INTERNATIONAL CONFERENCE **Dehli, India, 2026**
https://eijmr.org/conferences/

PERFORMANCE-BASED ASSESSMENT FOR YOUNG LEARNERS: REDEFINING VALIDITY AND RELIABILITY IN CEFR-ALIGNED TESTING

Xoshimova Diyora Azamjon kizi

Trainee lecturer at the Is'hoqxon Ibrat Namangan

State Institute of Foreign Languages

khoshimovadiyora1109@gmail.com

Abstract. In recent decades, the Common European Framework of Reference for Languages (CEFR) has become a cornerstone for language assessment worldwide. While it provides descriptors for communicative competence across proficiency levels, applying CEFR principles to young learners raises unique challenges. Traditional standardized tests often fail to capture children's developing skills authentically, resulting in questions of fairness, motivation, and validity. This study investigates the potential of performance-based assessment (PBA) as a more child-centered approach to measuring young learners' language abilities while maintaining psychometric soundness. Through a mixed-methods design that combined classroom-based performance tasks with statistical analysis of reliability measures, the research examines whether PBAs can provide both valid and reliable outcomes when aligned with CEFR descriptors. Results reveal that PBAs, when carefully designed and scaffolded, enhance construct validity by reflecting communicative competence in authentic contexts. However, ensuring inter-rater reliability remains a challenge, requiring rigorous rubrics and assessor training. The study concludes with pedagogical and policy implications for integrating performance-based assessment into CEFR-aligned testing frameworks for young learners.

Introduction

Language assessment in the twenty-first century is evolving beyond standardized, decontextualized testing toward approaches that value authentic communication. This shift is particularly relevant in the assessment of young learners, whose developmental characteristics, affective needs, and learning styles often clash with traditional testing formats. The Common European Framework of Reference for Languages (CEFR), widely adopted as a benchmark for language proficiency, offers descriptors that emphasize real-life communicative competence. However, when applied to children in primary school contexts, the challenge lies in designing assessment tasks that both engage learners and maintain psychometric rigor. Performance-based assessment (PBA) represents a promising alternative, as it requires learners to demonstrate language use in authentic, task-oriented situations rather than through isolated test items. Such tasks may include role-plays, story retellings, or collaborative problem-solving, all of which mirror the communicative acts described in CEFR can-do statements. Nevertheless, the adoption of PBAs has been hindered by concerns regarding their reliability and scalability. Teachers and policymakers often question whether such assessments can deliver consistent results across learners, raters, and contexts. This thesis addresses the central tension between validity and reliability in performance-based assessment for young learners. Specifically, it examines how CEFR-aligned PBAs can be designed and implemented to capture

children's communicative competence authentically, while ensuring reliability through systematic rubrics and training. The study thus contributes to both theoretical and practical debates in language assessment, offering insights for teachers, curriculum designers, and testing authorities.

Literature Review

The theoretical underpinnings of language assessment are rooted in the balance between validity and reliability. Validity refers to the extent to which an assessment measures what it intends to measure (Messick, 1989), while reliability pertains to consistency of results across administrations and raters. Traditional large-scale tests have prioritized reliability, often at the expense of validity, by employing discrete-item tasks such as multiple-choice questions. These methods provide statistical stability but fail to represent language as social action (Bachman & Palmer, 2010).

The CEFR, with its focus on communicative competence, has challenged such reductive practices by advocating descriptors that describe language use in real-world contexts (Council of Europe, 2001; 2020). While adult learners can be evaluated through complex performance tasks such as oral interviews or written portfolios, the assessment of young learners requires adaptation. Children's shorter attention spans, developing literacy, and sensitivity to affective factors demand assessments that are engaging, meaningful, and developmentally appropriate (Moon, 2005; Cameron, 2001). Performance-based assessment, as conceptualized by O'Malley and Valdez Pierce (1996), seeks to bridge this gap by placing learners in communicative tasks that require integration of skills. Studies by McKay (2006) and Rea-Dickins and Gardner (2000) highlight how PBAs foster positive washback, encouraging learners to view assessment as part of learning rather than as an external imposition. However, critiques of PBAs often target their subjectivity. Inter-rater reliability remains problematic, especially when teachers serve as both instructors and assessors (Fulcher, 2010).

Recent research on young learners has suggested that with clear analytic rubrics, careful task design, and systematic moderation, PBAs can achieve acceptable levels of reliability (Hasselgreen, 2005; Nikolov, 2016). The CEFR Companion Volume (Council of Europe, 2020) also provides descriptors for young learners, though these require operationalization into classroom-appropriate tasks. Thus, the literature suggests a potential but under-explored space for redefining validity and reliability in young learner assessment through performance tasks.

Methodology

This study employed a mixed-methods approach combining classroom-based task implementation with quantitative reliability analysis and qualitative reflections from teachers and learners. The research took place in two primary schools in Uzbekistan, involving 48 learners aged 9–11 at CEFR levels A1–A2. Three performance-based tasks were designed: (1) a role-play based on daily situations (shopping, asking directions), (2) a storytelling activity using picture prompts, and (3) a collaborative problem-solving game. Each task was aligned with CEFR descriptors for spoken interaction and production at A1–A2 levels. Assessment rubrics were developed with analytic scales covering fluency, accuracy, vocabulary use, interactional competence, and pragmatic appropriateness. Four trained raters scored each performance independently. To assess reliability, inter-rater consistency was measured using Cohen's kappa and intraclass correlation coefficients (ICCs). Validity

was examined by mapping task outcomes against CEFR descriptors and triangulating with classroom observations and teacher reflections.

In addition, semi-structured interviews were conducted with six teachers and twelve learners to gather qualitative perspectives on the fairness, engagement, and authenticity of the performance-based assessments.

Results

Quantitative analysis demonstrated moderate to high reliability across raters, with ICC values ranging between 0.72 and 0.83 depending on the task. Role-plays produced the most consistent scores, while storytelling showed more variability due to differences in rater interpretations of narrative coherence. Although reliability was not as high as in standardized multiple-choice tests, results fell within acceptable thresholds for classroom-based assessments. In terms of validity, task performances aligned strongly with CEFR descriptors. For example, learners at A1 demonstrated basic exchanges in role-plays but struggled with sustaining extended speech in storytelling, while A2 learners displayed emerging abilities to negotiate meaning in collaborative problem-solving tasks. These outcomes mirrored CEFR can-do statements, supporting construct validity.

Qualitative data revealed that learners perceived PBAs as more enjoyable and less stressful than traditional tests. Teachers noted that tasks provided richer insights into learners' communicative competence, though they emphasized the challenge of consistent scoring. Some teachers expressed concern over the additional time required for implementation and training.

Discussion

The findings suggest that performance-based assessment offers a viable alternative to traditional testing for young learners in CEFR-aligned contexts. Validity was strengthened by situating assessment in authentic communicative activities that mirrored CEFR descriptors, thus capturing language use more accurately than decontextualized tasks. At the same time, reliability—while slightly weaker than in standardized tests—proved acceptable when analytic rubrics and rater training were employed. This study contributes to the debate on the trade-off between validity and reliability. For young learners, privileging validity through authentic tasks is crucial, as assessments must reflect not only language ability but also developmental appropriateness and affective engagement. Reliability can be safeguarded through systematic rubric design, double-marking, and moderation processes. The qualitative findings further underscore the pedagogical value of PBAs. Children's positive attitudes toward tasks suggest that PBAs can reduce test anxiety and promote intrinsic motivation. For teachers, PBAs provided actionable insights into learners' strengths and weaknesses, informing instructional planning. However, concerns about scoring consistency and administrative feasibility highlight the need for ongoing professional development and institutional support.

Conclusion

This thesis has examined the potential of performance-based assessment to redefine validity and reliability in CEFR-aligned testing for young learners. Findings demonstrate that PBAs, when carefully designed, offer higher construct validity than traditional tests by authentically representing

communicative competence. Reliability, while more challenging, can be achieved through analytic rubrics and rater training. The study highlights the importance of moving beyond psychometric traditions that prioritize reliability at the expense of validity, particularly in young learner contexts where engagement and authenticity are essential. It also points to the need for assessment policies that support teacher training, moderation systems, and integration of PBAs into broader CEFR-aligned frameworks. Future research could extend the scope by examining PBAs across different cultural contexts, exploring longitudinal impacts on learning outcomes, and investigating digital technologies that may facilitate scoring consistency. Ultimately, performance-based assessment offers a promising pathway toward more meaningful and equitable language assessment for young learners.

References.

1. Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford University Press.
2. Cameron, L. (2001). *Teaching languages to young learners.* Cambridge University Press.
3. Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.
4. Council of Europe. (2020). *CEFR Companion Volume.* Council of Europe Publishing.
5. Fulcher, G. (2010). *Practical language testing.* Routledge.
6. Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22*(3), 337–354.
7. McKay, P. (2006). *Assessing young language learners.* Cambridge University Press.
8. Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan.
9. Moon, J. (2005). *Children learning English.* Macmillan.
10. Nikolov, M. (2016). *Assessing young learners of English: Global and local perspectives.* Springer.
11. O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* Addison-Wesley.
12. Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*(2), 215–243.