RESEARCH SHOP

"know what's out there!"

Comparing Literacy and Numeracy Assessment Tools for Youth

Prepared for

Beyond the Bell

In

April 2023

By

Rebecca Correia
Megan Devoe
Tony Nguyen
Maria Durrani
Amna Kataria
Evan Gravely

rshop@mcmaster.ca researchshop.mcmaster.ca

Office of Community Engagement | McMaster University

Contents

Executive Summary

“Beyond the Bell” is an after-school and summer academic program with educational components focused on literacy and numeracy skills. Program staff regularly assess academic performance of youth enrolled in “Beyond the Bell,” but there are concerns about the continued use of assessment tools, their ease of use, and effectiveness. Therefore, we aimed to review evidence-based tools to assess youth’s literacy and numeracy skills and compare these alternative tools on several criteria.

This project was divided into two phases. First, we conducted an online search to identify evidence-based assessment tools. We collected relevant information about these tools using a variety of data sources and summarized the tools. In the second phase, we compared each tool using a set of criteria (i.e., timeliness, resource-intensiveness, duration to administer, cost, user-friendliness, quality of the output, validity, and reliability). We provided recommendations based on our comparison to inform “Beyond the Bell” leadership in selecting an appropriate tool. Our findings are limited by the scope of our online search and our subjective consensus-building process to compare tools.

We identified 13 texts that presented 12 assessment tools. Three tools assess literacy skills, two measure numeracy performance, and seven assess both literacy and numeracy skills. In our comparison, we identified variability in the strengths and limitations of tools across the criteria. We selected PM Benchmark Reading Assessment-2 as the strongest literacy assessment tool because many resources are available to educators through the subscription, the tool produces an informative output, and it has demonstrated high validity and reliability in testing. We recommend Mathematical Virtual Learning Environment as the best numeracy assessment tool due to minimal resources required to administer the assessment, high user-friendliness, and established validity and consistency. Lastly, we selected Student Learning Assessments as the greatest comprehensive tool because the assessment can be conducted over time, the output is highly interpretable and offers a parent dashboard, and it is currently used across the province of Alberta.

Overall, our suggestions are preliminary, and “Beyond the Bell” needs to investigate our shortlist of tools for themselves and select one that best fits their context.

Introduction

Context

"Beyond the Bell" is a fully funded after-school and summer academic program that aims to close the academic achievement gap experienced by students from low-income communities (YMCA, 2023). The educational components primarily focus on areas where children are most likely to fall behind: literacy (e.g., reading, writing, comprehension, speaking) and numeracy (e.g., math and number recognition). The program is offered free of charge and increases access to supports that help youth keep pace with others in their grade level.

Program staff regularly assess changes in literacy and numeracy skills of youth enrolled in "Beyond the Bell" by using the following tools:

- Dr. Fry's Informal Reading Assessment is currently used to assess the literacy skills of youth enrolled in "Beyond the Bell." For grade one students, the assessment involves letter recognition and associated skills. For older students, the tool is administered as two progressive tests in October and May, where students keep reading until they make errors. They then get a score based on where they stopped reading without errors. This tool is administered one-on-one, making it resource intensive.
- To assess numeracy skills, Prodigy Math is the core assessment tool used. Prodigy is an electronic game with avatars where students go through a series of questions evaluating different skills (e.g., telling time and interpreting ratios) until each skill is mastered. The game has a reporting feature, can be administered to multiple students simultaneously using iPads, and is an ongoing experience from September to May.

For more than 13 years, "Beyond the Bell" staff have used these tools to measure literacy and math skills. This has raised concerns about whether these tools are still accurate, easy to use, and effective. Technology requirements are a further concern with the current tools.

Study Objectives

This research project aimed to review and update assessment tools for "Beyond the Bell" to assess the literacy and numeracy skills of youth enrolled in the program. The specific objectives that were determined in consultation with our community partner included:

1. Conduct a review of evidence-based tools to effectively assess youth's literacy and numeracy skills (in grades 1 to 5) attending "Beyond the Bell."
2. Collect and analyze information about these alternative tools to compare them based on several criteria (e.g., validity, reliability, time to administer, cost).

3. Propose a shortlist of tools for “Beyond the Bell” to consider for evaluating their program.

Scope

“Beyond the Bell” comprises four key program areas: realizing academic potential, improving health and wellness, exploring culture and creativity, and developing social skills. This project focuses on skills attainment pertaining to academic achievement. Therefore, we did not examine how to assess or evaluate other components of the program, like recreation, nutrition, or socialization activities.

Report Structure

We begin by describing the methods used to identify studies examining assessment tools and then report summaries of the tools we identified in our review. We discuss the strengths and limitations of each tool according to our comparison criteria. Where possible, we present our results using conceptual diagrams, tables, and figures. Finally, we conclude by synthesizing the identified tools and providing recommendations based on our comparison to inform “Beyond the Bell” leadership in selecting an appropriate tool.

Methodology

Overview

This project was divided into two phases. First, we conducted an online search to identify evidence-based assessment tools that measure youth’s literacy and numeracy skills in grades 1 to 5. We collected relevant information about these tools using a variety of data sources. In the second phase, we established comparison criteria to guide our data collection. We compared each tool using these criteria, which we used to propose a shortlist of tools in our discussion.

Phase 1: Online Search

Search Strategy

We began by searching for academic and grey literature using well-defined terms. Our search terms aligned with the research objectives by specifying the study population (i.e., youth in grades 1-5), setting (i.e., after-school programs), and intervention (i.e., literacy and numeracy tools). We specified alternative ways to state these terms by identifying keywords and synonyms. We combined the complete set of search terms using Boolean operators (e.g., OR, AND) (Figure 1).

[child* OR student* OR kid* OR youth OR pupil]	AND	[primary OR elementary OR grade 1 OR grade 2 OR grade 3 OR grade 4 OR grade 5 AND school] OR [after school OR after-school OR after hours OR after-hours]	AND	[literacy OR reading OR writing] OR [numer* OR math OR arithmetic]	AND	[assess* OR exam* OR evaluat* OR instrument* OR test* OR tool* OR educat* OR quiz]

Figure 1. Search Terms Applied in Online Search

We selected seven databases to search for academic and grey literature. We searched large academic databases (Ovid, Web of Science, ERIC, and PsycINFO), the McMaster library catalogue, Google Scholar, and conducted a traditional Google search. We expected the academic databases to identify peer-reviewed journal articles and theses (e.g., experimental studies testing the tools). In contrast, the web search would yield educational resources and curriculum documents (e.g., guidance from school boards).

We extracted the titles, abstracts, and hyperlinks of articles from each database into our spreadsheet for data collection. Due to time constraints, we restricted our review to the first 10 articles listed in the search results. This process resulted in 70 titles and abstracts we screened for relevance.

Inclusion/Exclusion Criteria

We developed a set of inclusion and exclusion criteria aligned with our study objectives (Table 1).

Table 1. Inclusion and Exclusion Criteria for Online Search

Inclusion Criteria	Exclusion Criteria
• The tool should be appropriate for youth in grades 1 to 5. • The study should have been published within 15 years (2008 to 2023). • The tool must evaluate literacy (reading and/or writing) OR numeracy (math) capacity. • The tool should be flexible enough to be adapted to different classroom or after-school settings. • The study should be published as a report, article, or thesis.	• The study was published outside of Canada or the United States. • The study was published in a language other than English. • The tool measures skills other than literacy or numeracy (e.g., social or behavioural skills). • The tool applies to a limited youth group (e.g., gifted students). • Blogs, conference abstracts, or research protocols.

23 texts aligned with our eligibility criteria and were included for full-text review. The texts were published in various forms, including peer-reviewed articles, reports, theses, and government websites.

We then reviewed the texts that initially aligned with our eligibility criteria. Upon further review, we excluded ten additional texts that did not discuss assessment tools related to our study objectives. Therefore, we ended up with 13 included texts for data extraction (Table 2).

Table 2. Overview of Literature Screening Process

Database	Results	Screening Results and Reasons for Exclusion
Ovid	10	INCLUDED: 2 texts
		EXCLUDED: 8 texts • Wrong setting (e.g., home school, outside Canada/US) • Wrong population (e.g., kindergarteners) • Does not assess literacy or numeracy skills
Web of Science	10	INCLUDED: 2 texts
		EXCLUDED: 8 texts • Outdated • Wrong setting (e.g., outside Canada/US) • Wrong population (e.g., kindergarteners) • Does not assess literacy or numeracy skills
ERIC	10	INCLUDED: 1 text
		EXCLUDED: 9 texts • No mention of a specific assessment tool • Wrong population (e.g., focused on migrant children who are English learners) • Does not assess literacy or numeracy skills
PsycINFO	10	INCLUDED: 3 texts
		EXCLUDED: 7 texts • Wrong setting (e.g., outside Canada/US) • No mention of a specific assessment tool • Does not assess literacy or numeracy skills • Unable to retrieve the full text
McMaster library catalogue	10	INCLUDED: 0 texts
		EXCLUDED: 10 texts • Outdated • Wrong population (e.g., gifted learners) • Does not assess literacy or numeracy skills
Google Scholar	10	INCLUDED: 3 texts
		EXCLUDED: 7 texts • Unable to retrieve the full text • No mention of a specific assessment tool • Wrong setting (e.g., outside Canada/US) • Does not assess literacy or numeracy skills
Google search	10	INCLUDED: 2 texts
		EXCLUDED: 8 texts • No mention of a specific assessment tool • Not a scientific resource (e.g., blog post)
TOTAL	**70**	**13 texts aligned with our eligibility criteria for inclusion**

Data Extraction

We documented the title, author, publication year, and reason for inclusion/exclusion for each text we screened. For included texts, we extracted the publication type, setting, study design/methodology, objective/research question, methodological strengths and limitations, and the name of the tool that was discussed.

For each tool that was presented in an included text, we extracted information about:

- Whether the tool assessed numeracy skills, literacy skills, or both;
- Grade level(s) the tool was appropriate for;
- Where the tool was developed; and
- How the tool was administered.

The above information was primarily obtained from the included studies identified in our online search. In some cases, we conducted an additional Google search to collect more information about the specific tool.

Phase 2: Comparing the Tools

Comparison Criteria and Data Extraction

We developed our comparison criteria iteratively and collaboratively with the community partner. In our project kickoff meeting, we facilitated a discussion with the community partner about aspects of assessment tools that would be important to investigate as part of the comparison. This consultative process resulted in eight criteria we used to guide our data extraction:

- **Timeliness:** When was the assessment tool developed?
- **Resource-intensiveness:** What resources are required to administer the tool? (e.g., staff, technology, educational requirements of staff)
- **Duration:** How long does it take to administer the tool? Does the tool require repeated assessments?
- **Cost:** What is the cost to purchase the tool or conduct each assessment?
- **User-friendliness:** What is the tool's ease of use?
- **Output:** What is the complexity of the output? (e.g., scores, graphics, tables, figures)
- **Validity:** What is the tool's validity? (e.g., how well a student's assessment score measured by "Beyond the Bell" staff aligns with their in-school performance)
- **Reliability:** What is the tool's reliability? (e.g., would re-assessing the same student multiple times result in the same score?)

When reading the full texts of included literature, we took note of information relating to each criterion in a table (presented in the findings).

Approach to Rating the Tools

Once we had extracted relevant information for our comparison criteria, our team came together in a synchronous meeting to compare the tools. We had discussions about each tool's strengths and limitations, which resulted in us developing a shortlist of

assessment tools that ranked highly and poorly across the criteria, which informed our recommendations for the community partner.

Limitations

First, our search was not comprehensive, as we only reviewed the first ten search results from each database. This approach was intentional based on time limitations for this project. We favoured breadth over depth by reviewing fewer results from each database, but searching multiple databases.

Second, some results obtained in our search were inaccessible for full-text review. Some texts obtained from PsycINFO and Google Scholar – mainly scholarly theses – could not be retrieved or were only available behind paywalls.

Third, our evaluation of the tools was based on secondary information from articles and studies about those tools, which may have incorporated some biases. For example, subjective reports of a tool's user-friendliness may not be generalizable to other users. Similarly, the time required to administer the tool may not be accurate for all users. Furthermore, we evaluated the tools through a consensus-building process within our team. We did not have the time, knowledge, or resources to develop objective scales to rate each criterion. Obviously, this means our evaluation of the tools was a subjective process informed by the opinions of team members rather than a more rigorous method. As such, our evaluation is only meant to serve as a springboard for thoughtful consideration by the community partner in selecting a tool that best fits their context. We consulted the community partner about this process to ensure this approach satisfied their expectations.

Findings

Phase 1: Online Search Findings

Overview of Included Literature

We identified 13 texts that met our inclusion criteria. Most texts were based in the United States (n=10), Canada (n=2), or elsewhere (n=1). The year of publication ranged from 2007 to 2022, with most published in the early 2010s. The publication types included seven peer-reviewed articles, four theses, and two government websites. The methodology of the peer-reviewed articles included four experimental studies, one observational study, one qualitative study, and one systematic review. Of the theses, three were observational studies, and one was experimental. We summarize these included studies in Table 3 and Table 4.

Table 3. Overview of Included Texts

Title (Publication Year)	Author	Setting	Objective or Research Question	Publication Type (Methodology)
Is it working? An overview of curriculum-based measurement and its uses for assessing instructional, intervention or program effectiveness (2007)	Cusumano, DL, et al.	United States	Is Curriculum Based Measurement effective inside and outside school settings?	Peer-reviewed article (Experimental study)
Promoting Understanding of Measurement and Statistical Investigation Among Second-Grade Students with Mathematics Difficulties (2022)	Doable, CT, et al.	United States	Does PM-2 improve the math achievement of second grade students?	Peer-reviewed article (Experimental study)
A randomized experiment of a mixed-methods literacy intervention for struggling readers in grades 4–6: effects on word reading efficiency, reading comprehension and vocabulary, and oral reading fluency (2010)	Kim, JS, et al.	United States	Does READ 180 improve reading literacy and comprehension?	Peer-reviewed article (Experimental study)
Trajectories of Math and Reading Achievement in Low Achieving Children in Elementary School: Effects of Early and Later Retention in Grade (2013)	Moser, SE, et al.	United States	To investigate the effects of retention vs promotion in first grade influence trajectories of achievement scores in math and reading through fifth grade	Peer-reviewed article (Observational study)
Item-level and construct evaluation of early numeracy curriculum-based measures (2012)	Lee, YS, et al.	United States	To extend research on valid early mathematics screening measures by introducing new measures, modifying existing ones, and replicating criterion validity studies with additional students.	Peer-reviewed article (Experimental study)
The design, implementation and evaluation of a desktop virtual reality for teaching numeracy concepts via virtual manipulatives (2013)	Daghestani, L	Saudi Arabia	Applying virtual reality manipulatives as tools in teaching addition and subtraction to elementary students and assessing their ability to connect concrete and abstract numeracy concepts.	Thesis (Experimental Study)
Assessing numeracy in the upper elementary and middle school years (2015)	Gittens, CA, et al.	United States	Introduce a scale for assessing numeracy in an applied form of critical thinking and enable educators to	Peer-reviewed article (Qualitative study)

			optimize math curriculum design and evaluation to enhance student achievement.	
Organizational citizenship behaviors, collective teacher efficacy, and student achievement in elementary schools (2010)	Jackson, JC	United States	To describe the relationship between teacher OCB, Collective Teacher Belief Scale (CTE) and Student Achievement (SOL) from a sample of Virginia elementary schools	Thesis (Observational study)
Elementary school size and academic performance: A multi-year study (2010)	Zoda, PR	United States	To determine the effects of school district size on the academic performance of Texas students	Thesis (Observational study)
An investigation into the relationship between cognitive ability, standardized achievement, and grades in middle school (2009)	Blue, LT	United States	To examine the relationship between cognitive ability and measure using standardized achievement scores on the New Jersey Assessment of Skills and Knowledge (NJ ASK) and school grades	Thesis (Observational study)
Evidence-Supported Interventions Associated with Black Students' Education Outcomes: Findings from a Systematic Review of Research (2018)	Same, MR et al.	United States	To examine what interventions have shown to be associated with improved academic achievement of Black students according to evidence tiers I (strong evidence), II (moderate evidence), and III (promising evidence) from Every Student Succeeds Act (ESSA)?	Peer-reviewed article (Systematic review)
Student Learning Assessments (2020)	Government of Alberta	Canada	To improve student learning and help identify students' strengths and areas for growth	Government website
Assessments and Evaluation (2022)	Government of New Brunswick	Canada	To improve teaching and learning and to keep the public informed about the educational system's general health	Government website

Table 4. Summaries of Included Texts

Title (Publication Year)	Summary
Is it working? An overview of curriculum-based measurement and its uses for assessing instructional, intervention or program effectiveness (2007)	Cusumano (2007) investigated a comprehensive, widely accepted tool that is available at low or no costs, which aligned with their objective of achieving proficiency in core academic areas across students of all genders, ethnicities, socioeconomic statuses, and dis/abilities.
Promoting Understanding of Measurement and Statistical Investigation Among Second-Grade Students with Mathematics Difficulties (2022)	Christian et al. (2022) conducted a randomized-control trial of students with mathematical difficulties, in which they introduced an intervention involving learning activities and assessments, to monitor achievement. A large sample size strengthened their study, but their participants predominantly white. Also, the study period was too brief to properly implement and develop skills for adequate measurement.
A randomized experiment of a mixed-methods literacy intervention for struggling readers in grades 4–6: effects on word reading efficiency, reading comprehension and vocabulary, and oral reading fluency (2010)	Kim et al. (2010) assessed whether multi-component literacy activities improved attitudes toward literacy instruction and academic achievement. The intervention included various activities to improve reading efficacy and comprehension, including a mix of teacher-led and computer-assisted learning. Missing data for 14 students may have biased their findings.
Trajectories of Math and Reading Achievement in Low Achieving Children in Elementary School: Effects of Early and Later Retention in Grade (2013)	Moser et al. (2013) led a longitudinal study to investigate the long-term effects of retention in grade one among low-achieving students on their trajectories in grade five for math and reading. The long observation window allowed the researchers to evaluate the effects of retention over time, and many baseline factors were collected about the study population. The findings may have been affected by missing data that occurred in later years (e.g., if participants withdrew from the study over time or changed schools).
Item-level and construct evaluation of early numeracy curriculum-based measures (2012)	Lee et al. (2012) conducted an experimental study among 137 kindergarten and first grade students. The study aimed to replicate and add to existing research on reliable and valid early mathematics screening measures and introduce new measures. Another objective was to modify existing measures to examine the theory that early mathematics measures might contribute to the construct of early numeracy proficiency. The study addressed several CBMs but included only a small sample size of fewer than 70 students at each grade level.
The design, implementation and evaluation of a desktop virtual reality for teaching numeracy concepts via virtual manipulatives (2013)	Daghestani (2013) tested a virtual reality assessment in an experimental study. The objective was to apply virtual reality manipulatives as cognitive tools to determine whether second grade students can make connections between concrete and abstract

	numeracy concepts. The navigation feature of VR had a positive effect on students' conceptual understanding of numeracy concepts.
Assessing numeracy in the upper elementary and middle school years (2015)	Gittens et al. (2015) introduced a quantitative scale to assess numeracy skills as an applied form of critical thinking among children. The assessment allowed educators to guide mathematics curriculum design and evaluation to maximize student achievement. The intervention positively impacted creative problem-solving and mental focus to help students acquire numeracy skills. The main limitation was a lack of ethnic diversity in the sample.
Organizational citizenship behaviours, collective teacher efficacy, and student achievement in elementary schools (2010)	Jackson (2010) used a quantitative correlational study to describe student achievement from a sample of Virginia public elementary schools. The study was limited by a small sample size and did not represent the region.
Elementary school size and academic performance: A multi-year study (2010)	Zoda (2010) conducted a non-experimental, quantitative, causal-comparative quantitative research design to determine the effect of school district size on the academic performance of Texas students.
An investigation into the relationship between cognitive ability, standardized achievement, and grades in middle school (2009)	Blue (2009) used a cross-sectional study design to examine the relationship between cognitive ability, measures, and school grades. The study used standardized tools to assess students' cognitive ability and academic performance. The main limitation of the study is the small sample size. Most of the students were white and from privileged backgrounds.
Evidence-Supported Interventions Associated with Black Students' Education Outcomes: Findings from a Systematic Review of Research (2018)	Same et al. (2018) conducted a systematic review to understand what interventions are associated with improved academic achievement among Black students. The study was strengthened by the extensive number of databases, abstracts, and studies reviewed. The main limitation was that the list of interventions was not exhaustive, and the different types of interventions may not have demonstrated similar associations in all settings.
Student Learning Assessments (2020)	The two government resources we identified (Governments of Alberta and New Brunswick) simply commented on teaching, learning, and assessment plans to inform the public about the provincial education system. Both sources listed assessment tools that were appropriate for different grade levels; we extracted the names of tools that were relevant to our study population.

Included Tools

We identified 12 assessment tools from our included texts. This number is smaller than the number of included texts because two texts discussed the same tool. The tools included:

1. Curriculum Based Measurement (CBM)
2. PM Benchmark Reading Assessment-2 (PM-2)
3. Read180
4. Woodcock Johnson-III Tests of Achievement (WJ-IV)
5. Mathematical Virtual Learning Environment (MAVLE)
6. Numeracy Scale Insight Assessment
7. Virginia Standards of Learning (SOL) exams
8. The State of Texas Assessments of Academic Readiness (STAAR)
9. The Partnership for Assessment of Readiness for College and Careers (PARCC)
10. Elementary School Success Profile Model of Assessment and Prevention (ESSP MAP)
11. Student Learning Assessments (SLA)
12. Early Grades Literacy Assessment (EGLA)

In the following sub-sections, we will briefly describe each tool, organized by whether the tool assesses literacy skills, numeracy skills, or both.

Tools to Assess Literacy Skills

PM-2 can be administered to students in grades 1 to 5 to assess reading abilities. The assessment is administered one-on-one and involves students reading passages of increasing difficulty to advance to more challenging levels. Students read the texts aloud while the instructor records their performance. The instructor then analyzes the data by referencing a guide to determine the student's reading level. The student's reading level is inputted into "EdPlan Insight," which uses information to develop instructional planning based on each student's performance.

Read180 is a blended instructional model for students in grades 3 to 5, with instructional software that provides intensive, individualized skills practice. The four areas of focus in the instructional software are: reading zone, word zone, spelling zone, and the success zone. Read180 involves independent reading of audiobooks and paperbacks of increasing levels of difficulty. Students practice their language fluency and reading comprehension skills with a virtual reading coach. The assessment is delivered through instructors where students are taught in group and independent instruction. Instructors use software to input data and access reports for individual students, classes or schools. The output for instructors can be complex as the program includes multiple components and measures of students' progress.

EGLA is a literacy tool applied in more than 60 countries and is currently being piloted in New Brunswick schools for students in kindergarten to grade 2. ELGA measures five foundational reading skills (i.e., phonological awareness, phonics, fluency, vocabulary, and comprehension) through 15-minute individual oral assessments. The output is oral

interviews that assessors must be able to listen to, code, and monitor elapsed time simultaneously. All assessors must code students' responses in the same way to ensure that data consistently reflect collective standards for reading abilities, rather than assessors' personal evaluations of skills.

Tools to Assess Numeracy Skills

MAVLE is an online application to assess numeracy skills among students in grades 1 to 5. The application offers a three-dimensional (3D) interactive environment where students manipulate objects using a mouse and keyboard. The students themselves create the exercises, and they can choose the level of difficulty. The output for instructors is easy as teachers are able to use a virtual manipulative program of their choosing and directly incorporate into numeracy and math lessons and target areas for improvements.

Numeracy Scale Insight Assessment enables instructors to guide mathematics curriculum for students in grades 1 to 5. The assessment encourages students to represent and interpret data, solve problems, engage in tasks requiring measurement, and estimate and interpret numerical expressions through a series of evaluation questions. The assessment is delivered in-person through text booklets or by online software where students are given 75 minutes to complete the assessments. The output is easily understood as support materials and instructions are provided.

*Tools to Assess **both** Literacy and Numeracy Skills*

CBM assesses numeracy and literacy skills and is appropriate for students in grades 1 to 5. CBM is useful for progress monitoring and identifying students at risk of not meeting curriculum objectives. It is recommended to administer the tool three times per year (e.g., fall, winter, and spring), as the results can inform decisions about further instruction and allow instructors to evaluate progress over a sequence of assessments.

WJ-IV assesses the strengths and weaknesses of test-takers in key cognitive, academic, and linguistic areas. The tool is appropriate for anyone over the age of two. WJ-IV consists of three separate tests that can be administered independently or in combination: (i) Test of Achievement (WJ-IV ACH), (ii) Test of Cognitive Abilities (WJ-IV COG), and (iii) Test of Oral Language (WJ-IV OL). WJ-IV COG involves 18 question types to assess the cognitive strengths and weaknesses that may be related to learnt materials from school. WJ-IV ACH contains up to 20 types of questions which mainly focus on material learned at school, including subjects such as science or social studies. Skills assessed by WJ-IV may include verbal ability, quantitative reasoning, and pattern recognition.

SOL exams establish minimum expectations for learning and achievement at the end of each grade or course in English, mathematics, science, history, social science, and other subjects. To complete the test, students log into an online testing system called TestNav. This secure platform can be assessed through desktops, laptops, or tablets. SOL offers Computer Adaptive Testing (CAT), which is a customized testing

assessment that provides questions of increasing difficulty based on how well students respond to previous questions.

STAAR is a series of standardized tests used in primary and secondary schools to assess skills and knowledge in reading and math. The test is designed to test the Texas Essential Knowledge and Skills (TEKS), which is the state-mandated curriculum standard for public schools. The test is administered annually, each taking 2-3 hours to complete. Parents and guardians can access the government portal to access students' grades.

PARCC is a district annual assessment of mathematics and English language skills based on Common Core State Standards. The assessment measures the knowledge and skills that matter most for students — understanding complex texts, evidence-based writing, and mathematical problem-solving. The test is provided to students in grades 3 to 12 and is taken online each spring. The assessment results are delivered like report card grades, including feedback on classroom performance and subjective comments. The online exam can also accommodate students with disabilities or learning challenges.

ESSP MAP is an assessment and intervention strategy designed to improve student academic performance and behaviour. The test is administered online and involves collecting information from multiple sources. Teachers, parents, and students create online profiles, which help instructors interpret and establish specific interventions appropriate for each student. The test takes around 20 minutes for students, 30 minutes for parents, and 10 minutes for teachers.

SLA is an exam developed by the Alberta government to enable parents and teachers to identify the strengths and areas of growth in the coming academic year for students in Grade 3. The test is administered from August to October and delivered in digital and paper format. SLAs consist of different tests that can be administered independently and take four to four and a half hours to complete.

Phase 2: Tool Comparison Findings

Table 5 summarizes the key features of each assessment tool broken down by the comparison criteria. Following the table, we compare our findings for each tool across these criteria.

Table 5. Description of Assessment Tools by Comparison Criteria

Tool	Comparison Criteria							
	Timeliness	Resource-intensiveness	Duration	Cost	User-friendliness	Output	Validity	Reliability
Tools to Assess Literacy Skills								
PM-2	2010	Teacher resources, books, assessment cards	Length depends on student's progress	Assessment kit costs $687 (one-time)	Organized materials designed to administered by trained teachers	Easy to understand as it highlights student's areas of strengths and weaknesses	Strong validity in accurately predicting student's future reading achievement	Reliable for measuring student's reading level
Read180	2016	Lesson plans, audiobooks, paperbacks, assessment technology, practice material, teacher resources	23-week period	Approximately $11,000 for 30 students (annually)	Can vary depending on user's level of familiarity, require initial training	Can be complex as the program includes multiple measures of student progress	High validity	Consistent reports of positive effects on comprehension and reading fluency
EGLA	2014	Assessors able to conduct oral interviews	15-minute individual oral interviews followed by assessor coding (once a year)	Not specified	Requires assessors to be well-trained in coding	Complex, results must be coed consistently	Depends on coding expertise	Reliable if there is consistency across coding
Tools to Assess Numeracy Skills								
MAVLE	2013	Runs on standard computer hardware	Use repeatedly throughout the school year	No cost	Can vary depending on experience and familiarity with	Easy to interpret with visual and interactive concepts of	High validity	Students consistently improve in numeracy

					virtual manipulatives	numeracy learning		learning achievement
Numeracy Scale Insight Assessment	2015	Test booklets, questionnaire, graphics, online interface	75 minutes	Not specified	In-person and online assessment with skill evaluation questions	Output is easily understandable with supporting materials	Numeracy score strongly correlated with International Baccalaureate math scores	Low reliability with a Flesch-Kincaid Grade Level of 4.2 (may be due to sample size limitations)
Tools to Assess BOTH Literacy and Numeracy Skills								
CBM	Mid 1970s; Some newer measures available	Graphing tool to plot scores	Length varies based on measure (3 times per year)	No cost	Customizable based on needs	Easy to interpret and calculate statistics	Depends on measures used	Reliable for student performance due to standardized measures
WJ-IV	2014	Online test requires computers/tablet	60-90 minutes (can be administered repeatedly year-round)	$95 per student	Online test with engaging and interactive process. May require basic computer skills	May be complicated as different tests have different scoring scales	Shown to have high levels of validity in testing	Demonstrated high reliability
SOL	1997	Online test requires computers/tablet	90-120 minutes (once a year)	Free of charge for public school in Virginia students (Cost for other settings is unknown)	Online test with engaging visualization and different types of question formats	Clear graded scale that makes it easier to interpret	Valid measures of student achievement	High levels of internal consistency and test-retest reliability
STAAR	2012	Paper test format	3 hours (once a year)	Free of charge for public school students in Texas	Easily understandable questions with demonstrations	Different types of question formats which increases with grades	High validity	High level of test-retest reliability for math and reading

				(Cost for other settings is unknown)				
PARCC	2015	Online format that requires computer/tablet and audio system	60-90 minutes (once a year)	Free of charge for public school students in Texas (Cost for other settings is unknown)	The exam tests both critical thinking and lesions learned from school	Different types of question formats with visualization	High validity	High reliability
ESSP MAP	Between 1990-2002	Computerized format requires computer/tablet	20 minutes for children, 30 minutes for parents, 10 minutes for teachers	Not specified	System generates a summary score	Easy to interpret and compare scores against a benchmark	High content, developmental, convergent, criterion, and construct validity	Reliable scales
SLA	2015	Online format requires completed SLA application, computer/tablet, audio system, internet connection	4-4.5 hours in total, can be administered in several short sessions on different days (once a year)	Not specified, government funded	Digital questions marked by Alberta Education; performance tasks marked locally by teachers	Easily interpreted reports for teachers and parents	High, performance	Medium reliability, numeracy and literacy models perform well but students at grade level 3 are a challenging population to assess reliably

Timeliness: When was the assessment tool developed?
Four tools were developed or revised within the past 10 years – Read180 was updated in 2016, EGLA in 2015, SLA in 2014, and WJ-IV in 2014. The most outdated tools are PM-2 (developed in 2010), SOL (1997), ESSP MAP (1990s), and CBM (1970s).

Resource-intensiveness: What resources are required to administer the tool?
Most tools require some combination of human resources and technology to administer. PARCC, MAVLE, STAAR, Numeracy Scale Insight Assessment, and WJ-IV rely less on staff resources; they only require technology or paper to conduct the assessment. EGLA requires staff to administer the assessment orally and code the results to develop an output. All other tools require both technology and staff (ESSP MAP, SOL, Read180, SLA, CBM, and PM-2).

Duration: How long does it take to administer the tool? Does the tool require repeated assessments?
The length of time to administer each tool varies. While EGLA and ESSP MAP have shorter durations (less than 30 minutes to complete), other tools take one to three hours to administer (WJ-IV, Numeracy Scale Insight Assessment, SOL, STAAR, PARCC). SLA takes four to four and a half hours to administer but can be broken into shorter sessions to complete the assessment over several sittings. Read180 is administered over 23 weeks and delivered in three twenty-minute group sessions and one twenty-minute- direct session. The length of time required for CBM and PM-2 varies based on students' progress while completing the assessment. Many tools (e.g., PARCC, SOL, STAAR, WJ-IV) are administered once a year, but some (e.g., MAVLE, Read180) are conducted repeatedly throughout the school year.

Cost: What is the cost to purchase the tool or conduct each assessment?
We identified some tools that were publicly funded and administered by governments or school boards, and, as such, the costs were not specified. For example, STAAR, SOL and PARCC are offered free of charge to students for state-wide assessments. CBM and MAVLE require no upfront costs to administer. PM-2 has a one-time fee of $687 for an assessment kit of resources including books and student records for 67 students. Read180 was drastically the most expensive tool, costing around $11,000 for a license and supplementary resources. WJ-IV had one-time fees of $95, respectively. The costs for ESSP, SLA, Numeracy Scale, and EGLA were not specified.

User-friendliness: What is the tool's ease of use?
Our findings regarding user-friendliness were mixed; some tools require minimal to no training to complete the assessments or review outputs, while others are more complicated. CBM allows educators to easily customize assessments based on their (or student's) needs and goals. PM-2 provides a comprehensive toolkit to educators with several training materials to orient them to the assessment tool and output. While Read180 offers many assistive resources (e.g., workshops, audiobooks, manuals), this may be initially overwhelming to educators. Some tools, like SLA, provide an interactive dashboard for parents or guardians to view assessment results.

Output: What is the complexity of the output

We identified diversity across the tools in the complexity of their outputs. To produce the EGLA output, educators must complete some back-end coding, which may be complex for those without prior coding knowledge. The results from WJ-IV may appear complicated because multiple scores are generated for the different components of the assessment. PM-2, SOL, and SLA offer plain language reports or dashboards for parents/guardians that highlight the main results. Read180 allows educators to monitor student progress over the 23-week learning period; most other tools just provide scores at a single time that would need to be manually tracked by staff over time.

Validity: What is the tool's validity?

Most tools claimed to have high validity. CBM reported some diversity in validity, stating that some of its testing components are more valid than others. ESSP MAP appeared to be the most rigorously tested as the researchers reported on multiple different types of validity (content, developmental, convergent, criterion, and construct).

Reliability: What is the tool's reliability?

Similarly, most tools claimed to be highly reliable. Only the Numeracy Scale assessment reported poor reliability. Some tools involved rigorous testing (e.g., STARR, PARCC), although the findings from studies should be cautioned due to sample size limitations. Tools that were adopted by provincial or state governments were often tested across multiple sites and samples before implementation, which suggests greater rigour (e.g., SOL, SLA).

Discussion

Our online search resulted in 12 assessment tools for consideration by "Beyond the Bell" leadership to assess the literacy and numeracy skills of youth enrolled in the program. Three tools assess literacy skills (i.e., PM-2, Read180, EGLA), two assess numeracy skills (i.e., MAVLE, Numeracy Scale Insight Assessment), and seven assess both literacy and numeracy (i.e., CBM, WJ-IV, SOL, STAAR, PARCC, ESSP MAP, SLA).

In our comparison, we identified variability in the strengths and limitations of tools across the criteria. In our group discussions, it was challenging to identify tools that satisfied all the criteria; some fared well for some criteria but were limited in other aspects. It was also challenging to make an informed judgement about tools because some had more easily available information about each of our comparison criterion than others. For example, the costs of tools that were publicly funded and offered state- or provide-wide were often not reported.

Despite this variability, we selected three tools that we recommend "Beyond the Bell" to consider for future learning assessments: PM-2, MAVLE, and SLA. PM-2 is an established tool to assess literacy skills and offers many resources after purchasing a

one-time license. While the tool does require staff to administer the tool, we believe the informative output and high validity and reliability make this tool suitable for the program. We believe PM-2 was stronger than the other literacy tools we identified. Specifically, Read180 has an incredibly high cost, assessments span several weeks, and the number of staff resources and training required may be overwhelming. EGLA requires complex coding expertise to develop outputs, and its cost to administer was not specified.

We believe MAVLE is the strongest numeracy assessment tool. There is no cost to use this tool, it requires basic computer hardware, the assessment can be conducted online (requiring minimal staff resources), the output is considered user-friendly, and testing has demonstrated high validity and consistency. In contrast, the Numeracy Scale Insight Assessment requires staff to administer some in-person components of the assessment, testing has shown low reliability, and costs were not specified.

We selected SLA as the best comprehensive assessment tool. The key features of SLA are its ability to be administered over time (it does not need to be completed in one sitting), the output is easily interpretable with an engaging parental dashboard, and a provincial government stands behind the tool as it is currently being administered across Alberta. The limitations of this tool are its medium reliability and unspecified cost. We also believe CBM, ESSP MAP, and WJ-IV are strong candidates as tools to assess literacy and numeracy. With minimal costs, CBM could be considered by “Beyond the Bell.” We recognize the easily interpretable outputs, high reliability, and options to customize the assessment as its main strengths. ESSP MAP is a relatively short, computer-based assessment that boasts high validity and reliability, although the costs are unknown, and the assessment requires staff and parental input (resource-intensive). Lastly, WJ-IV is frequently updated (the assessment is currently on its fourth iteration), tests multiple knowledge domains, and does not require staff to administer the assessment or develop the output. However, the cost per student could be high given new students entering “Beyond the Bell” throughout the school year, and the output may be complex given the number of areas the tool assesses.

Our comparison of the tools may be impacted by the sources of information we used to identify and gather information about each tool. Since our search strategy was not comprehensive and some search results were inaccessible, we may have excluded some tools that would have been suitable candidates for “Beyond the Bell” to consider. Our comparison was informed by secondary information (articles and studies), which may have biased our assessments. For example, we did not have information about the costs of many tools, and user-friendliness is a subjective measure that may differ based on the audience. Also, some studies were conducted using more rigorous methods than others (e.g., ESSP MAP tested and reported on many types of validity), which made it challenging to compare our criteria across tools. Overall, our suggestions are preliminary, and “Beyond the Bell” needs to investigate our shortlist of tools for themselves and select the one they feel is most appropriate for their circumstances.

Conclusion

In this study, we conducted an online search and identified 12 assessment tools to assess literacy and numeracy skills. We reviewed 13 academic articles and grey literature that applied various methodologies to test the efficacy of tools. We used this information to compare the tools across eight criteria: timeliness, resource-intensiveness, duration to administer, cost, user-friendliness, quality of the output, validity, and reliability. The included literature was of varying methodological quality and provided different levels of detail about each tool, which hindered our comparisons. Our discussion of the strengths and limitations of tools across the evaluative criteria resulted in a shortlist of three highly ranked tools (PM-2, MAVLE, and SLA).

We advise “Beyond the Bell” leadership to consider the findings and recommendations we present in this report and select a tool that aligns best with their needs and preferences.

Bibliography

Blue, L. (2009). *An investigation into the relationship between cognitive ability, standardized achievement, and grades in middle school* (Order No. AAI3359614). Available from APA PsycInfo®. (622093152; 2009-99230-462). Retrieved from http://libaccess.mcmaster.ca/login?url=https://www.proquest.com/dissertations-theses/investigation-into-relationship-between-cognitive/docview/622093152/se-2

Cusumano, D. L. (2007). Is it working?: An overview of curriculum based measurement and its uses for assessing instructional, intervention, or program effectiveness. *The Behavior Analyst Today*, *8*(1), 24.

Daghestani, L. (2013). *The design, implementation and evaluation of a desktop virtual reality for teaching numeracy concepts via virtual manipulatives* (Doctoral dissertation, University of Huddersfield).

Doabler, C. T., Clarke, B., Kosty, D., Sutherland, M., Turtura, J. E., Firestone, A. R., ... & Jungjohann, K. (2022). Promoting understanding of measurement and statistical investigation among second-grade students with mathematics difficulties. *Journal of Educational Psychology*, *114*(3), 560.

Gittens, C. A. (2015). Assessing numeracy in the upper elementary and middle school years. *Numeracy*, *8*(1), 15-28.

Government of Alberta. (2023). Student Learning Assessment. Retrieved from https://public.education.alberta.ca/assessment/

Government of New Brunswick. (2023). Assessment and Evaluation (Anglophone Sector). Retrieved from https://www2.gnb.ca/content/gnb/en/departments/education/k12/content/anglophone_sector/assessment_evaluation.html

Jackson, J. C. (2010). *Organizational citizenship behaviors, collective teacher efficacy, and student achievement in elementary schools* (Order No. AAI3392567). Available from APA PsycInfo®. (754060313; 2010-99131-039). Retrieved from http://libaccess.mcmaster.ca/login?url=https://www.proquest.com/dissertations-theses/organizational-citizenship-behaviors-collective/docview/754060313/se-2

Kim, J. S., Samson, J. F., Fitzgerald, R., & Hartry, A. (2010). A randomized experiment of a mixed-methods literacy intervention for struggling readers in grades 4–6: Effects on word reading efficiency, reading comprehension and vocabulary, and oral reading fluency. *Reading and Writing*, *23*, 1109-1129.

Lee, Y. S., Lembke, E., Moore, D., Ginsburg, H. P., & Pappas, S. (2012). Item-level and construct evaluation of early numeracy curriculum-based measures. *Assessment for Effective Intervention*, *37*(2), 107-117.

Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of Math and Reading Achievement in Low Achieving Children in Elementary School: Effects of Early and Later Retention in Grade. *Journal of educational psychology*, *104*(3), 603–621. https://doi-org.libaccess.lib.mcmaster.ca/10.1037/a0027571

Same, M. R., Guarino, N. I., Pardo, M., Benson, D., Fagan, K., & Lindsay, J. (2018). Evidence-Supported Interventions Associated with Black Students' Education Outcomes: Findings from a Systematic Review of Research. *Regional Educational Laboratory Midwest*.

YMCA (2023). YMCA Beyond the Bell. [Online]. Accessed on March 24, 2023 via https://www.ymcahbb.ca/housing-community/ymca-beyond-bell

Zoda, P. F. (2010). *Elementary school size and academic performance: A multi-year study* (Order No. AAI3399011). Available from APA PsycInfo®. (759138846; 2010-99150-385). Retrieved from http://libaccess.mcmaster.ca/login?url=https://www.proquest.com/dissertations-theses/elementary-school-size-academic-performance-multi/docview/759138846/se-2