

Language of Assessment and the Visibility of Reasoning in the Early Grades: A Study of Construct Validity in Pakistan's Multilingual Classrooms

Ghulam Ali

Allama Iqbal Open University Islamabad

Uzma Asmat

Rawalpindi Women University

Ansar Ali

Allama Iqbal Open University Islamabad

Amir Haider

`amir.haider@nu.edu.kz`

Scholar Nazarbayev University Astana

Research Article

Keywords:

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11191816/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Tests are frequently treated as impartial apparatuses in multilingual schooling arrangements, but the language a child is tested in truly vagaries what the test evaluates. For millions of children across the globe who acquire education and take exams in a native language, the test language bounds both what they can elucidate and what instructors can notice. This study emphasizes on young students in Pakistan to response a critical question: when learners are compelled to reason and resolve problems in a non-native language taught at school, to what extent it restricts their genuine knowledge and rational remains unexpressed, or not fully utilized and how unprofessionally does that impact, how efficiently and fairly they are evaluated? Pakistan is one the country in the world which provides the most acute mismatch case of language in schooling. Ethnologue (2026) records 68 living languages in Pakistan, yet only two—Urdu and English—hold official status. After independence in 1947 the country moved only gradually from a largely monolingual Urdu orientation toward a more multilingual education policy in 2009, which began to acknowledge both the social demand for English and the case for including regional languages in the curriculum. Yet English and Urdu still dominate policy and practice, while the regional home languages of most children remain largely excluded as media of instruction (Ashraf, 2023). Urdu—the first language of under a tenth of the population (Gallup, 2024; Pakistan Bureau of Statistics, 2023)—is the primary medium in government schools, while English serves elite and many low-fee private schools (Soroya et al., 2024). The Federal Foundational Learning Policy (2024) benchmarks early learning in Urdu and English but stops short of designating a language of instruction, leaving children in the foundational grades to learn, and to demonstrate learning, in a language that is frequently foreign to them. The educational stakes are well documented for achievement; what remains unexamined is the problem of the measurement this creates. When a reasoning assignment is assigned to a child in a low resource language, their test score reflects a blend of their real discerning skills and their language capability. Our education systems promote rote learning than profound comprehension, similarly the evaluation system in Pakistan, can effortlessly misinterpreted how intelligent multilingual children truly are. They conclude misunderstood child's inadequate vocabulary for a lacking aptitude to deliberate. To correct this misconstrued approach, this research considers the test language as a factor that directly influences the final score. Investigators test each learned twice using undistinguishable thinking tasks initially in their native language, and then in the official school language. This methodology helps researchers distinguish a child's general language skills from the test itself, viewing precisely how much the language of the exam restricts real knowledge of a child. The study also examines how common teaching methods and instructor grading practices either worsen the problem or resolve it.

Research Questions

1. How huge is the gap within-student gap in cognitive efficiency when children are evaluated in native language as compared to the school language, and does it differ by proficiency, reasoning type, and province?

2. How do classroom language performance—teacher code-switching, trans-language, support, and classroom conversation value—shape children's prospects to demonstrate rational?
3. How do teachers' scores and elucidations of similar responses vary across the two language situations?

Consider a Grade 3 child in countryside Balochi who can elucidate, in Baluchistan, why an arrangement of objects follows a pattern, but who remains silent when the similar task is assigned in Urdu or English. A test accomplished only in an outside language, Urdu or English, records a failure; the child's cognition is integral but indiscernible. Multiplied across a national assessment, such episodes do not simply mismeasure individuals. They generate biased evidence about which children, regions, and languages are "low achieving," shaping teacher expectations, resource allocation, and policy conclusions about the quality of education itself.

The significance of this work is both scholarly and practical. Scholarly: the dominant research on language in education has examined achievement outcomes by medium of instruction, and the measurement literature on construct-irrelevant variance has developed largely in monolingual, high-income contexts and around content knowledge rather than reasoning. This project brings the two together and extends them to the early grades of the Global South particularly Pakistan, asking not whether multilingual children score lower but why, and whether those scores are valid measures of reasoning at all. Practical: by working with national and provincial assessment authorities, the project ensures that evidence about assessment-language effects reaches the actors who build and reform national assessment.

Rationale: Relevant Literature and Conceptual Framework

Relationship to existing literature

Three bodies of scholarship frame this study, and the project sits precisely in the space among them. The first concerns the effects of language of instruction on learning. A significant worldwide substantiation base specifies that children learn introductory skills more efficiently in a native language (Benson, 2004; Thomas & Collier, 2002) and spontaneous instruction in an unfamiliar language may reduce their performance especially at early stages. UNESCO has also advocated mother tongue-based education system as a base to equity (Ball, 2011). But in Pakistan recent analysis revealed that learners taught in native languages perform better than those who are instructed in Urdu or English languages and teaching in an unacquainted language can restrict a multilingual classroom platform which obstructs learning and participation (Coleman, 2010; Coleman & Capstick, 2012; Kwok et al., 2024; Soroya et al., 2024). Other Pakistani evidence, though concentrated at the secondary and tertiary levels, is consistent: the medium of instruction and students' English proficiency shape academic performance and classroom participation, with language anxiety depressing expression (Fiaz et al., 2025), and English—valued as a language of prestige and mobility—mediates how learners express themselves and are judged (Rehman et al., 2020). This literature establishes that the medium of instruction matters for what

children learn and how they are perceived. It does not, however, isolate the distinct question this study addresses—how the language of the test distorts the measurement of what children already know and can do—least of all in the early grades, where the evidence is thinnest.

The second body of work addresses precisely that distortion, but largely outside the Global South. Research on the assessment of emergent bilingual and English-language-learner populations has shown that linguistic complexity unrelated to the target construct depresses performance and threatens score validity (Abedi, 2006; Solano-Flores & Trumbull, 2003; Solano-Flores & Nelson-Barber, 2001). Where such assessments carry consequential stakes, they can operate as *de facto* language policy that systematically disadvantages emergent bilingual children (Menken, 2008). The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) name language as a recurrent source of construct-irrelevant variance and call for evidence that scores carry comparable meaning across linguistic groups. Differential item functioning (DIF) methods provide tools for detecting items that behave differently across groups after matching on ability (Holland & Wainer, 1993). This literature has matured around accommodations and item design, chiefly for older students in high-income systems. It has rarely been brought to bear on the early grades, on reasoning as opposed to content knowledge, or on low and middle-income multilingual systems where the entire instructional language may be unfamiliar (Heugh, 2011). Pakistan improves the gap. The NAT evaluates learning results at Grades 4 and 8, and national analyses describe a curricular and evaluation culture concerned with reminiscence rather than higher-order thinking (Soroya et al., 2024). In such a system, the conflation of language and reasoning is doubly consequential: tasks that could reveal reasoning are scarce to begin with, and the few that exist are administered in a language many children do not command. The specific cell that remains empirically dark—early-grade reasoning, measured in the home versus the school language, with the score gap decomposed into language and cognition—is precisely the cell this study illuminates.

The third body concerns the social and cognitive role of language in classroom learning and in testing as a social practice. Studies of classroom dialogue show that the quality of talk—exploratory rather than merely cumulative or disputational—shapes children's reasoning and its expression (Mercer, 2000; Mercer & Littleton, 2007), and that multilingual classrooms routinely deploy translanguaging and code-switching as resources for meaning-making (García & Wei, 2014). Critical scholarship on language testing demonstrates that assessments encode power relations and confer legitimacy on some languages and not others (Shohamy, 2001; Pardesi, 2022). These traditions explain why the measurement problem is not merely technical but consequential.

The contribution this project expects to make is to integrate these literatures into a single empirical investigation that has not, to our knowledge, been conducted: a within-student test, in an early-grade multilingual Global-South setting, of how assessment language affects the visibility of reasoning, accompanied by classroom and teacher-judgment evidence that situates the measurement effect in instructional practice. The new knowledge is a defensible, empirically grounded estimate of the magnitude of construct-irrelevant variance attributable to assessment language in early-grade reasoning, together with a model of how that variance is produced and interpreted in real classrooms. Its

contribution to the improvement of education is direct: assessment systems that misattribute language effects to cognitive deficit make invalid inferences with high stakes for children, teachers, and policy. Methodologically, the within-student decomposition of a reasoning score into language-attributable and reasoning-attributable components offers a transportable approach that other multilingual systems can adopt to audit the validity of their own assessments.

Conceptual framework

The study builds on four interlocked traditions that can all be construed as language as a constitutive condition for cognitive assessment and on which it operates, but it does not just stand on its rhetoric or bleed into it. Vygotsky's (1978, 1986) theory of mediated cognition is that mediated mental processes constitute the development of higher mental functions and the externalization of these functions – mediated by a number of cultural tools, primarily language – and that the language is not just transparent as a vehicle of thought but also serves as a control mechanism that regulates the possibility of the existence of the regulated process of knowledge in somebody else. Identification made in a language which minimizes expression can, therefore, reflect only somewhat linguistic performance and not cognitive ability. The realization of this insight in the project is achieved by eliciting reasoning with oral work and work of limited literacy level; post task verbal recall.

Linguistic interdependence hypothesis and common underlying proficiency models (Cummins, 1979, 2000) have been put forward as models of the way conceptual knowledge and reasoning acquired in one language are cognitively resident in other languages as well. The project operationalizes this by measuring each child's reasoning in both languages and by including a short measure of oral language proficiency as a covariate, allowing us to distinguish a language-of-assessment effect on the visibility of reasoning from a general proficiency difference.

The project's main analytic framework is provided by Messick's (1989, 1995) unified theory of construct validation, and the argument-based approach to validation that builds upon this framework (Kane, 2013). Validity requires that variance or differential performance related to the target construct (construct-relevant variance) predominates over variance or differential performance unrelated to the target construct (construct-irrelevant variance) in the assessment. The administration of a reasoning task in a child's weaker language brings in just this – scores combine reasoning with language ability. In addition to consequential validity's connection to the larger social context of utilizing scores, Messick also states that it is a part of the validity itself. The project implements the within-student contrast value estimates directly—the DIF analysis directly tests whether items have comparable meaning in the both language version(s)—the teacher-judgment strand is implemented directly, examining the consequences of the validity activity that has to do with how scores are produced and read.

Finally, Shohamy's (2001) critical language testing (CLT) framework places these technical queries in relation to power and reveals the institutionalizing of the privilege of dominant language assessment. The four traditions go together to back up a single thesis: Thinking is not language-neutral, and the medium of assessment makes thinking visible and makes thinking count. Table 1 outlines the mapping

of each tradition to a specific design decision; the argument is not considered decided but is submitted as a hypothesis to be tested by a designing machine which can measure the size of the argument and give the mechanisms for its implementation.

Research Design and Methods

Design overview and context

This study is a three-year convergent mixed-methods design where the quantitative strand is the within-student design, and the qualitative strand is embedded within the quantitative strand. The quantitative core discusses each child as his/her own control, the reasoning in each language condition, and differences between conditions; the qualitative strand (based on classroom observation, child verbal recalls, and teacher think-aloud scoring tasks) provides an understanding of how the observed differences in reasoning are produced and interpreted by teachers in practice. At the analysis stage the two strands come together to construct a validity argument that includes evidence to support the statistical estimates of construct-irrelevant variance, together with the evidence gathered with respect to its sources and consequences, in Messick's sense of validity. Table 2 aligns each research question with its data source, the way it is analyzed, and the inference to which it leads.

In the study, institutional partners coordinate research strategy, instrument development, validation, and analysis; assessment bodies facilitate school access and policy interpretation; and field organizations coordinate community engagement. This is, foremost, education research: non-research talents are used in service of a research agenda rather than the other way around; capacity-building and dissemination are to follow the research, not to precede it.

Sample, sites, and selection

The fieldwork has been conducted in three provinces to capture the polygenic diversity and demographic weight of different languages, namely Punjab (Punjabi), Khyber Pakhtunkhwa (Pashto), and Sindh (Sindhi). In each province we select government primary schools that fulfil a defining inclusion criterion (an inclusion criterion is a criterion that determines whether a policy, program, or service is included or excluded; in this study, this criterion is a product of our view that in Pakistan, the great majority of children in government primary schools speak not Urdu as the language of instruction and assessment, but another dominant home language). This criterion establishes structural condition–reasoning that is taught and tested in a non-dominant language, which the research question requires.

Eight schools per province (24 schools overall) and around 25 Grade 2/3 children per school or about 600 children were suggested as the quantitative sample of children. This sample meets the study's inferential needs: for every student, both conditions are sampled, resulting in high power in the multilevel models for the school-level random effect, this being around 200 school-level units of sampling; for items, a DIF screening based on logistic regression and Mantel-Haenszel was adequate (the minimum is 200 children per language). A nested qualitative subsample consists of one classroom observed in each

of these 24 schools (about 24 children), 3–4 children from each of the observed classrooms participating in verbal recalls (with a final n of approximately 80 children) and 2 teachers from each school engaging in the think aloud task with children (with a final n of approximately 48 teachers).

The deliberate sample size is carefully sized to be rigorous but manageable for individually administered, oral assessment across three provinces; the final sample size will be substantiated by the pilot of Year 1 and a formal power analysis. It is easy to see the sample size, at roughly 600 children in each of a within-student design which controls for stable between-child variation when computing error terms, such a design would be well powered to observe a small standardized effect for assessment language ($d < 0.20$), and would still have the power needed to observe the essential cross-level interactions (reasoning construct and province; proficiency band is built into to accommodate attrition and incomplete sessions as an approximation). We do not make any claims of the statistical representativeness of the school population of Pakistan. One design approach used for inferences of analytic generalization is to keep the structural conditions the same (assessment medium = non-dominant language), while varying linguistic context across the three primary language communities and extrapolating this across other multilingual assessment contexts via language-mediated measurement. Recruitment is carried out through district and school facilitation. Institutional ethical approval for all protocols is secured in Year 1, and bilingual community liaisons ensure informed consent/carer support that is culturally appropriate, as well as ongoing assent from children.

Key constructs, measures, and data sources

The target construct is reasoning, construed quite narrowly, explicitly measurable for the early grades as reasoning ability to draw inferences and solve novel problems; and operationalised as three task families: quantitative reasoning (mathematical word problem tasks); ordering and pattern completion (sequencing tasks); and relational mapping (analogical tasks). Instead of the more general and controversial term 'critical thinking', we specifically identify these constructs as they are directly measurable and relatively easily scorable, have lower confounds with maturity and schooling, and are developmentally appropriate.

The point is, all tasks are administered verbally by a trained, bilingual assessor, there is minimal reading (no extended writing) and no decoding (transcription skill) being measured, and are pictorial with simple manipulatives, which again is an important safeguard in relation to the construct-validity argument. Develop two parallel forms of each language with core surface content mixed but reasoning demand remains constant so as to prevent the contamination from item memory in the within-child comparisons. A mini oral language-proficiency test in each language is included as a covariate.

Qualitative measures include a structured classroom-observation schedule coded for code-switching, teacher scaffolding, and peer-dialogue quality (based on three levels of Mercer, 2000: exploratory, cumulative or disputational talk), surface interpretive accuracy (both HLM and SLL responses), the immediate post-task verbal recall (in the child's preferred language; Ericsson & Simon, 1993), and teacher think-aloud (scoring both HLM and SLL responses for accuracy on the surface level). Table 3

presents a summary of the mapping between constructs, task families, scoring and data sources. Each reasoning construct is associated with particular oral task type and there is a dual scoring scheme; for each reasoning construct there is also an equivalent verbal-recall probe and counterpart that is included in the observational subsample. By aligning, all three types of information (quantitative scores, process data, and classroom evidence) reflect the same competence; the within student contrast (elevated by low performances or vice versa) reflects a single construct; the item-level analysis reflects a single construct; the classroom evidence reflects a single construct.

Instrument development and validation

It is quality of instruments that will comprise the backbone of the design, and Year 1 to a significant extent is dedicated to the development of the instruments themselves. The proposed tasks first come in English and Urdu, then are back-translated independently in the individual languages (Punjabi, Pashto and Sindhi) and reconciled with subject and language experts who are bilingual. Before cognitive pretest questions on the draft cognitive forms, a small sample of children is given the draft form to evaluate its cultural sensitivity and comprehension; the draft form is then piloted for approximately 120–150 children who are not part of the main sample.

The data from the pilot tests are used to estimate item difficulty, item reliability, and preliminary DIF analysis comparing language versions to ensure that there are no items which differ across language versions in reasons unrelated to reasoning that are subsequently revised or removed prior to the main study. The construct-irrelevant-variance analysis is checked by a panel of bilingual expert raters, which compares the quality of reasoning. Scoring is standardized across different raters and is monitored throughout.

Procedures for data collection

Each child completes both language conditions over a two-week window, in counterbalanced order, with assessor and language condition assigned so that order and form are crossed and can be modelled. Spacing the two sessions reduces fatigue and carryover, and the use of parallel forms further guards against practice effects; the post-task verbal recall is conducted after, rather than during, the task to avoid the reactivity that concurrent think-aloud can introduce with young children (Ericsson & Simon, 1993).

Sessions and recalls are audio-recorded with written caregiver consent and transcribed; recall interviews are conducted in the child's preferred language by a bilingual researcher. Wherever feasible, the scoring of recorded responses is conducted by raters blinded to the language condition, to prevent expectancy effects. Classroom observations in the 24 focal classrooms document naturally occurring language practices across several lessons. Teacher think-aloud interviews are conducted individually. Field operations are organized as three provincial teams, each responsible for eight schools, working within a defined data-collection window in Year 2; teams are trained centrally and monitored for protocol fidelity, with regular data-quality checks and secure data handling.

Scoring and procedures for data analysis

Every reasoning item contains two scores: one with full marks if the answer is correct or partial if it is correct; one with no mark if it is incorrect, partial correct solutions and reasoning-process score that represented the presence and quality of the relevant strategy, justification, and self-correction that occurred as the child worked through the task and remembered at the end was coded into answers to the rubric, to reflect reasoning moves, which was developed and refined based on task performance in the Year-1 pilot. Exemplar responses are provided with rubrics, raters are tested for reliability to a criterion, and a percentage of the responses is double coded across the entire assessment to determine inter-rater agreement.

The quantitative analyses estimate the effect of language of assessment on children's reasoning scores based on the nested, repeated-measures structure (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), by fitting children's responses as random effects in multilevel (mixed-effects) models, in which children's reasoning scores are the a priori dependent outcomes and both language condition (home vs. school language) and control variables (oral language proficiency, task form, administration order) are treated as fixed effects. The within-student condition contrast is the most important estimate of construct-irrelevant variance due to the assessment language, and interactions are used to measure differences across regions, proficiency levels, and reasoning constructs. The DIF (item difference) analyses are used to compare the language versions on an item level.

The qualitative analysis is based on the techniques of reflexive thematic analysis (Braun & Clarke, 2006) for the verbal recall and teacher interviews and the technique of discourse analysis for the transcription of the language of classroom observations and is used to classify the functioning of the language of talk and code-switching with regard to the expression of reasoning and teachers' interpretation of responses in the different languages. Integration is convergent and shares joint displays that compare statistical effects to process evidence to help understand a measured score gap in a given province by specific classroom practice and scoring judgement (Creswell & Plano Clark, 2018). The composite product is an account or evidence-based construct validity argument: an estimating of the extent to which teachers and students are misled by the language in their reasoning scores from grades K-5, specification of what this would mean, and an explanation of what this means for the interpretation of and use of the reasoning scores.

A threat tree is used to identify anticipated threats to inferences and is addressed both through design and analysis. Maturation effects over the brief time frame, as well as effects of practice and carryover are reduced through parallel forms and intervals spanning two weeks, and through counterbalancing, and the range of effects is explicitly modelled; within-student effects absorb maturation effects over such a short time. Standardized administration, bilingual assessor training, and recording and having raters score recorded items blinded to language condition reduces assessor and expectancy effects. Selection policy provided for by the school is made transparent and recorded if a student doesn't participate in it. In the qualitative strand, observer effects are minimized by repeated visits during which

classrooms familiarize themselves with the context, while the potential for socially desired responding during the teacher interviews is reduced by focusing on an anchored think-aloud task instead of on the teacher's general beliefs.

The design yields actionable evidence in any direction of results. If achievement in early grade reasoning assessment exhibited a significant home language effect, it would offer direct evidence for assessing the impact of construct-irrelevant variance, with implications for assessment policy. A small difference, when appropriately powered, would offer the same information, namely that in these questions and these ages, the medium of assessment does not seem to be a major obstacle, and shifting of focus from assessment to teaching/item design is needed. Finally, province- and proficiency-level interactions would help clarify the focus of assessment and for whom language responsive assessment is most important. We are upfront about our scope: Our findings concern government schools in three provinces and the reasoning constructs assessed, and the verbal-recall and observational strands are not representative. The boundaries are why integration is so important and statistical estimates have meaning when used in the context of the classroom and scoring evidence that explains the estimates.

Timeline and Milestones

The work proceeds over three years.

Year 1 establishes governance, secures ethical approvals across the three provinces, and is devoted to instrument development in five languages—drafting, independent back-translation, bilingual expert reconciliation, cognitive pretesting, and a pilot with approximately 120–150 children to estimate reliability, verify parallel-form equivalence, and screen items for DIF—culminating in finalized instruments and trained, reliability-certified raters.

Year 2 is the full data-collection year: three provincial teams administer paired oral reasoning sessions and language-proficiency screens to the roughly 600 children across 24 schools, conduct observations in 24 focal classrooms, and complete approximately 80 child verbal recalls and approximately 48 teacher think-aloud interviews, with transcription and data-quality monitoring running in parallel.

Year 3 integrates the quantitative and qualitative analyses and concentrates on output: peer-reviewed publications, conference presentations, the co-produced policy brief and practitioner guidance, and a research-dissemination event with national and provincial assessment actors. A project timeline accompanies the proposal and mirrors these milestones.

Dissemination

Dissemination is planned along scholarly and policy-practice channels and is integral to, not appended to, the research. Scholarly output includes peer-reviewed articles targeting journals in educational assessment, multilingual and comparative education, and Global-South education research, and presentations at major conferences such as AERA, CIES, and UKFIET. To advance open scholarships, key articles will be published open access.

Policy and practice dissemination is anchored in project partnerships. With assessment specialists we will co-produce a concise policy brief translating the construct-validity findings for national and provincial assessment design, timed to inform the ongoing implementation of foundational-learning policy and the design cycles of national and provincial assessments. With provincial partners we will produce practitioner-facing guidance on language-responsive assessment for early-grade teachers and convene a research-dissemination event bringing together assessment officials, provincial education staff, and teacher educators to discuss implications and uptake.

Throughout, the team adheres to the professional and ethical expectations of the field, ensuring that claims are commensurate with evidence and that findings about children's competence are communicated responsibly. The intended impact is tangible: to shift how evaluation systems differentiate intellectual language expertise, so that multilingual children's capability is acknowledged rather than misunderstood.

Conclusions

This study addresses a fundamental question of educational equity and assessment validity: whether tests administered in an unfamiliar school language mismeasure the cognitive reasoning ability of young multilingual learners. By decomposing early-grade reasoning performance into language-attributable and cognitive components through a rigorous within-student design across three diverse provinces in Pakistan, the project establishes the extent and mechanisms of construct-irrelevant variance. The findings highlight the critical need to distinguish language proficiency from intellectual potential in foundational assessments, offering practical and transportable methodologies for multilingual education systems worldwide to design fairer, more valid assessments.

Abbreviations

- **AERA:** American Educational Research Association
- **APA:** American Psychological Association
- **AIU:** Allama Iqbal Open University
- **CIES:** Comparative and International Education Society
- **CLT:** Critical Language Testing
- **DIF:** Differential Item Functioning
- **HLM:** Home Language Medium
- **L1:** First Language / Home Language
- **L2:** Second Language / School Language
- **NAT:** National Achievement Test
- **NCME:** National Council on Measurement in Education
- **PIE:** Pakistan Institute of Education

- **SAQE:** Society for Access to Quality Education
- **SLL:** School Language Medium
- **UKFIET:** The Education and Development Forum
- **UNESCO:** United Nations Educational, Scientific and Cultural Organization

Declarations

Author Contribution

G.A. led the conceptualization and study design, coordinated the research, developed the methodology, and wrote the initial manuscript draft. U.A., A.A., and A.H. contributed equally to refining the methodology, collecting and analyzing the data, interpreting the findings, and critically revising the manuscript. All authors reviewed and approved the final manuscript.

References

1. Abedi, J. (2006). Language issues in item development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Lawrence Erlbaum.
2. AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
3. 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>
4. Ball, J. (2011). *Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in the early years*. UNESCO.
5. Benson, C. (2004). *The importance of mother tongue-based schooling for educational quality*. UNESCO.
6. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
7. Coleman, H. (2010). *Teaching and learning in Pakistan: The role of language in education*. British Council.
8. Coleman, H., & Capstick, T. (2012). *Language in education in Pakistan: Recommendations for policy and practice*. British Council.
9. Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
10. Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, 19, 121–129.
11. Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Multilingual Matters.

12. Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2026). *Ethnologue: Languages of the world* (29th ed.). SIL Global. <https://www.ethnologue.com/country/PK/>
13. Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). MIT Press.
14. Federal Foundational Learning Policy. (2024). *Government of Pakistan*. Ministry of Federal Education and Professional Training.
15. Fiaz, H., Farah, S., & Khan, W. A. (2025). Impact of instruction medium on students' academic performance: A case of university-going students in Karachi, Pakistan. *Journal of Social Horizons*, 2(3), 406–416. <https://doi.org/10.5281/zenodo.17063860>
16. Gallup Pakistan (2024, August 7). *Gallup big data analysis of the 2023 census, Volume 1: A comparative analysis of Pakistan's mother tongues using census data from 1998, 2017 and 2023* [Press release]. <https://gallup.com.pk/post/37156>
17. García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
18. Heugh, K. (2011). *Theory and practice—language education models in Africa: Research, design, decision-making and outcomes*. UNESCO Institute for Lifelong Learning & ADEA.
19. Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum.
20. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
21. Kwok, P. K. P., Thomas, N., Young, C., & Veitch, A. (2024, July 1). *The impact of the language of instruction on foundational literacy and transitions*. UKFIET (The Education and Development Forum). <https://www.ukfiet.org/2024/the-impact-of-the-language-of-instruction-on-foundational-literacy-and-transitions/>
22. Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Multilingual Matters.
23. Mercer, N. (2000). *Words and minds: How we use language to think together*. Routledge.
24. Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach*. Routledge.
25. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
26. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
27. Pakistan Bureau of Statistics. (2023). *Population and housing census 2023*. Government of Pakistan.
28. Pardesi, H. Y. (2022). The role of language policy in nation-building in Pakistan. *Journal of Literature Languages and Linguistics*.
29. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE.

30. Rehman, M., Altaf, S., & Nadeem, M. (2020). The impact of English second language learning on students' identity in Pakistani colleges. *Elementary Education Online*, 19(3), 4857–4879. <https://doi.org/10.17051/ilkonline.2020.03.735639>
31. Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.
32. Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.
33. Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
34. Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
35. Soroya, S. M., Diaz, A. S., Ali, H., Asad, M. I., & Janjua, Z. Q. (2024). *Learning in a familiar language: An impact analysis of the language of instruction on student outcomes and learning experiences in Pakistan*. Pakistan Institute of Education.
36. Thomas, W. P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Center for Research on Education, Diversity & Excellence.
37. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
38. Vygotsky, L. S. (1986). *Thought and language*. Trans.). MIT Press. (Original work published 1934)A. Kozulin.

Tables

Table 1
Conceptual Framework: How Each Tradition Maps onto the Design

Tradition	Core Claim	How the Study Operationalizes It
Vygotsky (1978, 1986): Mediated Cognition	Higher thinking is developed and made visible through language and cultural tools.	Reasoning is elicited orally with pictures and manipulatives, plus post-task recall, so expression is supported rather than blocked.
Cummins (1979, 2000): Interdependence / Common Underlying Proficiency	Reasoning built in one language is available across languages; weak second-language proficiency can mask sound reasoning.	Each child is assessed in both languages; a short oral-proficiency screen is a covariate that separates reasoning from language.
Messick (1989, 1995) & Kane (2013): Construct Validity / Argument-Based Validation	Language unrelated to the construct is construct-irrelevant variance; the use and interpretation of scores have consequences.	The within-student contrast estimates that variance; DIF tests item comparability; the teacher-scoring strand tests consequences.
Shohamy (2001): Critical Language Testing	Tests encode power and confer legitimacy on dominant languages.	The design foregrounds home vs. school language and asks whose reasoning gets counted.

Table 2
Research Questions Mapped to Data, Analysis, and Inference

Question	Data Source	Analysis	What It Establishes
RQ1	Paired oral reasoning tasks in home vs. school language (~ 600 children); oral-proficiency screen	Within-student multilevel models; differential item functioning (DIF)	The size of the language-of-assessment effect and whether items are comparable across languages
RQ2	Classroom observations in 24 focal classrooms	Discourse analysis of talk types; coding of code-switching and scaffolding	How classroom language practice enables or limits the expression of reasoning
RQ3	Teacher think-aloud scoring of paired responses (~ 48 teachers)	Reflexive thematic analysis; comparison of scores across languages	How teachers interpret and, at times, misjudge reasoning across languages

Table 3
Constructs, Task Families, Scoring, and Data Sources

Reasoning Construct	Oral Task Family (Example)	Scoring	Verbal-Recall Probe (Subsample)	Observational Counterpart
Quantitative Reasoning	Mathematical word problems (change/comparison, with picture support)	Product (correct/partial) + reasoning-process score	"How did you work that out?"	Math talk; teacher scaffolding in number activities
Logical Sequencing	Ordering and pattern completion ("what comes next?")	Product + reasoning-process score	"Why does it go in that order?"	Exploratory vs. cumulative talk in sequencing
Analogical Reasoning	Relational mapping, picture-based	Product + reasoning-process score	"Why does that one match?"	