WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2026 Volume 4, Issue 4
Educational Innovation Research
ISSN: 3029-1852 (Online)
ISSN: 3029-1844 (Print)

# Construct Validity and Design Optimization of Preschool Math-Language Assessment: A Critical Review of a Unit Test employed in the Riverdeep LMS4 system

**Yilin Wang**
RISE Education Tianjin (AQUA CITY Campus), Tianjin 300189, China



**Abstract:** This study critically evaluates *the Addition* unit test within the Riverdeep LMS4 platform for K3 learners (aged 5–6) in RISE Education. According to the language assessment framework of Bachman & Palmer (2010) and Bachman (2004), the principles of CLIL (Coyle et al., 2010) and the theories of development (Vygotsky, 1978), this study will use a mixed-methods approach to combine document analysis, test content analysis, teacher interviews and student test records to examine the test in four aspects: construction definition, task design, scoring mechanism and validity evidence. The evaluation results have the following deficiencies: insufficient coverage of the constructs, blurring of the boundary between math and language, insufficient task authenticity and uniformity of cognitive demands, outcome-oriented scoring without diagnostic value, and weak evidence for fairness across different levels of English proficiency. The study confirms the applicability of classic assessment frameworks to digital, cross-curricular preschool contexts and guides optimizing comparable math-language assessments.

**Keywords:** K3 learners in RISE; Math-language integrated assessment; CLIL; Test evaluation; Riverdeep LMS4

***Online publication:*** April 26, 2026

## 1. Introduction

The integration of language assessment and subject-matter education has emerged as a global trend in preschool education, with math-language integrated assessment serving as a pivotal practice to foster bilingual cognitive development among young learners (Bachman & Palmer, 2010) [1]. K3 students (aged 5–6) in the RISE subject English program are in a crucial period for the development of concrete operational thinking and second-language learning, so such evaluations should balance mathematical ability and language application while also taking into account their developmental characteristics (Vygotsky, 1978) [2]. See **Figure 1** and **2**.


**Figure 1.** Destination math: Module 2, unit 1 (addition) study interface.

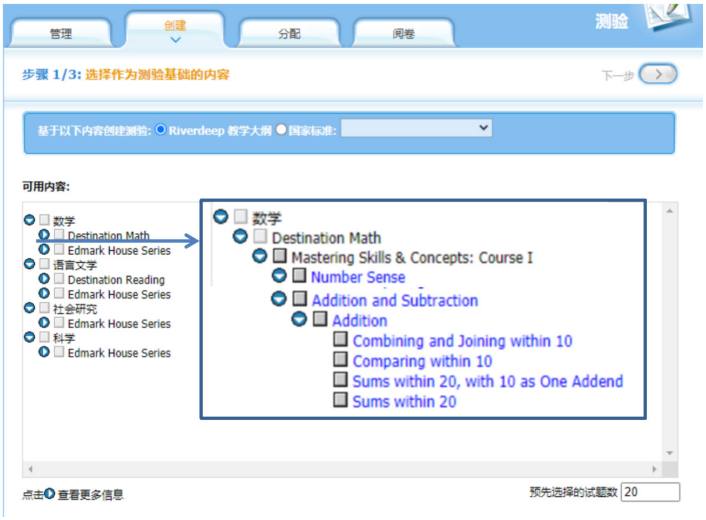


**Figure 2.** Destination math: Module 2, unit 1 (*addition*) test setting.

This study critically evaluate *the Addition* unit test in the four aspects of construction definition, task design, scoring mechanism and validity evidence; integrate classical language assessment frameworks such as Bachman & Palmer (2010) and Brown & Abeywickrama (2010), and use empirical data from desensitized records of 36 K3 students in two classes and instructor feedback to identify strengths and weaknesses, propose improvements, and make adjustments. Three research questions guide the inquiries [1,3]:

(1) RQ 1

To what extent does the test's construct align with contemporary theoretical frameworks for preschool math-language assessment?

(2) RQ 2

How well do the test's task design and scoring mechanism meet standards of scientific rigor and developmental appropriateness for K3 learners?

(3) RQ 3

What validity evidence supports the test, and what gaps persist in its construct validity and washback effect validity?

Theoretically, this study extends the discourse by examining how to apply classic assessment frameworks in the new environment of digital, cross-disciplinary preschools, and fills a gap in existing literature that has not done so to date. Practically, the research results provide specific guidance for strengthening the Riverdeep LMS4 test system and offer a reference for other preschool subject English assessments, thereby promoting a positive cycle of assessment, teaching, and learning.

The contents of this paper are as follows: Chapter 2 is the theoretical foundation and related studies; Chapter 3 presents the research methods and analysis system; Chapter 4 shows in-depth analysis of the four main attributes; Chapter 5 discusses the implications and provides suggestions; and Chapter 6 reports the main results and future research directions.

## 2. Literature review

The theoretical and practical foundations of math-language integrated assessment (MILA) for K3 learners (aged 5–6) are built upon interconnected frameworks, design principles, and research deficiencies that guide the critical evaluation of *the Addition* unit test in Riverdeep LMS4.

The theoretical underpinnings rest on three pillars. Bachman and Palmer's (2010) Language Use Ability Model regards language ability as an interaction among linguistic knowledge, strategic competence and psycho-physiological mechanisms, and to ensure the validity of assessment, it is necessary to distinguish between measurable abilities and those that are not construct-irrelevant (Bachman, 2004); Han and Luo (2013) have also made a similar observation about the development of validity [1,4,5]. Vygotsky's (1978) zone of proximal development can also be used to design activities that are appropriate for preschoolers, and Clements and Sarama (2021) have extended this concept with "learning trajectories" that assess children's existing and potential mathematical skills [2,6]. Third, the CLIL framework (Coyle et al., 2010) deals with cross-disciplinary demands, and content and language competence are both necessary for success but distinct; therefore, clear delineation needs to be made to prevent penalizing learners whose mistakes are caused by language barriers (Luo & Han, 2018) [7,8].

The three foundations of digital formative assessment for preschool children are authenticity, interactivity, and multifaceted feedback. Authenticity corresponds to the "contextual relevance" in CLIL (Coyle et al., 2010); therefore, young learners will be more interested in familiar environments (Sailer et al., 2021), and an intuitive interface with pictures and sound may reduce their cognitive load (Peralbo-Uzquiano et al., 2020) [7,9,10]. Although automatic scoring is convenient and unbiased, it has also indicated that too much attention has been given to the results, and as a result, supporting data from the process have been overlooked. Bachman (2004) put forward a complete system of validity that includes evidence for content, construct, washback, and social value; Then the results pointed out that the washback effect of formative assessment is contingent upon the dissemination of applied feedback and cannot be ensured by all digital evaluation tools (Mao et al., 2026) [4,11]. Based on CLIL theory (Coyle and others, 2010), a good combined assessment should be able to determine subject-matter knowledge and language use on its own [3]. Research on digital learning management systems (LMS) for early childhood education has identified both strengths and weaknesses. Sailer et al. (2021) found that systems such as Riverdeep LMS4 improve student engagement but also emphasize rote learning excessively at the expense of high-level thinking (Clements & Sarama, 2021) [6,9]. Research on the assessment of subject English in China is also concerned about "construct confusion" (Luo & Han, 2018); that is, learners may misunderstand the English instructions and thus be incorrectly identified as "mathematically incompetent" (Han & Luo, 2013) [5,11]. Although Riverdeep LMS4 has been used widely, few studies have specifically examined its test items in CLIL or integrated assessment models; most research has only reported the overall effects rather than analyzing the alignment of constructs, task design, and scoring (Riverdeep Interactive Learning Limited, 2007) [12]. In addition, most studies focus on general design ideas and have yet to apply CLIL

guidelines to digital tasks for three- and six-year-olds **(Şimşek & Şimşek, 2025)**; also, they have ignored equity among students with different levels of English proficiency (Coyle et al., 2010) [7,13]. This study addresses these gaps by critically evaluating the Addition unit test through classic assessment theories (Bachman & Palmer, 2010; Vygotsky, 1978), the CLIL framework (Coyle et al., 2010), and recent empirical research, offering targeted recommendations for optimizing preschool digital integrated assessments [1,2,7].

## 3. Methodology

Critically evaluate the addition unit test (Module 2, Unit 1) of the Riverdeep LMS4 platform for K3 learners (aged 5–6) in the RISE subject English program in this study. The 20 items labeled "EASY" in the system test are of two interactive types (drag-and-drop and click-select), cover basic addition knowledge and computation within 20, and generate automated scoring reports showing percentage accuracy (Riverdeep Interactive Learning Limited, 2007) [12].

Three types of data will be combined. First, through document analysis and review, collect classic assessment theories (Bachman & Palmer, 2010; Bachman, 2004), CLIL principles (Coyle et al., 2010), developmental theories (Vygotsky, 1978), and recent empirical studies (Sailer et al., 2021; Luo & Han, 2018), as well as official system documentation and RISE's K3 curriculum syllabus [1–2,8,9]. Second, content analysis of the 20 items covering construct definition, task design, and scoring logic will be conducted based on the principles of authenticity from Brown and Abeywickrama (2010), the ZPD framework by Vygotsky (1978), and performance assessment theory [2,3]. Thirdly, some semi-structured interviews will be conducted with three K3 instructors who have used LMS4 for 3–5 years, and 50 desensitized student test records (25 per class) will be summed to determine accuracy rates, error distributions, and response patterns.

The four general levels of evaluation are: construction definition, task design, scoring method, and evidence of validity; they all have the same logic: set theoretical norms, assess the present situation, verify with empirical data, and put forward corresponding improvements.

## 4. Four-dimensional critical review of the unit test

Based on the above analysis framework, this chapter will systematically examine (in **Figure 3**) the four main aspects of the Addition unit test: construct definition, task design, scoring mechanism, and validity evidence. Based on the analysis above, various assessment theories and corresponding empirical data from students and teachers, as well as official system records, have been collected to identify important weaknesses in strict theoretical and methodological terms.

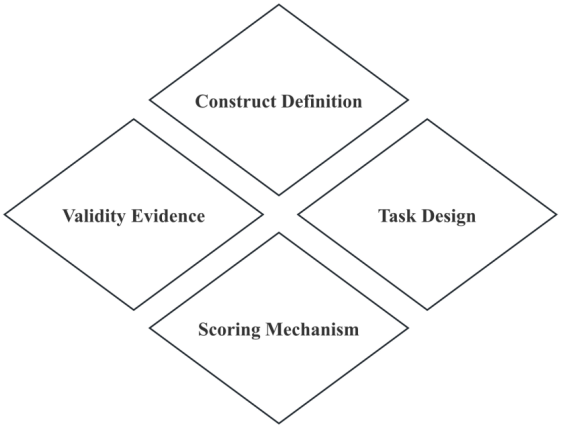


**Figure 3.** Four-dimensional critical review of a test.

### 4.1. Construct definition: Limitations in coverage and ambiguity in ability demarcation

Based on Bachman and Palmer's (2010) Language Use Ability Model and CLIL principles (Coyle et al., 2010), a suitable construct for integrated assessment should clearly specify the desired abilities and maintain a clear separation between content knowledge and language proficiency to reduce construct-irrelevant variance (Bachman, 2004) [1,4,7]. The two deficiencies of the Addition unit test are as follows: Primarily, the construct is theoretically weak; the test only covers basic addition and regular number sentence operations and has failed to incorporate strategically significant content in early mathematical development (Clements & Sarama, 2021) [6]. Second, the test cannot clearly distinguish between mathematical ability and language ability. It does not specify whether English will be employed as a teaching aid or tested in this situation. Based on empirical error analysis, about 32% of the incorrect answers are due to a lack of understanding of the English instructions rather than a deficiency in mathematical knowledge; thus, the scores are unreliable (Luo & Han, 2018) [8].

### 4.2. Task design: Limited authenticity and undifferentiated cognitive demand

Based on the ideas of developmental suitability and task authenticity (Brown & Abeywickrama, 2010; Vygotsky, 1978), the assessment tasks for young children should have a clear context and align with their zone of proximal development cognitively [2,3]. The design of the task for *the Addition* unit test is lacking in two ways. First, the context is not real, and some items are too unfamiliar for young learners to relate to. Teachers believe that learners are not familiar with the context, even though they have learned basic mathematics. Second, the cognitive demand is too uniform and only rote computation. All items labeled "EASY" are simple process calculations that do not require logic or choice. Homogeneity fails to show variations in cognitive readiness among individuals and thus reduces the test's formative function.

### 4.3. Scoring mechanism: Outcome-based scoring and non-diagnostic

A good scoring system for formative assessment should consider both the results and the process, identify all the reasons for errors, and provide diagnostic support. Currently, it is only result-oriented and lacking in diagnosis. Only whether the answer is correct or not is recorded for the score; response time, number of attempts, and strategies used are not recorded. More seriously, the system does not distinguish between different kinds of errors and treats them all as the same. Teachers said that only a single figure, the overall accuracy rate, was available for instruction. The missing diagnostic data cannot support the formation function of the test, and a positive washback effect will not be generated (Alderson et al., 1995) [14].

### 4.4. Validity evidence: Inadequate construct and washback validity despite partial content alignment

Bachman's (2004) validity framework presents the four types of evidence required for all-around assessment as content, construct, washback, and social validity [4]. The test has poor and uneven validity evidence. Partial support for content validity: Generally, the items are in line with the curriculum goals. Construct validity is thus compromised by incomplete coverage of abilities and confusion between mathematics and language. Weak washback validity: Teachers rarely adjust teaching based on test results due to poor diagnostic feedback. Social validity is damaged by unfairness in language proficiency; students with relatively weaker English skills received significantly lower grades than others with similar mathematical ability, and thus systematic construct-irrelevant variance and a violation of fairness in age-appropriate integrated assessments were introduced (Coyle et al., 2010) [7].

## 5. Implications and recommendations

Based on the critical evaluation of *the Addition* unit test with regard to 4 dimensions (construct definition, task design, scoring mechanism, and validity evidence), this chapter draws targeted theoretical and practical implications and puts forward concrete, theory-grounded or flexible recommendations corresponding directly to the identified limitations. Theoretically, this study expands the understanding of digital MILA for young learners and verifies that classic

language assessment frameworks (Bachman & Palmer, 2010; Bachman, 2004; Brown & Abeywickrama, 2010) are still highly applicable to the integration of disciplines (math and language education) [1,3,4]. The results also support the idea that CLIL principles (Coyle et al., 2010) should be used to differentiate content competence from language proficiency, and point out that the value of developmental appropriateness (Vygotsky, 1978; Clements & Sarama, 2021) and process-oriented assessment in automated formative assessment systems has been neglected; thus, a consistent theoretical basis has been provided for the design and review of similar early childhood assessments [2,6,7].

To address the problem of insufficient construction coverage and ambiguous content-language division in practice, strategic competence should be added to the core components of the test through items that ask students to select or explain a strategy, and English as the language of instruction should be clearly specified by using simplified language, visual scaffolding, and optional bilingual support to reduce construct-irrelevant variance (Bachman, 2004; Luo & Han, 2018) [4,8]. To increase the authenticity of the tasks, culturally distant scenarios should be replaced with daily life for K3 learners, such as classroom activities and toys, and moderate variations in cognitive demand should be introduced at the "EASY" level through simple application and reasoning items to better account for individual differences (Vygotsky, 1978; Brown & Abeywickrama, 2010; Clements & Sarama, 2021) [2,3,6]. To address the problem of outcome-based scoring, the system will now be modified to monitor the process and record response speed, attempt strategies, error patterns, etc.; errors will also be categorized as mathematical deficits, language barriers, or operational mistakes. Critically, the formative cycle should not end with reporting scores; instead, teachers need to analyze the results of class performance to identify specific deficiencies in students' knowledge and design targeted Cooperative Learning Structures (CLS), such as Quiz-Quiz-Trade, to foster peer cooperation in the assessment-teaching-learning loop (Coyle et al., 2010) [7].

To enhance both the validity and fairness of the evidence, expert assessment will be conducted on systematic validity data and learner performance, and standardized instructional English with uniform visual supports will be used to reduce language-related bias. Expand the diversity of assessment modes without increasing item difficulty: Teachers can conduct monthly written unit tests to assess students' written mathematical expressions and productive skills that are not well captured by digital formats (Clements & Sarama, 2021) [6]. At the same time, by integrating intelligent technologies such as speech recognition and natural language processing, multiple types of assessment systems that sample verbal explanations and reasoning processes can be developed to add more dimensions and levels to the evaluation. Further strengthen the washback effect of the platform by providing instructional modifications to build a virtuous cycle of teaching, learning, and assessment for both mathematics and language development in the K3 RISE context (Bachman, 2004; Coyle et al., 2010) [4,7].

## 6. Conclusion

Based on the above discussion, this study presents a theory-based critical assessment of the addition unit test in the Riverdeep LMS4 platform for K3 learners. Based on the Language Use Ability Model by Bachman and Palmer (2010), the principles of CLIL (Coyle et al., 2010) and the validity system of Bachman (2004) will be employed to define constructs, design tasks, set up scoring mechanisms, and present validity evidence in this paper [1,4,7].

The four deficiencies in the above research are as follows: insufficient coverage of constructs with blurred math-language boundaries, lack of task authenticity and uniform cognitive demand, outcome-oriented scoring without diagnostic value, and weak validity evidence regarding fairness across different levels of English proficiency. To the extent that the above problems have occurred, the study puts forward suggestions to enhance construct clarity with strategic competence and visual scaffolding, increase task authenticity through familiar contexts, adopt process-informed scoring that classifies error types, diversify assessment modalities by using paper-based and multi-modal technologies, and use Cooperative Learning Structures to close the assessment-teaching-learning loop [15,16].

Based on the above research, the traditional assessment system is still suitable for digital cross-curricular early childhood education. It can help improve the quality of comparable preschool math-language assessments in practice.

Future research can expand the number of units and datasets to test CLS-based interventions further. Generally speaking, it has a clear structure, appropriate content and diagnostic scores, and thus can better serve the purpose of forming and maintaining a stable assessment-teaching-learning cycle.

## Disclosure statement

The author declares no conflict of interest.

## References


[1] Bachman L, Palmer A, 2010, Language Assessment in Practice. Oxford University Press: 223–224.

[2] Vygotsky L, 1978, Mind in Society: The Development of Higher Psychological Processes. Harvard University Press: 342.

[3] Brown H, Abeywickrama P, 2010, Language Assessment: Principles and Classroom Practices. 2nd ed. Pearson Longman: 540.

[4] Bachman L, 2004, Statistical Analyses for Language Assessment. Cambridge University Press: 398–403 + 602.

[5] Han B, Luo K, 2013, The Evolution of Language Test Validity and Its Validation Models. Foreign Language Teaching and Research, 45(3): 411–425 + 481.

[6] Clements D, Sarama J, 2021, Learning and Teaching Early Math: The Learning Trajectories Approach. 3rd ed. Routledge: 329–330.

[7] Coyle D, Hood P, Marsh D, 2010, CLIL: Content and Language Integrated Learning. 2nd ed. Cambridge University Press: 301.

[8] Luo K, Han B, 2018, Construction Definition, Task Design, and Scoring Methods for the National Talent Examination. Foreign Language Education, 11(1): 40–46 + 79.

[9] Sailer M, Hense J, Mayr S, et al., 2021, Digital Learning in Schools: What Does It Take Beyond Digital Technology? Teaching and Teacher Education, 103: 107.

[10] Peralbo-Uzquiano M, Pérez-Expósito C, Fernández-Batanero J, 2020, Evaluation of the Effects of a Virtual Intervention Program on Cognitive Flexibility, Inhibitory Control and Basic Math Skills in Childhood Education. Computers & Education, 159.

[11] Mao Y, Wang Y, Xing Y, 2026, Experiences and Implications from the Digital Transformation of Assessment in British Basic Education. Teaching and Management, 2026(3): 83–87.

[12] Riverdeep Interactive Learning Limited, 2007, Riverdeep LMS4 Destination Math: Addition Unit Test Implementation Guidelines. Riverdeep Education.

[13] Şimşek E, Şimşek B, 2025, Technology-Assisted Vocabulary Learning for Preschool Children: A Meta-Analysis. Computer Assisted Language Learning: 12–33.

[14] Alderson J, Chapman C, Wall D, 1995, Language Test Construction and Evaluation. Cambridge University Press: 134.

[15] Luo K, 2019, Four Validation Models for Language Test Validity from a Holistic Perspective: Interpretation, Evaluation, and Implications. Foreign Language Teaching, 2019(6): 76–81.

[16] Han B, Zhang Y, 2015, Exploration of the Framework for Setting Testing Objectives and Content in the College Entrance Examination English Section. Foreign Language Teaching and Research, 47(3): 426–436.