084894	[-0.04, 0.30]	0.1348366
Expressive	Conceptual (spontaneous)	T1_score_z	0.195469117	0.06181	[0.07, 0.32]	0.002027
Expressive	Conceptual (spontaneous)	interest_T1_z	0.130468093	0.062763	[0.01, 0.25]	0.0399883
Expressive	Conceptual (spontaneous)	input_T1_z	0.149245664	0.062465	[0.03, 0.27]	0.0185987
Expressive	Conceptual (spontaneous)	eng_exposure_T1_z	0.043585331	0.095213	[-0.15, 0.23]	0.648031
Expressive	Conceptual (spontaneous)	SES_z	0.262657515	0.088185	[0.09, 0.44]	0.0035709
Expressive	Conceptual (spontaneous)	interest_T1_z:eng_exposure_T1_z	0.077802568	0.081337	[-0.08, 0.24]	0.3409105
Expressive	Conceptual (spontaneous)	input_T1_z:eng_exposure_T1_z	-0.034530268	0.074709	[-0.18, 0.11]	0.6448595
Expressive	Conceptual (spontaneous)	interest_T1_z:SES_z	-0.192472878	0.078256	[-0.35, -0.04	0.0154842
Expressive	Conceptual (spontaneous)	input_T1_z:SES_z	0.107014558	0.079444	[-0.05, 0.26]	0.1807625
Expressive	English-only (spontaneous)	T1_score_z	0.143469406	0.068033	[0.01, 0.28]	0.0372498
Expressive	English-only (spontaneous)	interest_T1_z	0.06336505	0.059609	[-0.05, 0.18]	0.2901226
Expressive	English-only (spontaneous)	input_T1_z	0.122043128	0.060751	[0.00, 0.24]	0.0470184
Expressive	English-only (spontaneous)	eng_exposure_T1_z	0.417306106	0.082933	[0.25, 0.58]	1.93E-06
Expressive	English-only (spontaneous)	SES_z	0.17747355	0.082973	[0.01, 0.34]	0.0346746
Expressive	English-only (spontaneous)	interest_T1_z:eng_exposure_T1_z	0.15874862	0.087025	[-0.01, 0.33]	0.0708645
Expressive	English-only (spontaneous)	input_T1_z:eng_exposure_T1_z	-0.085706682	0.076387	[-0.24, 0.07]	0.2643235
Expressive	English-only (spontaneous)	interest_T1_z:SES_z	-0.019050784	0.07702	[-0.17, 0.13]	0.8051038
Expressive	English-only (spontaneous)	input_T1_z:SES_z	0.007555222	0.068127	[-0.13, 0.14]	0.9118997
Expressive	Urdu-only (spontaneous)	T1_score_z	0.051808072	0.06587	[-0.08, 0.18]	0.4332682
Expressive	Urdu-only (spontaneous)	interest_T1_z	0.084018698	0.055441	[-0.03, 0.19]	0.132552
Expressive	Urdu-only (spontaneous)	input_T1_z	0.036179877	0.060199	[-0.08, 0.16]	0.5490826
Expressive	Urdu-only (spontaneous)	eng_exposure_T1_z	-0.363768789	0.072821	[-0.51, -0.22	2.25E-06
Expressive	Urdu-only (spontaneous)	SES_z	0.091224282	0.070009	[-0.05, 0.23]	0.195306
Expressive	Urdu-only (spontaneous)	interest_T1_z:eng_exposure_T1_z	-0.116520773	0.071528	[-0.26, 0.03]	0.1061961
Expressive	Urdu-only (spontaneous)	input_T1_z:eng_exposure_T1_z	0.061008581	0.077458	[-0.09, 0.21]	0.4326177
Expressive	Urdu-only (spontaneous)	interest_T1_z:SES_z	-0.101307058	0.063077	[-0.23, 0.02]	0.1111492
Expressive	Urdu-only (spontaneous)	input_T1_z:SES_z	0.03967999	0.082323	[-0.12, 0.20]	0.6307709

Table 11 shows separate models are reported for receptive and expressive (spontaneous) outcomes under conceptual, English-only, and Urdu-only scoring. b = unstandardized coefficient; SE = standard error; 95% CI = confidence interval for b. Predictors are z-scored (suffix _z) and include T1_score, interest_T1, input_T1, eng_exposure_T1, and SES; interaction terms test moderation by English exposure or SES.

Sensitivity analysis (expressive ability)

Ability-based expressive scoring, which credited words produced after the language-specific follow-up prompt, produced the same overall pattern as the spontaneous-scoring analysis. In the English-only ability model, caregiver input (b = 0.11, 95% CI [-0.01, 0.23], p = .078) and interactive English exposure (b = 0.44, 95% CI [0.25, 0.62], p < .001) remained positively associated with growth, although caregiver input was marginal under this definition (Table 12).

outcome	scoring	term	b	SE	p	term_label	95% CI
Expressive (ability)	Conceptual (ability)	T1_score_z	0.195469	0.06181	0.002027	Baseline (T1) score	[0.07, 0.32]
Expressive (ability)	Conceptual (ability)	interest_T1_z	0.130468	0.062763	0.039988	Child interest (domain-matched)	[0.01, 0.25]
Expressive (ability)	Conceptual (ability)	input_T1_z	0.149246	0.062465	0.018599	Caregiver domain input (interaction-based)	[0.03, 0.27]
Expressive (ability)	Conceptual (ability)	eng_exposure_T1_z	0.043585	0.095213	0.648031	Interactive English exposure	[-0.15, 0.23]
Expressive (ability)	Conceptual (ability)	SES_z	0.262658	0.088185	0.003571	Household SES	[0.09, 0.44]
Expressive (ability)	Conceptual (ability)	interest_T1_z:eng_exposure_T1_z	0.077803	0.081337	0.34091	Interest Ã— English exposure	[-0.08, 0.24]
Expressive (ability)	Conceptual (ability)	input_T1_z:eng_exposure_T1_z	-0.03453	0.074709	0.644859	Input Ã— English exposure	[-0.18, 0.11]
Expressive (ability)	Conceptual (ability)	interest_T1_z:SES_z	-0.19247	0.078256	0.015484	Interest Ã— SES	[-0.35, -0.04]
Expressive (ability)	Conceptual (ability)	input_T1_z:SES_z	0.107015	0.079444	0.180762	Input Ã— SES	[-0.05, 0.26]
Expressive (ability)	English-only (ability)	T1_score_z	0.104153	0.071414	0.147595	Baseline (T1) score	[-0.04, 0.25]
Expressive (ability)	English-only (ability)	interest_T1_z	0.097572	0.061225	0.11391	Child interest (domain-matched)	[-0.02, 0.22]
Expressive (ability)	English-only (ability)	input_T1_z	0.107466	0.060427	0.078121	Caregiver domain input (interaction-based)	[-0.01, 0.23]
Expressive (ability)	English-only (ability)	eng_exposure_T1_z	0.437109	0.09349	8.45E-06	Interactive English exposure	[0.25, 0.62]
Expressive (ability)	English-only (ability)	SES_z	0.198844	0.086954	0.024141	Household SES	[0.03, 0.37]
Expressive (ability)	English-only (ability)	interest_T1_z:eng_exposure_T1_z	0.14903	0.087604	0.091761	Interest Ã— English exposure	[-0.02, 0.32]
Expressive (ability)	English-only (ability)	input_T1_z:eng_exposure_T1_z	-0.06605	0.0789	0.404318	Input Ã— English exposure	[-0.22, 0.09]
Expressive (ability)	English-only (ability)	interest_T1_z:SES_z	-0.02377	0.074887	0.751573	Interest Ã— SES	[-0.17, 0.12]
Expressive (ability)	English-only (ability)	input_T1_z:SES_z	0.010852	0.071738	0.880036	Input Ã— SES	[-0.13, 0.15]
Expressive (ability)	Urdu-only (ability)	T1_score_z	0.064998	0.063055	0.304913	Baseline (T1) score	[-0.06, 0.19]
Expressive (ability)	Urdu-only (ability)	interest_T1_z	0.135511	0.061827	0.030522	Child interest (domain-matched)	[0.01, 0.26]
Expressive (ability)	Urdu-only (ability)	input_T1_z	0.061707	0.06265	0.326834	Caregiver domain input (interaction-based)	[-0.06, 0.19]
Expressive (ability)	Urdu-only (ability)	eng_exposure_T1_z	-0.18318	0.077578	0.019993	Interactive English exposure	[-0.34, -0.03]
Expressive (ability)	Urdu-only (ability)	SES_z	0.172198	0.074186	0.022138	Household SES	[0.03, 0.32]
Expressive (ability)	Urdu-only (ability)	interest_T1_z:eng_exposure_T1_z	-0.07652	0.084069	0.364727	Interest Ã— English exposure	[-0.24, 0.09]
Expressive (ability)	Urdu-only (ability)	input_T1_z:eng_exposure_T1_z	0.073286	0.079479	0.358522	Input Ã— English exposure	[-0.08, 0.23]
Expressive (ability)	Urdu-only (ability)	interest_T1_z:SES_z	-0.10674	0.076881	0.167843	Interest Ã— SES	[-0.26, 0.05]
Expressive (ability)	Urdu-only (ability)	input_T1_z:SES_z	0.070755	0.088784	0.42722	Input Ã— SES	[-0.11, 0.25]

Table 12 shows expressive (ability) scoring. Outcome is expressive ability. Models are reported for conceptual, English-only, and Urdu-only scoring. b = unstandardized coefficient; SE = standard error; 95% CI = confidence interval for b; p = p-value. All predictors are z-scored: Baseline (T1) score, Child interest (domain-matched), Caregiver domain input (interaction-based), Interactive English exposure, and Household SES. Interaction terms (e.g., Interest × SES) test moderation.

Discussion

Main findings

This longitudinal study examined Urdu-English bilingual toddlers' vocabulary growth in Karachi over six months across five everyday semantic domains. Three main findings emerged. First, children showed clear vocabulary gains from T1 to T2, and receptive knowledge consistently exceeded expressive production. This receptive-expressive gap is consistent with evidence that comprehension and production draw on partly different processes: comprehension requires recognition of a familiar label-referent mapping, whereas production also requires retrieval, phonological encoding, and contextually

appropriate articulation (Bates, 1993; Hoff, 2021; Giguere & Hoff, 2022).

In bilingual toddlers, expressive performance is further shaped by the language in which a word is usually heard and used. The present results thus reflect both the general growth that occurs during the toddler years and the language ecology of Urdu-English bilingual households. The findings should not be taken to reflect that one language is easier or "better" than the other. Rather, they show that vocabulary outcomes depend on the distribution of opportunities to hear, use, and respond to labels in meaningful interaction. Second, the scoring method had a profound impact on inferences of development. Conceptual scoring yielded higher estimates of vocabulary knowledge and growth than Urdu-only and English-only scoring. This finding reflects the distributed nature of bilingual lexical representations: bilingual children do not necessarily know the translation equivalent for every concept they know, because lexical knowledge is distributed based on language-specific experience, need, and use (Tsui et al., 2022; Byers-Heinlein et al., 2024).

This is supported by the misclassification analysis, which demonstrates that single-language scoring can misclassify children as having low vocabulary when they know the concept in the other language. Third, vocabulary development was domain-specific and experience-related. Domains in which child interest and caregiver input aligned showed stronger growth, supporting an interactionist and usage-based interpretation of early lexical learning. Caregiver input was especially important for expressive outcomes, suggesting that real-time opportunities for labeling, elaboration, expansion, and contingent response help children move from comprehension to production (Tomasello, 2003; Clark, 2018; Donnelly & Kidd, 2021). Child interest should therefore be interpreted as a facilitative condition rather than an independent cause: interest opens a window for attention, but caregiver interaction determines whether that window becomes linguistically productive.

Domain-Structured Lexical Growth

The domain-specific approach is a key strength of this study because it reveals why vocabulary increased, rather than just how much vocabulary increased overall. Children did not learn

all words equally. Words related to everyday routines, such as body parts and food, may provide more opportunities for labeling, correction and child response than words such as vehicles and kitchenware, which may be dependent on family routines, toys, books, transportation, and household practices. This finding is in line with usage-based theories of language development, which postulate that children acquire words through meaningful use in interaction (Tomasello, 2003; Rowe & Weisleder, 2020). The fact that the receptive-expressive gap is consistent across domains and scoring systems also indicates that the asymmetry is a reliable feature of early bilingual development, rather than an artifact.

Importantly, that the receptive expressive gap does not shrink over time in terms of domains and across scoring systems suggests that this is not a domain-specific anomaly of early bilingual development. It is also more interpretable due to its robustness.

Scoring Choice and Interpretive Consequences

These findings demonstrate that scoring decisions are not merely technical. Conceptual scoring estimates the child's cross-language conceptual knowledge, whereas single-language scoring indicates only whether the child can access a particular label in a particular language during testing. The misclassification analysis makes this consequence practical: English-only scoring identified more children as low vocabulary than conceptual scoring. In an Urdu-dominant environment, this pattern could lead to unnecessary concern or referral if English-only tools are interpreted as measures of total language ability. The results therefore support recommendations that bilingual assessment include both languages and report conceptual scores alongside language-specific scores (Bedore et al., 2005; Gross et al., 2014; Fitton et al., 2023; Durán et al., 2025).

The consequences of the misclassification are especially significant. In practice, however, within an Urdu dominant setting, English-only methods of screening and assessing education can result in excessive identification of language delay. This highlights the importance of assessment models that help to differentiate between conceptual knowledge and language-specific retrieval. A bilingual child which forms kutta and not dog has learned the concept but has no ready mapping in

the experimental language situation. In the absence of conceptual scoring, these differences are not evident.

Borrowing and Shared Lexical Status

The major methodological insight of the research is its clear treatment of borrowing and shared words. Such objects as car, bottle, plate, etc. are in the grey linguistic position in multilingual urban settings. Instead of considering this noise, the study has incorporated borrowing in design and analysis by using pre-specified coding rules and strong checks.

The similarity of the outcome of alternative coding strategies shared-status inclusion and item exclusion proves that the core findings are not artefacts of lexical ambiguity. This enhances the soundness of the conceptual versus single language comparison and offers an explicit framework of dealing with the process of borrowing in bilingual research. Notably, it also indicates ecological validity in that it recognises that the two language lexical lists are constructed based on actual language practices, and not on the perfect language constraints.

Domain-Specific Predictors of Growth

The predictor analyses help explain why some children learned more words than others. Caregiver input in the same domain was a strong and consistent predictor, particularly for expressive vocabulary, suggesting that children grew their vocabulary when caregivers discussed the same categories. Child interest was also important, but in a more nuanced way: interest may focus attention, but vocabulary gains depend on whether caregivers provide labels, expansions, and responsive interaction in that domain. English vocabulary growth was also linked to interactive English exposure and SES. Crucially, English exposure was defined as interactive conversational English between caregivers and children, not background noise. SES is therefore a proxy for a bundle of opportunities (caregiver education, resources, language attitudes, English-medium school aspirations, access to English books/media) rather than a causal factor (Dailey & Bergelson, 2022; Rowe & Weisleder, 2020; Rahman, 2005; Ashraf, 2023). Child interest had a more context-specific effect. While this can direct attention and can lead to more engagement, it cannot lead to lexical growth without language input. This interaction between interest and input suggests that interest provides learning opportunities, but

the interaction between the parents determines if the opportunities are linguistically useful. Exposure to interactive English and SES also predicted English vocabulary. The exposure effect suggests a distinction between passive exposure to English and interactive use of English, given that exposure was defined as interactive use. The effects of SES are likely to reflect the ecological factors of resource allocation, parental education, language attitudes, and expectations for English-medium education. Thus, SES should be understood as an indicator of related factors.

Methodological Implications

The study has several implications for bilingual assessment. The two language-mode blocks eliminated uncertainty by providing children with contextual opportunities to respond in Urdu and English without assuming that the two languages are separate in everyday life. The two-step elicitation of expressive vocabulary also separated spontaneous production from ability-based production, allowing for a more complex understanding of expressive vocabulary. The triangulation of exposure measures through interaction coding and caregiver report enhanced the validity of the exposure measures by integrating observed behaviour with estimates of daily practices. The consistency of findings across sensitivity analyses indicates that the main conclusions were not dependent on a single operational definition of borrowing or expressive ability.

Limitations and Boundary Conditions

The study has several limitations. The study participants were Urdu-English bilingual toddlers from one urban center of Karachi, and cannot be generalised to all Pakistani children or to children from other language backgrounds. The domain lists were culturally relevant but short and could not capture the full range of children's vocabulary. The testing environment may also have affected performance, particularly for expressive measures, as young children can be shy or uncomfortable in the testing environment. Parental input was assessed in a structured play interaction, which may not reflect home routines. Finally, the purposive sampling strategy also means that engagement measures may be influenced by families' willingness to engage in the study.

Recommendations for Future Research

Future studies should add more time points to the longitudinal design to allow more accurate descriptions of development via growth-curve models. Ecological validity would be enhanced by including home recordings and comparing these to clinic- or researcher-administered measures. Future studies should also treat borrowing as a graded phenomenon rather than a simple binary category. Detailed coding of turn-taking, contingent response, expansions, and domain-specific elaboration would help identify which aspects of caregiver input are most predictive of vocabulary growth. Increasing sample sizes to include children exposed to other Pakistani languages (Punjabi, Sindhi, Pashto, Balochi, regional dialects) would enhance representativeness. Additional analyses could also explore gender, rural-urban settings and school transitions to understand variability in early bilingual development.

Conclusion

This research demonstrates that Urdu-English bilingual toddlers in Karachi showed vocabulary growth over six months, and that their receptive vocabulary knowledge was greater than their production across topics. Conceptual scoring was a better measure of children's vocabulary knowledge than single-language scoring because it counted concepts known in either Urdu or English. English-only scoring was especially likely to over-identify low vocabulary in an Urdu-dominant sample, demonstrating why assessment practices must distinguish between limited access to an English label and limited conceptual knowledge. Vocabulary growth was also explained by domain-specific experience: children gained more when caregiver input, child engagement, and interactive English exposure supported the relevant semantic domains. The results support a bilingual approach to assessment that tests both languages, provides conceptual and language-specific scores, and considers sociolinguistic factors when interpreting outcomes. For Pakistan and similar multilingual settings, the study cautions against single-language testing and offers a practical framework for more equitable and developmentally valid vocabulary assessment.

References

Ackermann, L., Hepach, R., & Mani, N. (2020). Children learn words easier when they are interested in the category to which the word belongs. Developmental Science, 23(3), e12915. https://doi.org/10.1111/desc.12915

Ahmad, N., & Guijun, Z. (2022). Inclusion or exclusion: Language ideology and language-in-education policies in Pakistan. European Journal of Language Policy, 14(2), 181–203.

Ashraf, H. (2023). The ambivalent role of Urdu and English in multilingual Pakistan: A Bourdieusian study. Language Policy, 22(1), 25–48. https://doi.org/10.1007/s10993-022-09623-6

Bates, E. (1993). Comprehension and production in early language development. Monographs of the Society for Research in Child Development, 58(3–4), 222–242. https://doi.org/10.1111/j.1540-5834.1993.tb00403.x

Bedore, L. M., Peña, E. D., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? Language, Speech, and Hearing Services in Schools, 36(3), 188–200. https://doi.org/10.1044/0161-1461(2005/020)

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., Rague Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Van Alphen, P. M., & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. Proceedings of the National Academy of Sciences, 120(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Alam, F., Migdalek, M. J., & Casillas, M. (2025). A cross-linguistic examination of young children's everyday language experiences. Journal of Child Language, 52(4), 786–814. https://doi.org/10.1017/S030500092400028X

Byers-Heinlein, K., Gonzalez-Barrero, A. M., Schott, E., & Killam, H. (2024). Sometimes larger, sometimes smaller: Measuring vocabulary in monolingual and bilingual infants and toddlers. First Language, 44(1), 74–95. https://doi.org/10.1177/01427237231204167

Byers-Heinlein, K., Kircher, R., & Lew-Williams, C. (2025). A systems framework of bilingual language acquisition: How development, experience, and contexts interact to shape outcomes. Annual Review of Developmental Psychology, 7, 219–242. https://doi.org/10.1146/annurev-devpsych-010923-110936

Byers-Heinlein, K., & Lew-Williams, C. (2013). Bilingualism in the early years: What the science says. LEARNing Landscapes, 7(1), 95–112.

Bylund, E., Antfolk, J., Abrahamsson, N., Haug Olstad, A. M., Norrman, G., & Lehtonen, M. (2023). Does bilingualism come with linguistic costs? A meta-analytic review of the bilingual lexical deficit. Psychonomic Bulletin & Review, 30(3), 897–913. https://doi.org/10.3758/s13423-022-02136-7

Clark, E. V. (2018). Conversation and language development: A pragmatic approach. Language Learning and Development, 14(3), 170–185. https://doi.org/10.1080/15475441.2017.1340843

Coffey, J. R., & Snedeker, J. (2025). How strong is the relationship between caregiver speech and language development? A meta-analysis. Journal of Child Language, 53(2), 435–470. https://doi.org/10.1017/S0305000924000692

Côté, S. L., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2022). Multilingual toddlers' vocabulary development in two languages: Comparing bilinguals and trilinguals. Journal of Child Language, 49(1), 114–130. https://doi.org/10.1017/S030500092000077X

Dailey, S., & Bergelson, E. (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. Developmental Science, 25(3), e13192. https://doi.org/10.1111/desc.13192

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. Child Development, 92(2), 609–625. https://doi.org/10.1111/cdev.13511

Durán, L., Zegers, M., Siebert, J. M., Zopluoglu, C., Jasso, J., Pratt, A., Pei, F., & Gorno-Tempini, M. L. (2025). Innovations in conceptual scoring of a bilingual expressive vocabulary measure. American Journal of Speech-Language Pathology, 34(5), 2618–2631. https://doi.org/10.1044/2025_AJSLP-24-00524

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.). Brookes.

Fitton, L., Goodrich, J. M., Thayer, L., Pratt, A., & Luna, R. (2023). Bilingual vocabulary assessment: Examining single-language, conceptual, and total scoring approaches. Journal of Speech, Language, and Hearing Research, 66(9), 3486–3499. https://doi.org/10.1044/2023_JSLHR-22-00573

Giguere, D., & Hoff, E. (2022). Bilingual development in the receptive and expressive domains: They differ. International Journal of Bilingual Education and Bilingualism, 25(10), 3849–3858. https://doi.org/10.1080/13670050.2022.2087039

Gross, M., Buac, M., & Kaushanskaya, M. (2014). Conceptual scoring of receptive and expressive vocabulary measures in simultaneous and sequential bilingual children. American Journal of Speech-Language Pathology, 23(4), 574–586. https://doi.org/10.1044/2014_AJSLP-13-0026

Hoff, E. (2021). Why bilingual development is not easy. Advances in Child Development and Behavior, 61, 129–168. https://doi.org/10.1016/bs.acdb.2021.03.002

Lam, B. P. W., & Sheng, L. (2020). Taxonomic development in young bilingual children: Task matters, and so does scoring method. American Journal of Speech-Language Pathology, 29(3), 1162–1177. https://doi.org/10.1044/2019_AJSLP-19-00143

Madhavan, R., & Mani, N. (2025). Children's individual interests are sustained across development and predict later vocabulary development. Journal of Child Language, 1–25. https://doi.org/10.1017/S0305000925100378

Mansoor, S. (2003). Language planning in higher education: Issues of access and equity. The Lahore Journal of Economics, 8(2), 17–42.

Mitchell, L., Tsui, R. K.-Y., & Byers-Heinlein, K. (2024). Cognates are advantaged over non-cognates in early bilingual expressive vocabulary development. Journal of Child Language, 51(3), 596–615. https://doi.org/10.1017/S0305000923000648

Rahman, T. (2002). Language, ideology and power: Language learning among the Muslims of Pakistan and North India. Oxford University Press.

Rahman, T. (2005). Passports to privilege: The English-medium schools in Pakistan. Peace and Democracy in South Asia, 1(1), 24–44.

Rowe, M. L., & Weisleder, A. (2020). Language development in context. Annual Review of Developmental Psychology, 2, 201–223. https://doi.org/10.1146/annurev-devpsych-042220-121816

Siow, S., Lepadatu, I., Gillen, N. A., & Plunkett, K. (2025). Bilingual toddlers' vocabulary growth interacts with existing knowledge and cross-linguistic similarity. Journal of Child Language, 1–28. https://doi.org/10.1017/S0305000924000710

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Harvard University Press.

Tsui, R. K.-Y., Gonzalez-Barrero, A. M., Schott, E., & Byers-Heinlein, K. (2022). Are translation equivalents special? Evidence from simulations and empirical data from bilingual infants. Cognition, 225, 105084. https://doi.org/10.1016/j.cognition.2022.105084

Weisleder, A., Friend, M., Tsui, A. S. M., & Marchman, V. A. (2024). Using parent report to measure vocabulary in young bilingual children: A scoping review. Language Learning, 74(2), 468–505. https://doi.org/10.1111/lang.12617