MAP Growth Validation Study

An Evaluation of the Alignment of
Selected MAP Growth Assessments to the Virginia Standards of Learning
and an Exploration of the Utility of MAP Growth Reports for
Determining Student Performance Relative to Grade Level

Christopher R. Gareis, Ed.D.
William & Mary

James H. McMillan, Ph.D.
Virginia Commonwealth University

Amelie Smucker, M.A.T.
William & Mary

Ke Huang, M.S.W.
William & Mary

August 5, 2021

Acknowledgements

This study was funded by a competitive grant from the Virginia Department of Education (VDOE). The study consortium consisted of four school divisions in Virginia: Chesapeake Public Schools (CPS), King William County Public Schools (KWPS), Williamsburg-James City County Public Schools (WJCC), and York County School Division (YCSD). The consortium partnered with faculty from William & Mary (W&M), Virginia Commonwealth University (VCU), and Georgia Southern University (GSU) in designing and undertaking the study.

Steering Committee Members

Leeza Beazlie	Coordinator of Testing & Accountability	YCSD
Diane Edwards	Director of Department Assessment & Accountability	CPS
Chris Kelly	Coordinator of Accountability & Assessment	WJCC
Joe Schipper	Supervisor of Assessment for Teaching & Learning	CPS
Amy Siepka	Director of Curriculum & Instruction	KWPS

Study Leads

Chris Gareis	Professor of Educational Leadership	W&M
Jim McMillan	Professor Emeritus of Education	VCU

Review Team Leads

Kristin Conradi-Smith	Associate Professor	W&M
Sam Rhodes	Assistant Professor	GSU

Subject Matter Experts

Alanna Bragg	Teacher	WJCC
Ellena Baker	Title I Instructional Coach	CPS
Wanda Calhoun	Secondary Division Math Coach	YCSD
Francine Cooney	Title I Instructional Coach	CPS
Lisa Derbaum	Title I Instructional Coach	CPS
Brian Domroes	High School Math Teacher	KWPS
Heather Galgano	Middle School English Teacher	KWPS
Jillian Gillikin	Elementary School Testing Coordinator	YCSD
Angela Gogol	Math Specialist	WJCC
Tracy Holland	Reading Specialist K-12	WJCC
Sharon Huber	Elementary Math Supervisor	CPS
Hancy Jean-Baptiste	Grade 5 Teacher	CPS
Crystal Kelly	Middle School English Teacher	KWPS
Meagan King	Elementary Assistant Principal	YCSD
Lindsay Kurtz	Elementary School Testing Coordinator	YCSD
Kim Lawler	Primary School Teacher	KWPS
Amy Livingstone	School Improvement Specialist	WJCC
Kiondra Russell	Assessment Specialist	CPS
Nicole Showah	General Education Teacher	WJCC
Jennifer Stanley	School Improvement Specialist	WJCC
Vika Stephenson	Secondary English & History Instructional Coordinator	YCSD

Jaclyn Tolbert	Elementary Teacher	KWPS

Data Analysts

Amelie Smucker	Doctoral Student in Educational Leadership	W&M
Ke Huang	Doctoral Student in Educational Leadership	W&M

Table of Contents

Executive Summary

The purpose of this study was to gauge the degree to which selected NWEA MAP Growth assessments are aligned to the Virginia Standards of Learning (SOL) and the extent to which MAP Growth reports can be used by school divisions to gauge student achievement relative to grade level and to identify learning gaps.

The study was delimited to four MAP Growth assessments: Reading 2-5, Reading 6+, Mathematics 2-5, and Mathematics 6+. The review of technical adequacy was largely dependent upon the NWEA-created technical report. Additionally, the stratified random sample of items for review was delimited to a range of four grade levels per test and therefore not representative of the full grade-level range of possible items in the test item pools. The study drew item reviewers and focus group participants from four selected school divisions.

Three processes were used to collect information and data for analysis: (1) a review of technical information about MAP Growth assessments provided by NWEA, as well as other literature concerning MAP Growth assessments and computer adaptive testing; (2) a review of 2,400 MAP Growth test items in Grades 2-9 in Reading and Grades 2-Algebra in Mathematics for content and depth of knowledge match with Virginia SOLs; and (3) two focus group interviews with end-users of MAP Growth reports, including subject-matter experts, instructional specialists, and division-level leaders responsible for assessment and evaluation.

From the analysis of collected data, four conclusions were drawn:

1. MAP Growth assessments have very strong technical adequacy, and test items are well aligned to the Virginia Standards of Learning.
2. Due to limits of computer-based standardized testing and the purpose of the assessment itself, MAP Growth assessments do not assess the full range of SOL content, nor do they assess at the highest levels of cognitive demand; therefore, the use of MAP Growth assessments should be complemented by the use of other trustworthy indicators of student learning in order to assess the full range of the curriculum.
3. MAP Growth assessments provide end-users (namely, division- and school-level leaders and classroom teachers) accurate, useful, and timely information about student, small-group, and school-level achievement, which, when supported by additional evidence of student performance, can be used to make near- and long-term instructional decisions to further student growth.
4. Given the limitations of norm-based, computer-adaptive testing, results regarding categorical determinations of *at, above,* and *below* grade-level performance must be interpreted with caution and complemented by other trustworthy indicators of student learning.

More specific key findings and recommendations from the study follow:

Technical Adequacy of MAP Growth Tests

- NWEA provides strong evidence for the validity, reliability, and fairness of overall scores and scores by *instructional area*. Test and item development procedures are consistent with established professional standards. MAP Growth assessments do not report sub-area scores; therefore, there are no validity and reliability data reported at the level of instructional sub-areas.
- Overall subject MAP Growth test scores, as well as scores in instructional areas, should be interpreted with attention to appropriate standard errors that describe performance bands in which true student performance lies.
- NWEA has stated clearly that MAP Growth assessments and SOL tests are designed for different purposes and cannot be assumed to be interchangeable. MAP Growth assessments should not be used as a substitute for state summative tests.

Alignment to Virginia SOLs

- A review of 2,400 items from MAP Growth Reading and MAP Growth Mathematics found substantial agreement among the subject matter experts that a very high percentage of items clearly or likely matched both the SOL and the depth of knowledge (DOK) tagged by NWEA to individual items.
- Item reviewers demonstrated strong interrater agreement in conducting item reviews for content and DOK.
- While Virginia teacher judgements confirm the NWEA-supported claims of both content and DOK alignment of MAP Growth items, there is less than complete coverage of the full body of SOLs in the MAP Growth tests. In the review of a stratified random sample of items, 68% of elementary reading SOLs were represented, 76% of secondary reading SOLs, 86% of elementary mathematics SOLs, and 92% of secondary mathematics SOLs. This means that, especially for reading, a significant number of standards are not assessed by MAP Growth tests.

Utility in Identifying Gaps in Learning

- School division administrators find NWEA-stated purposes of MAP Growth assessments and uses helpful, especially in monitoring student achievement and growth over time, planning instruction, comparing student performance to national norms, and, in the main, predicting performance on SOL tests.
- Student classification of different levels of performance, expressed as RIT scores, are based on national norms, not school-, district-, or state-level normative data. The nature of the MAP Growth tests and data result in students being classified as "at," "below," or "above" grade level *based on these national norms*. This suggests that conclusions regarding grade-level student performance for Virginia or specific divisions, schools, or classes is not established or directly reported but can be accessed through NWEA linking studies. To determine student achievement relative to grade-level standards, the use of MAP Growth results should be complemented by the use of other measures, including teachers' classroom assessment and evaluation of student learning.

- In light of the unknown reliability and match of MAP Growth assessments with curriculum guides, caution should be used in drawing inferences about student performance relative to *instructional sub-areas*. There is an uneven match between what MAP Growth tests assess, local curriculum guides, and SOLs. This means that some interpretations of MAP Growth data must include an analysis of the alignment between the local curriculum and MAP Growth scores. This is particularly important in looking at results based on national norms that assume a certain number of weeks of instruction prior to test administration.
- Instructional specialists and division-level assessment leaders are generally very positive about the accuracy and usefulness of the MAP data and reports.
- NWEA support and training are individualized to district needs, and school- and division-based leaders generally consider the training to be effective, helpful, and necessary.
- The large number of MAP Growth reports, along with training needed to understand RIT scores and comparative data, means that teachers and instructional specialists may need years of experience and support to fully establish appropriate uses of MAP Growth reports.
- Currently, there is little to no systematic data supporting the efficacy of classroom teachers' use of MAP Growth test results. There is some anecdotal evidence among end-users that the data are useful for instructional decision-making.
- Due to the timing of this study, predictive evidence for validity in using fall and winter MAP Growth scores to predict spring SOL test scores was lacking, although anecdotal evidence was mostly positive.
- Use of MAP Growth test data should primarily be restricted to the two major purposes of the assessment: (1) to gauge student growth in achievement over time and (2) to identify possible strengths and weaknesses in student proficiency of what is tested.

Purpose and Scope of the Study

MAP Growth Assessments

NWEA MAP Growth assessments are computer-adaptive tests designed to measure students' academic achievement and growth at interim points in time, up to four times a year (NWEA, 2019a). MAP Growth tests are currently available in Reading, Language Usage, Mathematics, and Science. Tests are banded by grade levels, such as K-2, 2-5, and 6+, although not all subjects are assessed at all grade levels.

As computer-adaptive tests, MAP Growth assessments draw from item pools that span grade levels. Thus, a fourth-grade student would answer items aligned to fourth grade standards but could also respond to items at lower and higher grade levels, depending upon their real-time performance while taking the test. The computer-adaptive algorithm is designed to select items that sample a breadth of subject-area content while also adjusting by grade level. As such, a student's performance can be gauged relative to grade level. Through multiple administrations, a student's academic "growth" over time can be seen. Furthermore, MAP Growth assessments are nationally normed, thereby allowing for a comparison of a student's performance to a reference group (NWEA, 2019a).

NWEA offers versions of the assessments aligned to the Common Core Curriculum, as well as to state-specific standards, such as the Virginia Standards of Learning (SOLs). As growth measures, the MAP Growth tests are intended as interim assessments that can be used to predict performance on end-of-year state summative assessments (NWEA, 2016; NWEA, 2020).

Purpose of the Study

The purpose of this validation study was to gauge the degree to which selected NWEA MAP Growth assessments (a) are aligned to the Virginia SOLs and (b) can be used to identify growth and learning "gaps" in student achievement. More specifically, the study reviewed the nature of and stated uses for MAP scores and individual items on four MAP Growth assessments:

- Reading 2-5
- Reading 6+
- Mathematics 2-5
- Mathematics 6+

The aim of the study was two-fold:

1. To determine the degree of alignment between representative samples of items and the Virginia SOLs, and
2. To determine the extent to which it is reasonable to use MAP results to document student growth and learning gaps for students *on, above,* or *below* grade level relative to the Virginia SOLs.

To address these aims, the study was conducted in three phases:

- Phase 1: Clarification of the Purpose and Design of MAP Growth Assessments
- Phase 2: Validation of Samples of Items
- Phase 3: Examination of the Perceived Accuracy and Utility of Inferences Drawn by End Users of MAP Growth Reports

Scope of the Study

This study was limited to four MAP Growth assessments and was undertaken in Virginia, which has its own curriculum standards (i.e., Virginia has not adopted Common Core Standards); therefore, generalizability of the findings to other MAP Growth assessments or to other states should be made with caution. The study sought only to address the aims and questions articulated in the previous section. Notably, the study was not designed as an explicit investigation of concurrent validity, predictive validity, or consequential validity, nor did it investigate assessment issues related to user interface of the platform, item difficulty, or potential bias in item construction. However, some of these issues are tangentially addressed in the findings and discussion.

In addition to partnering with the four school divisions in the study consortium, the study leads coordinated with NWEA leadership and staff to amend the study methodology for feasibility purposes and relied upon NWEA to generate stratified random samples of items for Phase 2 of the study.

Neither the study leads nor any member of the study team hold financial interest in NWEA.

Methodology

The three phases of the study addressed distinct evaluation questions, each of which was investigated through a different methodology.

Phase 1: Clarification of the Purpose and Design of MAP Growth Assessments

In this phase, the intent was to clarify the purpose and design of the MAP Growth tests through a document review, as well as to critically examine the evidence of validity of the assessments through these documents. The study leads examined NWEA technical reports, previous validation studies, and extant literature to determine the purpose of the assessments, test design features, alignment claims, types of items, parameters of testing events for students (e.g., duration, location), and logic of the computer-adaptive algorithm. The guiding evaluation question was, "What are the validation claims of the MAP Growth assessments?"

The primary sources reviewed were the following NWEA publications, each of which was identified by NWEA as being an appropriate source for this phase of the study:

- *Instructional Areas: Standard Alignment—Virginia Mathematics 2-5* (NWEA, 2020b)
- *Instructional Areas: Standard Alignment—Virginia Mathematics 6+* (NWEA, 2020c)
- *Instructional Areas: Standard Alignment—Virginia Reading 2-5 and 6+* (NWEA, 2019b)
- *Linking the Virginia SOL Assessments to NWEA MAP Growth Tests* (NWEA, 2016)
- *Linking Study Report: Predicting Performance on the Virginia Standards of Learning (SOL) Mathematics Assessments Based on NWEA MAP Growth Scores* (NWEA, 2020a)
- *MAP Growth Technical Report* (NWEA, 2020a)

In determining the purpose and design of the assessments and examining the evidence of their validity, the "MAP Growth Technical Report" served as the most comprehensive source.

Phase 2: Validation of Samples of Items

The purpose of the second phase of the study was to determine the degree of alignment between representative samples of items and the Virginia SOLs they purport to assess. Phase 2 constituted the central focus of the validation study, for which the guiding evaluation question was, "How adequately do the MAP Growth assessments align with the Standards of Learning (SOL)?" Anchored in Webb's (2007) alignment review protocol, this phase of the study evaluated the degree to which items matched the Standards of Learning in content and cognitive demand.

Review Teams

The alignment review was conducted by subject matter experts (SMEs) recruited from each of the four consortium school divisions. The SMEs were selected by division-level leadership based

upon (a) demonstrated content knowledge in the subject and grade levels of the respective MAP assessments, (b) experience with the SOLs, and (c) expertise in assessment. Four teams of 5-6 SME reviewers were formed to review sample items for each of the four MAP Growth assessments of interest:

- Reading 2-5,
- Reading 6+,
- Mathematics 2-5, and
- Mathematics 6+.

The review teams were led by two university faculty members, each well qualified for this role given their respective (a) depth of pedagogical content knowledge in their subject area, (b) experience and expertise in unpacking Virginia SOLs, (c) expertise in principles of assessment, and (d) effectiveness in leading professional training.

Item Sample Sets

A total of 600 items was used as the sample set for each of the four MAP Growth assessments. MAP Growth assessments draw from item pools that can span as many as 13 grade levels (i.e., K-12); however, per NWEA advisement, the preponderance of items for a typical test event draws from the most immediate span of grades. Therefore, the sample sets were delimited to a span of four grade levels per test, as depicted in Table 1.*

Table 1. Sample Item Set Sizes by Subject Test and Grade Levels

READING	Item Sample Size	MATHEMATICS	Item Sample Size
Grade 2	150	Grade 2	150
Grade 3	150	Grade 3	150
Grade 4	150	Grade 4	150
Grade 5	150	Grade 5	150
TOTAL Sample for Reading 2-5	**600**	**TOTAL Sample for Mathematics 2-5**	**600**
Grade 6	150	Grade 6	150
Grade 7	150	Grade 7	150
Grade 8	150	Grade 8	150
Grade 9	150	Algebra I	150
TOTAL Sample for Reading 6+	**600**	**TOTAL Sample for Mathematics 6+**	**600**
GRAND TOTAL of Reading Items	**1,200**	**GRAND TOTAL of Mathematics Items**	**1,200**

* MAP Growth Reading 2-5 draws from an item pool that spans SOLs in Grades K-8, and MAP Growth Reading 6+ spans SOLs in Grades 3-12. MAP Growth Mathematics 2-5 draws from an item pool that spans mathematics SOLS in Grades K-8. MAP Growth Math 6+ draws from an item pool that spans mathematics SOLS in Grades 3-Algebra II (including Algebra and Geometry).

In order to reflect the categorical structure of the assessments, each set of items was a *stratified random sample*. Stratification occurred by *instructional area*. NWEA uses the term instructional area to mean content domain.[†] NWEA's instructional areas are similar to—but not exactly the same as—the term *strands* in the Virginia SOLs and the term *reporting categories* on the SOL tests.

NWEA provided access to the stratified random sample sets through an online review portal. The following elements and information were provided for each item:

- Item prompt, answer choices (if select-response), and correct answer
- SOL to which the item aligned, as tagged by NWEA
- Instructional area of the item, as tagged by NWEA
- Depth of knowledge (i.e., "DOK" designation) of the item, as tagged by NWEA

Alignment Review Process

Each SME reviewer was assigned a set of 100 – 150 items to review from the sample set. In order to reduce the demand of reviewing items across four grade levels, each SME was assigned items that spanned only two consecutive grade levels. This allowed SMEs to hone their understanding of the SOLs and thereby strengthen the accuracy of their reviews.

SMEs completed their review of items individually over the course of two weeks. For each item, three judgements were made:

1. To what degree does the item match the NWEA-tagged SOL? (Clear match, Likely match, Unlikely match, No match)
2. Does the item match another SOL in the same grade level? (Yes, No)
3. What is the depth of knowledge of the item? (DOK 1, 2, or 3)[‡]

Additionally, reviewers made brief written comments as appropriate for any item to note issues or details related to their judgements. Lead reviewers for each subject area periodically reviewed reviewer judgments to assure accuracy and reliability.

The complete "Item Review Protocol" is included as Appendix A.

[†] MAP Growth tests are defined by the structure of the respective disciplines they purport to measure, which NWEA refers to by the term *instructional areas.* Instructional areas are further defined by *sub-areas.* For example, MAP Growth Mathematics 2-5 is defined by five *instructional areas*: (1) Number and Number Sense, (2) Computation and Estimation, (3) Measurement and Geometry, (4) Probability and Statistics, and (5) Patterns, Functions, and Algebra. To further the example, the *instructional area* Number and Number Sense is more specifically defined by two *sub-areas*: (a) Whole Numbers: Place Value, Count, and Compare and (b) Fractions & Decimals: Represent and Compare (NWEA, 2019; NWEA, 2020).

[‡] Due to the inherent limitations of the current state of the art of online standardized testing, MAP Growth assessments do not include any items that tap DOK 4 (NWEA, 2020).

Interrater Agreement

Interrater agreement was established through several procedures. First, a general orientation session was held synchronously online for all SME reviewers and SME leads by the study leads, during which time the purpose of the study and the three phases of the study were explained. Then, SMEs met in subject-specific breakout groups with their respective SME leads to learn about the item review protocol.

Second, a half-day training session was conducted several days following the orientation. Two sessions were held simultaneously, one for Reading SMEs and the other for Mathematics SMEs, which allowed for differentiated examples and increased opportunity for discussion and calibration. As part of the training, SME reviewers were provided and made use of a written Item Review Protocol (Appendix A). The sessions were overseen by the two study leads, and the sessions were observed by members of the Steering Committee and the Data Analysts. The training sessions were comprised of direct modeling, guided practice, whole-group processing, initial interrater agreement checks with feedback, and additional cycles of practice and calibration. The aim was to establish interrater reliability of 80% or higher on average with no individual outliers among the SME teams.

Third, after establishing initial evidence of interrater reliability, a formal interrater agreement check was undertaken. Table 2 presents the interrater agreement as percentages by subject area and by review question.

Table 2: Interrater Agreement for Item Review

	To what degree does the item match the NWEA-tagged SOL?	Does the item align to a second standard at the same grade level?	What is the DOK level of the item?
Reading	96.6%	89.2%	83.3%
Mathematics	97.0%	72.7%	81.1%

The interrater agreement for the content alignment question was very strong for both subject areas. Notably, the interrater agreement for the question of alignment “to a second standard” was strong for Reading (89.2%) but modest for Mathematics (72.7%). However, the Mathematics SMEs reached an 81.8% interrater agreement on the final three items of the training sample set, indicating greater consistency among raters as the round concluded.

As a fourth step in ensuring interrater reliability, each of the two SME leads monitored SME reviewers’ individual item ratings by conducting random checks every 20-25 items (depending on the size of the reviewer's sample). When the SME lead disagreed with the SME reviewer on an item, the lead would make contact to discuss and resolve. On four occasions, SME leads asked a study lead to weigh in as a third reviewer for an item.

Phase 3: Examination of the Perceived Accuracy and Utility of Inferences Drawn by End Users of MAP Growth Reports

The final phase of the study examined K-12 practitioners' use of MAP Growth results. Practitioner perspectives were used to draw inferences about students' academic achievement and growth relative to their respective grade levels. The guiding evaluation question of this phase of the study was, "How adequately do the MAP Growth assessments identify learning gaps among students who perform *at, above,* and *below* grade level?" A structured focus group protocol (see Appendix B) was developed to obtain participants' perspectives based on previous experiences using MAP Growth reports. The practitioners commented on the practical significance of MAP Growth results in gauging student growth, identifying learning gaps, and making instructional decisions. A brief survey of SME reviewers provided additional information.[§]

Two focus groups were conducted—one with a selected sub-group of SME reviewers who had deep familiarity with using MAP Growth assessments, and the second with all of the members of the Steering Committee. For each group, a list of questions was provided in advance, as well as a copy of the NWEA (2021) publication *MAP Growth Reports V 2.0*, which provides a comprehensive explanation and instructive examples of all of the standard reports generated by MAP Growth. Additionally, focus group participants were asked to draw on their own experiences and material resources when responding to the questions.

Participants made written notes in response to the questions prior to their participation in their respective focus group. The purpose of the written notes was to prompt deep reflection in advance of the focus group, to provide notes from which to speak, and to submit initial and/or revised notes to the study leads following the focus group.

The focus group protocol followed the order of the written questions, but the intent of the focus group was to prompt thinking among the whole group in order to identify salient insights and practices of these "end users" of MAP Growth assessment results.

[§] In the original design of the study, the purpose of Phase 3 was to determine the adequacy of MAP Growth assessments in determining learning gaps among different performance levels. To that end, the study proposed to analyze the unique computer-adaptive testpaths of individual students representative of *on, above,* or *below* grade level performance. MAP Growth assessment determine low, medium, and high levels of performance based upon the algorithm that factors student responses (correct/incorrect) against item difficulty and item grade level. Since there are no unique "test forms" for varying levels of performance, determining "learning gaps" depends upon an analysis of the unique paths of item responses of students at respective performance levels. The evaluation planned to investigate this by reviewing samples of unique test events for a small sample of individual students at two grade levels, stratified by performance level relative to the median RIT (Rasch unit) score on the respective tests during that administration. For each student in the sample, the intent was to review each SOL to which each item was tagged, providing information to gauge the score report for an individual student to the sample of unique items in their test event. However, after conferring with NWEA staff on several occasions in response to concerns about both the intent and feasibility of our proposed methodology, it was determined that a more reasonable option was to use the focus group methodology previously described. Nevertheless, the intent of the original methodology of Phase 3 may be appropriate for future research.

The focus groups were co-conducted by the two study leads, one of whom facilitated the session and captured notes while the other took the lead on asking clarifying and probing questions.

Findings

Phase 1: Clarification of the Purpose and Design of MAP Growth Assessments

MAP Growth assessments are standardized achievement tests, matched to state standards, that measure K-12 student achievement and growth in Reading, Language Usage, Mathematics, and Science. NWEA claims that MAP assessments may be used for several purposes, including:

- Monitoring student achievement and growth over time, from kindergarten to high school
- Planning instruction for individual students and groups of students at the classroom, grade, school, and district levels
- Comparing student performances within normed groups
- Making universal screening and placement decisions within a response to intervention (RtI) framework or for talented and gifted programs
- Predicting student performance on external measures of academic achievement, such as the ACT®, SAT®, and on statewide summative achievement tests
- Evaluating programs and conducting school improvement planning
- Summarizing scores for district- or school-level resource allocation
- Combining RIT scores with other information (e.g., homework, classroom tests, state assessments) to make educational decisions (NWEA, 2019a, p. 7).

MAP Growth assessments are described as "interim" since they may be administered up to four times during the year (fall, winter, spring, and summer). NWEA (2019a) claims that this results in "[scores] that can be used to tailor instructional practices" (p. 3) to enhance achievement on year-end accountability tests.

NWEA is clear in indicating that MAP Growth assessments should not be substituted for year-end state accountability tests, including SOL tests:

> Virginia SOL and MAP Growth assessments are designed for different purposes and measure slightly different constructs even within the same content area. Therefore, scores on the two tests cannot be assumed to be interchangeable. MAP Growth may not be used as a substitute for the state tests. (NWEA, 2020a, p. 15)

Nature of MAP Growth Tests and Test Items

MAP Growth assessments are computer adaptive tests (CAT) that primarily use select-response items. The items for use in a particular state are identified based on the match between the items and state content standards. The online tests are untimed, taking approximately 45-60 minutes to complete, and typically consist of 40-50 items. As a CAT, the computer program instantly analyzes each response and, based on answering correctly or incorrectly, selects subsequent items that are at an appropriate level of difficulty for the student test-taker. This feature avoids items that are either very easy or very difficult for a particular student, ameliorating boredom or frustration. Students with the same score would have a 50% chance of answering a given item

correctly. Students with a higher score on the scale would have more than a 50% chance; students with a lower score on the scale would have less than a 50% chance. This limits the number of items students need to answer compared to non-CAT standardized achievement tests and means that each student answers a unique item set. While this characteristic of CAT testing lessens the amount of time needed to assess a general area, it results in a small number of items measuring any particular subskill.

Overall test and test item development procedures are consistent with established professional standards. The fairness of MAP Growth assessments is strong, with extensive item review procedures that eliminate bias and distracting sensitivity of item content, as well as text readability for each grade level. Appropriate accommodations are provided, and test security is established.

MAP items are designed to be challenging to students, economical in use of student time, precise, fair, aligned to content students presumably have had the opportunity to learn, and responsive in providing results quickly to educators and other stakeholders. Notably, MAP Growth tests utilize principles of Universal Design for Learning (UDL) to address needs of diverse student populations. Items are technology-enhanced, allowing the use of different types of formats, including traditional multiple-choice, multiple-correct choice, "drag-and-drop," "click-and-pop," "hot text" (selectable text), and text-entry (short constructed-response). According to an external review of MAP, the "process with which individual items were reviewed was impressively extensive and likely ensured high-quality test items" (Burns & Young, 2019, p. 666).

Reliability

Marginal reliability coefficients (similar to internal consistency) for overall RIT scores are approximately .97 for reading and .98 for mathematics. Marginal reliabilities for instructional areas in reading range from .90 to .91 and from .90 to .92 for instructional areas in mathematics. These high reliability coefficients indicate high consistency in answering items at one point in time. They do not reflect error attributed to sampling of content, nor error attributed to time between testing events. Test-retest reliabilities with alternate forms are reported to be between .75 and .85 for reading in Grades 2-5, and .85 to .92 for mathematics in Grades 2-5. Together, these reliability statistics indicate that students' overall RIT performance would be very close to the same if students took the test twice at about the same time, and also indicates reasonable stability of performance over time.

Validity

The *MAP Growth Technical Report* provides evidence of concurrent validity for overall composite RIT scores in broad subject areas, such as reading and mathematics (NWEA, 2019a). The results indicate that the classification accuracy between spring MAP Growth scores and SOL test scores across grade levels is 75-80% for reading and 80-90% for mathematics. This shows that there is substantial agreement with respect to classification of students as proficient/not proficient on tests given in the spring. Concurrent evidence for validity is indicated in the correlation between spring composite RIT scores and SOL scores. Correlations between

spring MAP Growth scores and SOL scores range, across grade levels, from 0.82 to 0.86 for reading and 0.81 to 0.86 for mathematics. Similar concurrent evidence is presented for other states and for other standardized tests. No validity evidence is included for instructional areas. Additionally, since MAP Growth does not report on student performance relative to instructional sub-areas, no validity evidence is available for that level of analysis.

Additional validity evidence is provided by the NWEA (2020a) *Linking Study Report*, which shows that MAP Growth mathematics scores, using thousands of Virginia students, correlated with Virginia SOL scores, ranging from 0.84 in Grade 3 to 0.80 in Grade 8. In addition, classification accuracy statistics ranged from 85% to 91%, suggesting that MAP Growth scores have a high accuracy rate of identifying *proficient* level of performance on the SOL test. These linking data are only applicable to total scores. Proficiency projections are indicated by applying established algorithms to fall and winter MAP Growth scores. This suggests that while fall and winter MAP overall scores may be predictive of overall spring SOL scores and categorization (e.g., proficient/not proficient), the prediction of spring SOL scores, based on previous fall and winter instructional area scores, has not been systematically established, and likely would be less accurate due to far fewer items used in reporting categories as compared to overall scores. Furthermore, interim MAP Growth tests assume designated instructional week intervals, but actual instructional weeks of each school division may differ from those designations. These differences may impact the estimations of student growth from fall or winter to spring, which can affect the classification accuracy of the fall and winter cut scores.

Other independent research has found modest correlations between fall MAP Growth testing and spring standardized test results. Jones (2015), in a study of 230 middle school students, reported a correlation of 0.67 between fall MAP Growth reading scores and spring reading accountability test scores in Georgia. In mathematics, Jones found a correlation of 0.75 between fall MAP Growth scores and spring scores on Georgia's accountability test. Klingbeil et al. (2015), using a sample of 520 third graders, found that the combination of fall MAP Growth reading scores with a measure of oral reading fluency accounted for 61% of the variance in spring MAP Growth reading scores. However, Mitchell (2019) reported that 78 of 389 eighth grade students from four Minnesota school districts were misclassified from fall academic progress MAP Growth scores as false negatives.

The NWEA (2020a) linking study did not report correlations of instructional area scores with reporting category scores (e.g., *Number Sense/Number Systems* RIT scores are not linked to SOL *Number and Number Sense* reporting category scores). The linking study is in progress by NWEA but has been delayed as of the time of this report due to the COVID-19 pandemic.

NWEA-Determined Alignment of MAP Growth to Virginia SOLs and to SOL Tests

NWEA content specialists conduct internal alignment analyses to establish a match to state standards, and third parties also conduct alignment studies. During this process, blueprints for each assessment are created to reflect the organization of the Virginia SOLs. Each blueprint is broken down into reporting categories known as instructional areas that mirror key concepts in the state standards. These instructional areas are further divided into instructional *sub-areas*. According to NWEA (2019a), an item is considered aligned when the item targets either the

whole standard or an integral part of a standard in a way that is both grade-appropriate and at a level of cognitive complexity addressed by the standard. NWEA-determined matches with Virginia reading and mathematics SOLs for Grades K-12 are summarized in the *MAP Growth Technical Report*. NWEA-determined matches provide one source of content-related evidence for validity with respect to alignment between what is tested with MAP Growth and Virginia SOLs. MAP Growth RIT instructional area scores, while based on SOLs, are not the same as SOL reporting category scores. This is especially true for reading, for which there are two SOL reporting category scores and three MAP instructional area scores. For mathematics, there is a clearer match between SOL reporting categories and MAP instructional area scores.

According to NWEA, the large item bank upon which MAP Growth assessments are based allows tests to be tailored to each state's learning standards. NWEA content specialists use knowledge of the SOL standards and content expertise to align items to specific Virginia SOLs. These aligned items then become part of the test pool for the assessment given to students. NWEA contends that this allows MAP Growth tests to serve two purposes simultaneously: (1) to measure achievement over time on state-determined learning standards, and (2) to compare achievement to national norms.

MAP Growth Test Scores

There are many types of scores reported for each MAP Growth test. These include overall content area scores (scores for a major area such as reading or mathematics) and instructional area scores.

Overall Content Area Scores

As a computer-adaptive test, MAP Growth content area scores are reported with a Rasch Unit (RIT) scale roughly between 100 and 350 across grade levels. This ability scale is unique for each subject and *continuous across grades*. The vertical scale allows NWEA to create growth norms within grades and spanning multiple grades (most recently calculated and released in 2020 based on a very large pool of nationally representative and diverse sample of students). The nature of the RIT scores shows how achievement on a given scale changes over time, as well as shows achievement compared to national norms. With the across-grade level emphasis, student achievement can be tracked for multiple years. For any given instructional season (e.g., beginning, mid-year, or end-of-year), students are classified as "at," "below," or "above" grade level *as compared to national norms.* As described by NWEA (2019a):

> [A] student's performance on the MAP Growth test, expressed as a RIT score, is associated with a percentile ranking that shows how well the student performed in a content area compared to students in the norming group. The relative evaluation of a student's growth from one period to another (e.g., from fall to spring) is provided by growth norms. (p. 54)

It is important to emphasize that some aspects of RIT score interpretation depend on percentile rank according to national norms. Since the primary purpose of MAP Growth

results is to show how students are achieving academically, independent of grade level, and show how student performance compares to national norms, a student may score well below or beyond the current grade level of instruction in a specific class or school.

Each RIT score has an associated standard error of measurement (SEM), a measure of random error. MAP Growth test standard errors are relatively consistent for 90% of RIT content area scores, and are generally smaller than what is observed for standardized achievement tests that are not CAT. As reported by NWEA (2019a), the mean standard error of measurement for composite content area scores in reading is 3.4 and for mathematics is 2.9-3.0. SEM is larger for students at the extreme ends of the scale, such as below 150 and above 260.

MAP Growth content area scores indicate achievement and growth over time, in comparison to the norm group, not necessarily achievement on specific SOLs. As such, the RIT score is not an indication of mastery, but the probability of a correct answer to items at a particular location of the scale based on the level of achievement. Achievement that is designated "at," "below," or "above" grade level is based on national norms, and is not normed for a specific district, school, class, or individual. This suggests caution in directly inferring needed instruction from RIT area scores since they are based on national norms, not the achievement of students in a particular class or school.

The use of national normative data with MAP tests also allows comparisons among different areas of proficiency. Fall administration assumes students are at week 4 from the beginning of the school year, winter at week 20, and spring at week 32.

Instructional Area and Sub-Area Scores

MAP Growth assessments provide RIT scores for reporting categories within each major content area (e.g., reading and mathematics). Each domain, in turn, is comprised of instructional sub-areas that are used to group test items to SOLs.

Comparing students on MAP-determined categories of performance at designated levels for an instructional area—such as "at," "below," or "above" grade level—may be used, with appropriate caution and verifying evidence, to show proficiency at different points in time (growth) or among instructional areas for diagnostic purposes. Instructional area standard errors are approximately 6 points (standard errors for instructional sub-areas are not reported). These points are used to create bands (score +/- SEM). When band scores do not overlap, there is greater confidence that a meaningful difference exists between proficiency in different instructional areas. For example, if a student's RIT score in the instructional area of Comprehension of Literary Texts is 187, with a 6-point standard error, and an RIT score of 181 in Comprehension of Nonfiction Texts, that 6-point difference suggests that there is not a meaningful, practical difference for this student, which suggests Comprehension of Literary Texts proficiency is not significantly higher than Comprehension of Nonfiction Texts proficiency. NWEA individual reports consider standard error in making determinations within instructional areas when labeling performance a "relative strength" or an "area of focus" for a student. It should be noted

that in comparing students, some RIT scores at the top of the range of the "at" categorization would not be meaningfully different from RIT scores at the bottom of the range of the "above" category. School level or small-group level RIT scores have a much smaller standard error than do individual scores.

MAP Growth tests utilize the organizatinon of state standards for determining instructional areas and sub-areas. Additionally, MAP Growth tests use a student's RIT ability value to compute the probability of answering an item correctly for any item in the item band, although a student may not actually be administered items aligned to a given SOL that is subsequently reported on. Therefore, it is very important for educators to use the scores in conjunction with other evidence of student performance on the areas tested, including classroom assessments, assignments, homework, and other measures for accurate interpretation and use.

Phase 2: Alignment of Items to Virginia's Standards of Learning

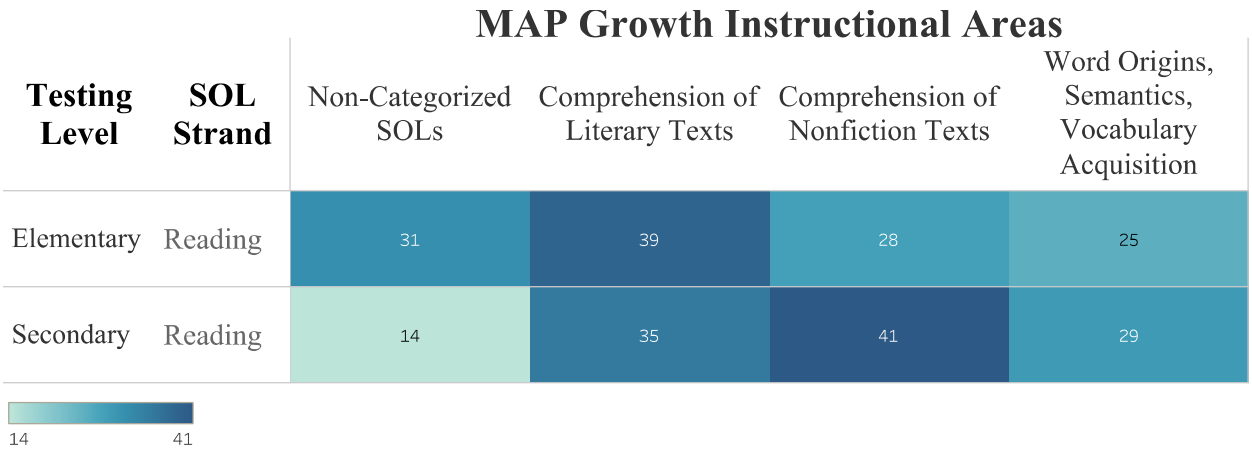
The purpose of Phase 2 of the study was to determine the degree to which MAP Growth assessments align to corresponding Virginia SOLs.

Comparison of MAP Growth Instructional Areas to SOL Strands and Standards

As a first level of analysis, NWEA-provided documents outlining instructional areas, instructional subareas, and corresponding SOLs (NWEA 2019b; NWEA, 2020b; NWEA, 2020c) were reviewed in comparison to the Virginia Standards of Learning. While the organization and naming of MAP Growth instructional areas do not completely correspond with the organization and naming of Virginia's reading and mathematics strands, the differences are modest and largely semantic.

In the area of reading, NWEA (2019b) reports that 81% of Virginia SOLs are assessed within their Reading 2-5 and Reading 6+ tests and are aligned to three NWEA instructional areas.[**] As shown in Figure 1, at the elementary level, the greatest number of Virginia standards (39) are categorized in NWEA's *comprehension of literary texts* instructional area, and 31 standards (25%) are not categorized. At the secondary level, the greatest number of standards (41) are categorized in the *comprehension of nonfiction texts* instructional area, and 14 standards (12%) are uncategorized. This distribution of standards in the MAP Growth instructional areas suggests a reasonable, but less than complete, association of reading standards across the instructional areas, with a notable number of standards not corresponding between the VDOE and NWEA reading categories.

Figure 1: Number of Reading SOLs Included in MAP Growth Reading 2-5 and 6+ Instructional Areas



Further analysis of reading standards by testing level and grade level found three trends. (See Table 3.) First, the majority of non-assessed SOLs were second grade standards. Second, only

[**] For this study, SOLs for the Reading strand were used for the comparison to the MAP Growth Reading assessments. SOLs from other English strands (namely, Multimodal Literacies, Writing, and Research) were *not* included. Additionally, since this study is not intended as an evaluation of concurrent validity, the comparison was not limited to SOLs considered to be assessable nor to SOLs identified in SOL test blueprints.

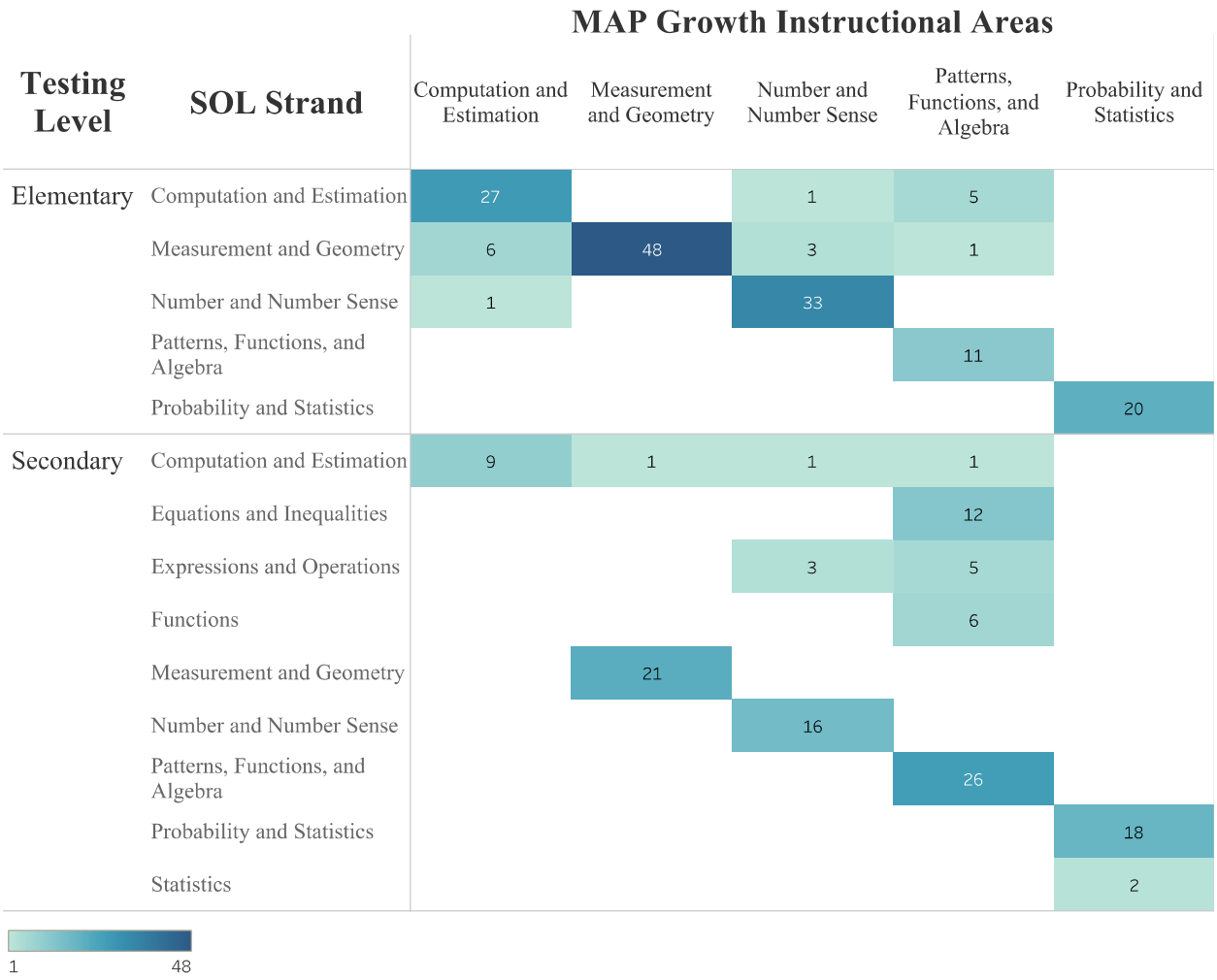
three standards were not assessed to any degreee or were not assessed completely (SOLs 2.3.a-e, 2.4.a-d, and 3.3.a-b), but these standards are related to phonemic awareness, phonetic strategies, and decoding. Third, as would be expected, a pattern of non-assessed standards was evident at multiple grade levels (e.g., 3.5.k, 4.5.k, 5.5.m, 6.5.k, 7.5.j, and 8.5.j). While the distribution of standards in the MAP Growth instructional areas is less than complete, these trends suggest that the majority of non-assessed standards represent non-assessable and/or repeated skills.

Table 3: NWEA-Reported Non-Assessed SOLs in Reading by Testing Level and Grade Level

Testing Level	Grade Level	Frequency of Non-Assessed Standards	List of Non-Assessed Standards
Elementary	2	15	2.3.a; 2.3.b; 2.3.c; 2.3.d; 2.3.e; 2.4.a; 2.4.b; 2.4.c; 2.4.d; 2.6.d; 2.7.b; 2.7.i; 2.8.c; 2.8.d; 2.8.h
	3	9	3.3.a; 3.3.b; 3.4.e; 3.5.a; 3.5.k; 3.5.m; 3.6.b; 3.6.i; 3.6.j
	4	4	4.5.k; 4.5.l; 4.6.h; 4.6.i
	5	3	5.5.m; 5.6.b; 5.6.k
Secondary	6	3	6.5.k; 6.6.a; 6.6.k
	7	3	7.5.j; 7.6.a; 7.6.m
	8	4	8.5.j; 8.6.c; 8.6.k; 8.6.m
	9	4	9.4.e; 9.4.g; 9.4.l; 9.5.l

For mathematics, NWEA (2020b, 2020c) reports that 100% of Virginia SOLs are assessed within their Mathematics 2-5 and Mathematics 6+ tests and are aligned to five NWEA instructional areas. As shown in Figure 2, it was confirmed that all of Virginia's Grades 2-9 math SOLs are categorized into at least one of NWEA's instructional areas. These results suggest strong correspondence of mathematics SOLs on the MAP Growth tests at both the elementary and secondary levels.

Figure 2: Number of Mathematics SOLs Included in MAP Growth Mathematics 2-5 and 6+ Instructional Areas



Representation of SOLs within the Sample of MAP Growth Items

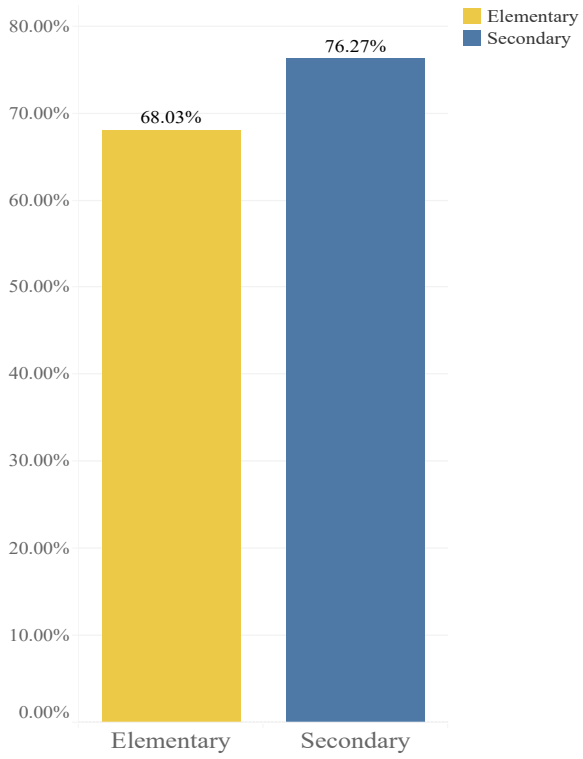
As a second level of analysis, the NWEA-provided sample of MAP Growth items was reviewed in comparison to the Virginia Standards of Learning. The extent to which Virginia SOLs are represented by MAP Growth items was analyzed by determining the percentage of standards addressed by the sample items. The percentages were calculated for total reading and mathematics, grouped by grade level. While the provided sample did not include items that addressed all SOLs, there was a very high percentage of items that clearly or likely matched both the NWEA-provided SOL and depth of knowledge (DOK) across all subject areas and grade levels.

Coverage of SOLs within the Sample of MAP Growth Items in Reading

The analysis of the sample items found that for reading a substantial, but less than complete, coverage of Virginia standards was indicated overall for all grade levels (72%).

As shown in Figure 3, across the testing levels in reading, 68% of the elementary reading standards were represented in the sample of MAP items, and 76% of secondary standards were represented.

Figure 3: SOLs Represented within the MAP Growth Item Sample Set by Testing Level in Reading



As depicted in Table 4, there was some variation across grade levels, from a low of 47% at the second grade level to a high of 81% at the fourth grade level.

Table 4: SOLs Represented within the MAP Growth Item Sample Set by Grade Level in Reading

Grade Level	Number of SOLs Aligned to MAP Growth Reading Sample	Total Number of Virginia SOLs	Percentage of SOLs Assessed
2	16	34	47.06%
3	22	32	68.75%
4	21	26	80.77%
5	24	30	80.00%
6	22	28	78.57%
7	23	30	76.67%
8	22	30	73.33%
9	23	30	76.67%

For MAP Growth Reading 2-5, 68% of the Virginia SOLs were represented by sample test items. Beyond the 31 standards not categorized, and therefore not assessed, according to NWEA, an additional eight standards were not represented by items within this study's sample. For MAP Growth Reading 6+, 76% of the Virginia SOLs were represented by sample items. While NWEA reported not assessing 14 SOLs, the sample items did not represent an additional 14 standards. Across all testing and grade levels, a total of 22 SOLs, in addition to those previously reported to not be assessed by NWEA, were not represented within the sample (Table 5).

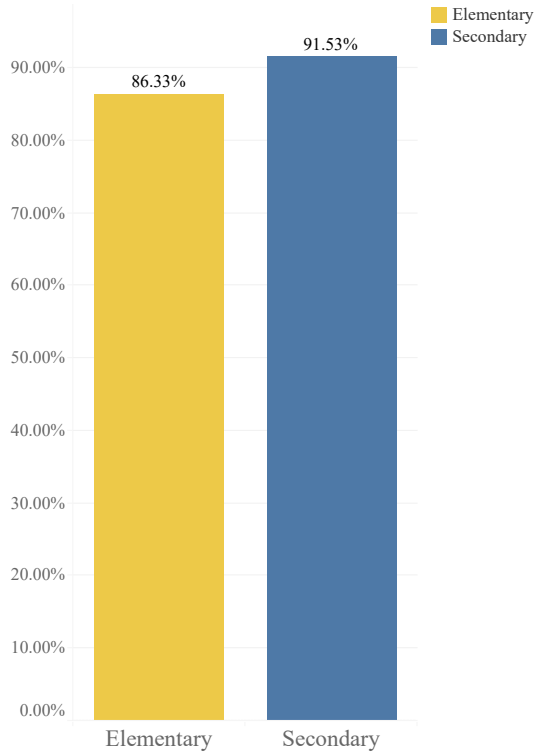
Table 5: Non-Assessed SOLs in Reading Sample by Testing Level and Grade Level

Testing Level	Grade Level	Number of Non-Assessed Standards	List of Non-Assessed Standards
Elementary	2	3	2.5.b; 2.6.a; 2.8.b
	3	1	3.5.l
	4	1	4.5.i
	5	3	5.5.e; 5.5.i; 5.5.l
Secondary	6	3	6.4.a; 6.5.i; 6.5.j
	7	4	7.4.a; 7.5.c; 7.6.j; 7.6.k
	8	4	8.5.g; 8.5.h; 8.6.g; 8.6.j
	9	3	9.4.d; 9.4.h; 9.4.k

Coverage of SOLs within the Sample of MAP Growth Items in Mathematics

The analysis in mathematics found more complete coverage of Virginia SOLs. For mathematics, there was substantial coverage of Virginia standards: 89% of the standards were addressed by MAP Growth items. As depicted in Figure 4, across the testing levels in reading, 86% of the elementary mathematics standards were represented in the sample of MAP Growth items, and 92% of secondary standards were represented.

Figure 4: SOLs Represented within the MAP Growth Item Sample Set by Testing Level in Mathematics



As shown in Table 6, there was some variation across grade levels, ranging from a low of 77% in Grade 5 to 100% in both Grade 3 and Grade 9 (i.e., Algebra).

Table 6: SOLs Represented within the MAP Growth Item Sample Set by Grade Level in Mathematics

Grade Level	Number of SOLs Aligned to MAP Growth Mathematics Sample	Total Number of Virginia SOLs	Percentage of SOLs Assessed
2	26	33	78.79%
3	34	34	100.00%
4	33	37	89.19%
5	27	35	77.14%
6	29	31	93.55%
7	23	26	88.46%
8	28	33	84.85%
9	28	28	100.00%

For MAP Growth Mathematics 2-5, 86% of the Virginia SOLs were represented by sample items. In this study's sample, 19 elementary mathematics standards were not represented by items. For MAP Growth Mathematics 6+, 91% of the Virginia SOLs were represented by items. The sample items did not represent 10 standards across secondary

grades in mathematics. Across all testing and grade levels, a total of 29 SOLs were not represented within the sample. (See Table 7.)

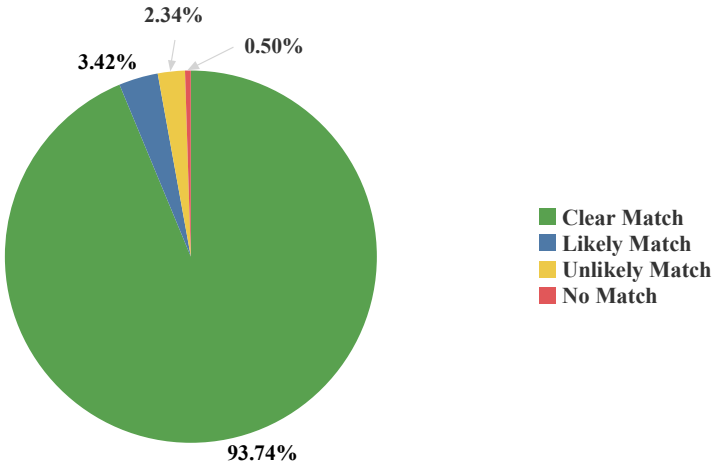
Table 7: Non-Assessed SOLs in the Mathematics Item Sample Set by SOL Strand

SOL Strand	Number of Non-Assessed Standards	List of Non-Assessed Standards
Computation and Estimation	1	5.6.b
Measurement and Geometry	6	2.8.b; 2.10.a; 2.10.b; 2.12.a; 8.7.b; 8.9.a
Number and Number Sense	6	2.3.a; 2.3.b; 4.3.d; 5.2.b; 7.1.a; 8.3.a
Patterns, Functions, and Algebra	4	5.19.a; 7.10.d; 8.15.b; 8.16.a
Probability and Statistics	10	2.14; 4.13.b; 4.13.c; 4.14.c; 5.16.c; 5.17.b; 5.17.c; 6.10.c; 6.11.a; 7.9.c

Content Alignment of MAP Growth Items to Virginia Standards of Learning

The extent to which MAP Growth items were aligned to Virginia SOLs was determined by item reviewer judgment of the match (Clear, Likely, Unlikely, or No Match) of the content of the specific item in relation to the NWEA-tagged Virginia SOL. Overall, for both subjects and across grade levels, there was substantial agreement among the raters that a very high percentage of items (n = 2,395) clearly or likely matched the NWEA-tagged standards. Across all areas and test levels, reviewers found that 97% of items clearly or likely matched the NWEA-tagged SOLs. Reviewers found that only 3% of items were unlikely to or did not match the NWEA-tagged SOLs, as depicted in Figure 5.

Figure 5: Percentage of Clear, Likely, Unlikely, and No Match of MAP Growth Items and SOLs for MAP Growth Reading 2-5, Reading 6+, Mathematics 2-5, and Mathematics 6+



Content Alignment of MAP Growth Items to Virginia SOLs in Reading

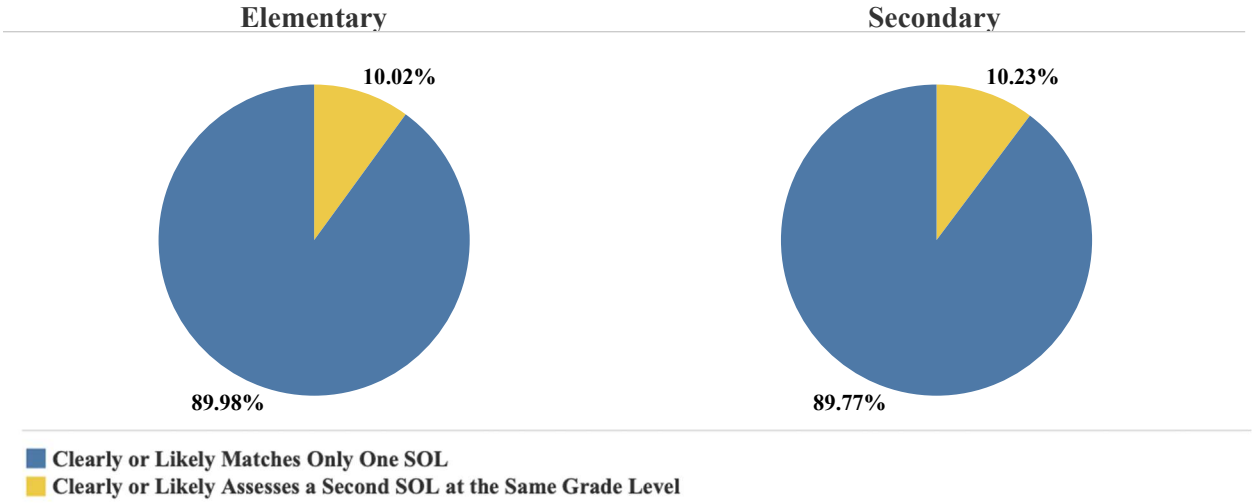
For reading, in which 1,195 items were reviewed, a clear or likely match was indicated for 97% of items at the elementary level and 98% at the secondary level. As depicted in Table 8, there was a small range across grade levels for reading, from a low of 94% for Grade 2 to a high of 100% for Grades 4, 8, and 9.

Table 8: Percentage of Reviewers' Agreement of NWEA-tagged SOLs in Reading by Grade Level

Reviewers' Agreement	Grade Level							
	2	3	4	5	6	7	8	9
Clear Match	87.33%	94.63%	96.67%	91.33%	92.62%	91.95%	99.33%	100.00%
Likely Match	5.33%	2.01%	3.33%	8.00%	4.70%	2.68%	0.67%	
Unlikely Match	7.33%	2.68%		0.67%	2.68%	5.37%		
No Match		0.67%						

Since test items that assess more than one standard can weaken the potential validity of an assessment, further analysis determined the number of items that aligned to a second SOL at the same grade level. (This analysis was delimited to considering only SOLs at the same grade level so as not to conflate the review with students' assumed prior knowledge from learning in previous grade levels.) In the total sample set for reading, only 10% of items were found to clearly or likely match a second SOL at the same grade, and, as depicted in Figure 6, similar percentages were found for both elementary and secondary testing levels. At both the elementary and secondary levels, this most commonly occurred "within" an SOL. For example, an item tagged as assessing SOL 6.6a would be identified by an SME reviewer as also clearly or likely assessing SOL 6.6b, thus not raising concerns since they are both within the same SOL.

Figure 6: Percentage of Items Clearly or Likely Matching a Second SOL at the Same Grade Level in Reading



Content Alignment of MAP Growth Items to Virginia SOLs in Mathematics

For mathematics, the review of 1,200 items showed that overall 97% were clearly or likely matched. As depicted in Table 9, there was a small range across grade levels, from a low of 93% for Grade 3 to high of 99% for Grades 2 and 8.

Table 9: Percentage of Reviewers' Agreement of NWEA-tagged SOLs in Mathematics by Grade Level

Reviewers' Agreement	Grade Level							
	2	3	4	5	6	7	8	9
Clear Match	89.33%	86.67%	90.00%	96.00%	94.00%	97.33%	98.00%	94.67%
Likely Match	10.00%	6.00%	6.67%	0.67%		0.67%	0.67%	3.33%
Unlikely Match	0.67%	4.67%	3.33%	1.33%	6.00%	1.33%		1.33%
No Match		2.67%		2.00%		0.67%	1.33%	0.67%

Given the number of strands in mathematics, the degree of match was also disaggregated by SOL strands, as shown in Table 10. The high degree of clear/likely match was evident across strands, ranging from a low of 93% for Patterns/Functions/Algebra to a high of 100% both for Functions and for Statistics.

Table 10: Percentage of Reviewers' Agreement of NWEA-tagged SOLs in Mathematics by SOL Strands

SOL Strand	Clear Match	Likely Match	Unlikely Match	No Match
Computation and Estimation	93.85%	4.10%		2.05%
Equations and Inequalities	86.79%	9.43%	1.89%	1.89%
Expressions and Operations	97.87%		2.13%	
Functions	100.00%			
Measurement and Geometry	94.72%	3.87%	1.41%	
Number and Number Sense	97.81%	0.55%	1.64%	
Patterns, Functions, and Algebra	87.02%	5.77%	5.29%	1.92%
Probability and Statistic	91.67%	2.78%	4.44%	1.11%
Statistics	100.00%			

In the analysis of math items for the presence of a second SOL at the same grade level, only 8% of the total sample set were found to clearly or likely match a second SOL at the same grade level: 5% at the elementary level and 11% at the secondary level. (See Figure 7.) Of the total, 52% were within the same standard, and 43% were across different mathematics standards. Four items were unaccounted for in the data. When analyzed by SOL strand, two stood out as having higher percentages of instances in which a second SOL was evident in the item, with those strands being Functions (36% of items) and Statistics (53% of items). (See Figure 8.)

Figure 7: Percentage of Items Clearly or Likely Matching a Second SOL at the Same Grade Level in Mathematics

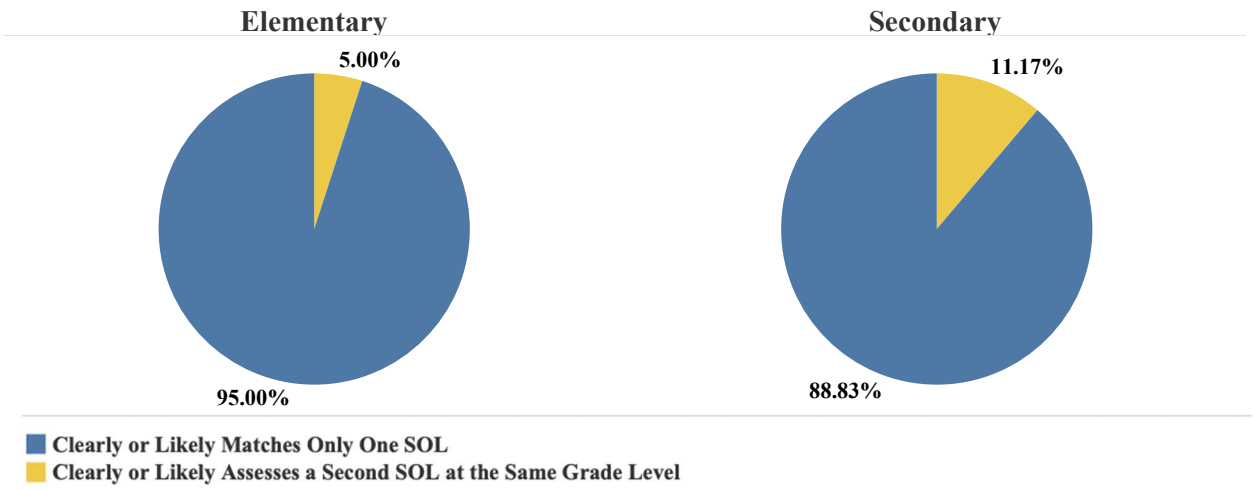
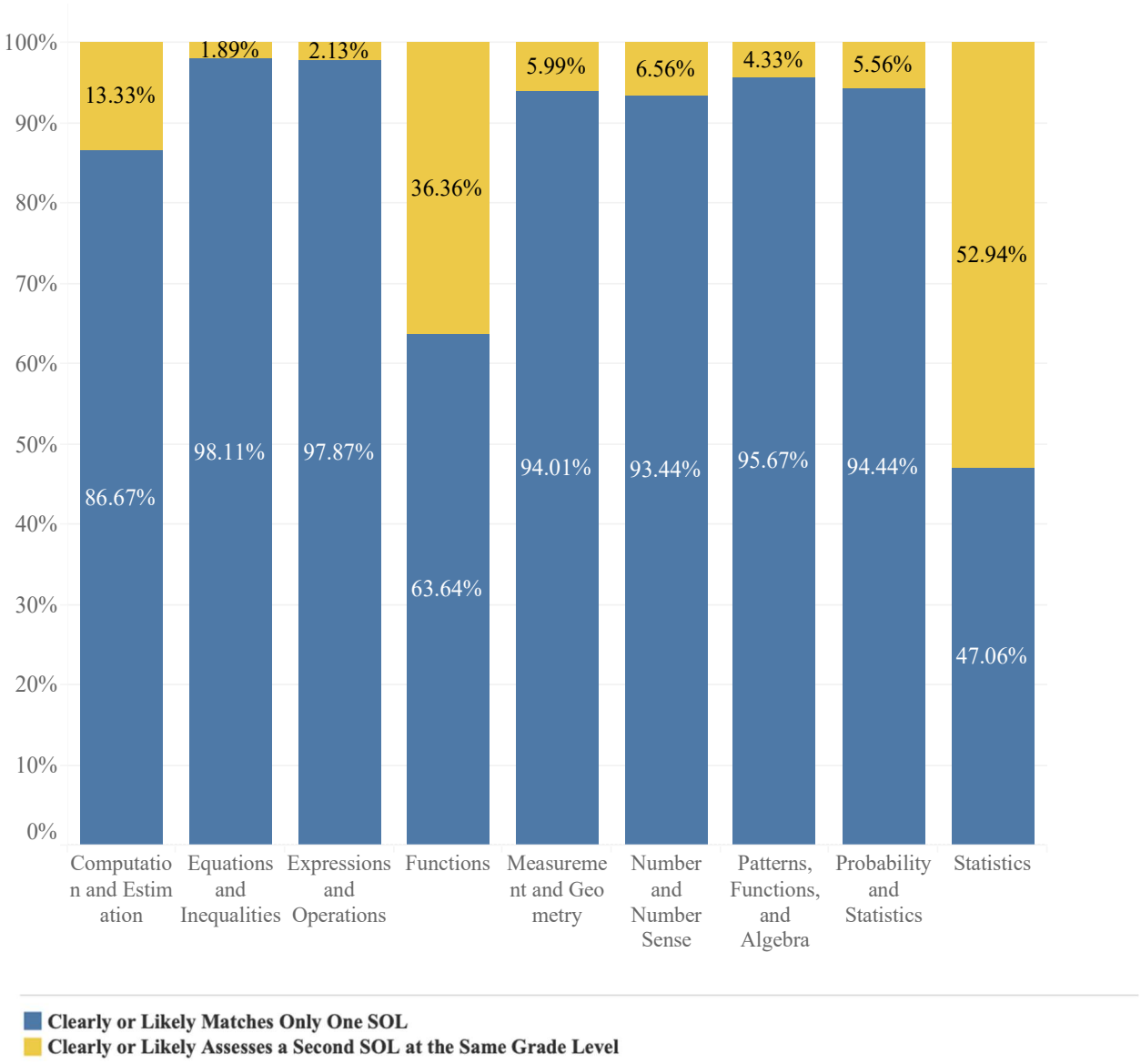


Figure 8: Relative Frequency of MAP Growth Items that "Clearly or Likely Match" a Second SOL at the Same Grade Level by Mathematics Strand



Depth of Knowledge Alignment of MAP Test Items to Virginia Standards of Learning

Reviewers' judgements of the depth of knowledge (DOK) assessed by each MAP Growth item was matched to the DOK assigned to the item by NWEA. The degree of alignment between the reviewers and NWEA was reported as percentage agreement. If there was disagreement, reviewers then judged whether items were at a lower or higher level than assigned by NWEA. Overall, across all subject areas and grade levels, reviewers reported that there was a substantial match (89%) between reviewers' and NWEA-tagged DOK levels.

DOK Alignment in Reading

In reading, reviewers found that 94% of the NWEA-tagged DOK level matched the reviewer' assigned DOK levels. This was consistent across grade levels: 93% at the elementary level and 94% at the secondary level. Table 11 presents the DOK match in reading by grade level, from a low of 87% in Grade 6 to a high of 99% in Grade 8.

Table 11: Percentage of Reviewers' Agreement of NWEA-tagged DOK in Reading by Grade Level

DOK Match	Grade Level							
	2	3	4	5	6	7	8	9
Match	90.67%	94.63%	92.67%	94.00%	86.58%	94.63%	99.33%	95.97%
Not Match	9.33%	5.37%	7.33%	6.00%	13.42%	5.37%	0.67%	4.03%

Further analysis investigated the degree to which SME reviewers judged the DOK level of items to be higher or lower than the DOK level tagged by NWEA. As shown in Table 12, reviewers of elementary items found that 9.77% of DOK 1 items could be considered higher as DOK 2. Reviewers found that 5.02% of DOK 2 items could be considered lower at DOK 1, and 1.37% of these items could be considered higher at DOK 3. Additionally, 3.57% of DOK 3 items could be considered lower at DOK 2.

Table 12: Frequency and Percentage of Reviewers' Agreement of NWEA-tagged DOK in Reading and Elementary

NWEA-tagged DOK	Reviewers' DOK	Frequency	Percentages
DOK 1	DOK 1	120	90.23%
	DOK 2	13	9.7%
DOK 2	DOK 1	22	5.02%
	DOK 2	410	93.61%
	DOK 3	6	1.37%
DOK 3	DOK 2	1	3.57%
	DOK 3	27	96.43%

As depicted in Table 13, reviewers of secondary items found that 15% of DOK 1 items could be considered higher as DOK 2, and 1.67% could be considered higher as DOK 3. Reviewers found that 1.19% of DOK 2 items could be considered lower at DOK 1, and 4.05% of these items could be considered higher at DOK 3. Additionally, 2.59% of DOK 3 items could be considered lower at DOK 2.

Table 13: Frequency and Percentage of Reviewers' Agreement of NWEA-tagged DOK in Reading and Secondary

NWEA tagged DOK	Reviewers' DOK	Frequency	Percentages
DOK 1	DOK 1	50	83.33%
	DOK 2	9	15.00%
	DOK 3	1	1.67%
DOK 2	DOK 1	5	1.19%
	DOK 2	398	94.76%
	DOK 3	17	4.05%
DOK 3	DOK 2	3	2.59%
	DOK 3	113	97.41%

DOK Alignment in Mathematics

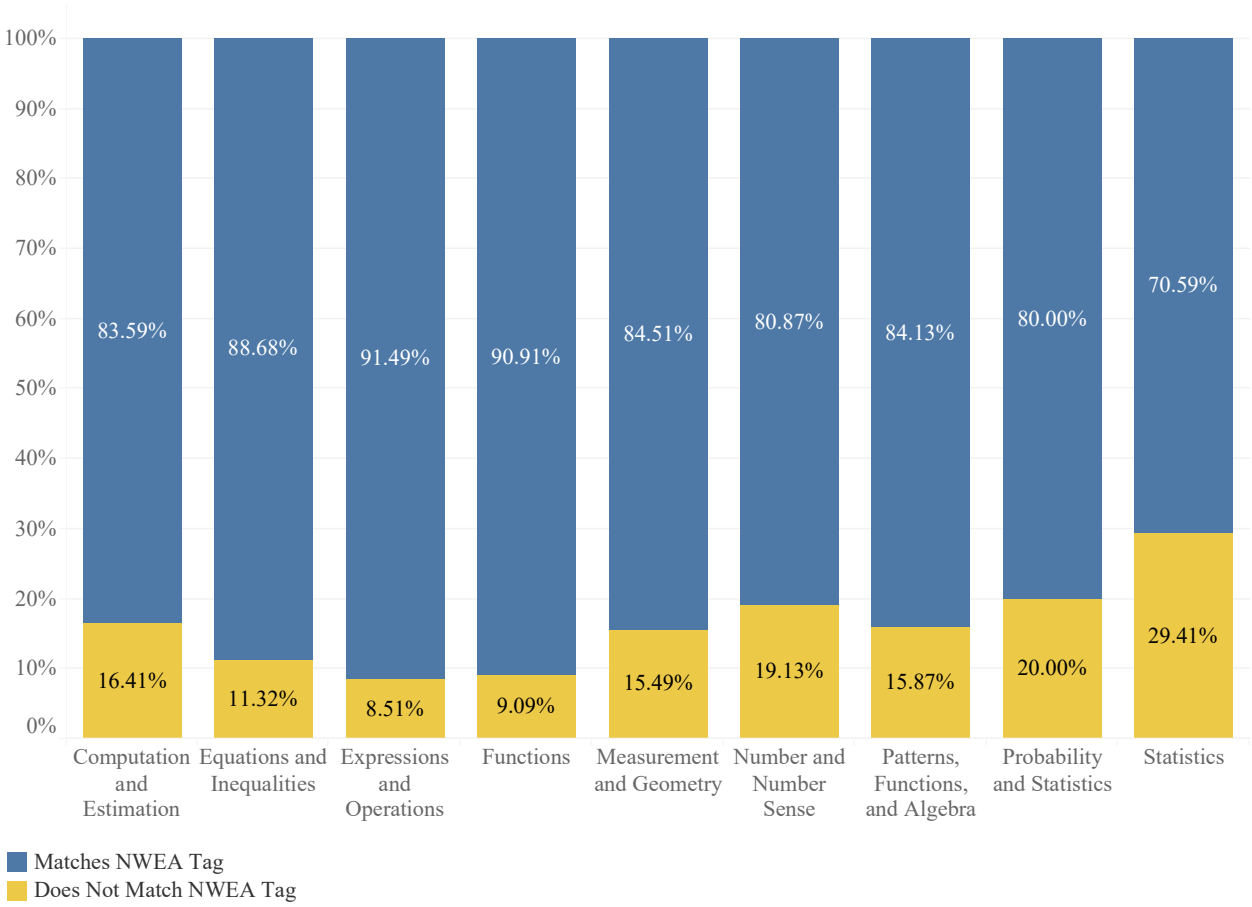
For math, overall, reviewers found that a somewhat lower, but still high, percentage of the NWEA-tagged DOK levels matched the reviewers' assigned DOK level (84%). At the elementary level, 81% showed a match, and 86% matched at the secondary level. Table 14 presents the DOK match in mathematics by grade level, from a low of 77% in Grade 3 to a high of 89% in Grade 8.

Table 14: Percentage of Reviewers' Agreement of NWEA-tagged DOK in Mathematics by Grade Level

DOK Match	Grade Level							
	2	3	4	5	6	7	8	9
Match	82.00%	76.67%	78.67%	86.00%	83.33%	84.67%	88.67%	88.00%
Not Match	18.00%	23.33%	21.33%	14.00%	16.67%	15.33%	11.33%	12.00%

Further analysis presented in Figure 9 shows the percentage match in DOK determination by SOL strand. Strongest agreement was evident in Expressions & Operations (91%) and in Functions (91%), whereas less agreement was evident in the strands of (a) Number & Number Sense (84%), (b) Probability & Statistics (80%), and (c) Statistics (71%).

Figure 9: Percentage of Reviewers' Agreement of NWEA-tagged DOK in Mathematics by SOL Strand



The degree to which SME reviewers judged DOK level of items to be higher or lower than the DOK level tagged by NWEA for elementary mathematics is shown in Table 15. Reviewers found that 20% of the DOK 1 items could be considered higher as DOK 2, and that 16% of DOK items could be considered lower at DOK 1.

Table 15: Frequency and Percentage of Reviewers' Agreement of NWEA-tagged DOK in Mathematics and Elementary

NWEA tagged DOK	Reviewers' DOK	Frequency	Percentages
DOK 1	DOK 1	321	80.45%
	DOK 2	78	19.55%
DOK 2	DOK 1	32	16.16%
	DOK 2	164	82.83%
	DOK 3	2	1.01%
DOK 3	DOK 2	3	100.00%

At the secondary level (see Table 16), 9% of the NWEA-tagged DOK 1 items could be considered higher as DOK 2, 16% of DOK 2 items could be considered lower at DOK 1,

and 2% of these items could be considered higher at DOK 3. No DOK 3 items were considered to be lower DOK levels than the NWEA tag.

Table 16: Frequency and Percentage of Reviewers' Agreement of NWEA-tagged DOK in Mathematics and Secondary

NWEA tagged DOK	Reviewers' DOK	Frequency	Percentages
DOK 1	DOK 1	250	90.91%
	DOK 2	25	9.09%
DOK 2	DOK 1	52	16.05%
	DOK 2	266	82.10%
	DOK 3	6	1.85%
DOK 3	DOK 3	1	100.00%

Phase 3: Examination of the Perceived Accuracy and Utility of Inferences Drawn by End Users of MAP Growth Reports

The final phase of the study examined K-12 practitioners' use of MAP Growth results to draw inferences about students' academic achievement and growth relative to their respective grade levels, including identifying learning gaps among students who perform *at, above,* and *below* grade level. Data were collected through an online survey of SME reviewers and two focus groups of selected SME reviewers and division-level leaders, respectively.

Survey Findings

The survey was completed by 19 of the 21 SME reviewers. Three questions on the survey were germane to the examination of the perceived accuracy and utility of inferences from MAP Growth results. Respondents indicated the degree to which they agreed or disagreed with the statements. Table 17 presents the percentages of respondents responding to different degrees of agreement. Regarding respondents' judgement about the accuracy and utility of MAP Growth assessments for identifying gaps in student learning, 84% indicated that they either agreed or somewhat agreed. Regarding respondents' confidence in using MAP Growth results to group students for instruction, 74% agreed or somewhat agreed. Regarding the accuracy of MAP Growth in determining student performance relative to at, below, or above grade level, 84% agreed or somewhat agreed.

Table 17: Percentage of Respondents Indicating Degree of Agreement with Statements about the Accuracy and Utility of MAP Growth Results (*n*=19)

	Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Disagree	No Basis for Judgement
A student's results on a MAP Growth assessment can be used to identify gaps in their learning.	68.4%	15.8%	0%	10.5%	5.3%	0%
I would feel confident grouping students for instruction based on their MAP results	57.9%	15.8%	5.3%	15.8%	5.3%	0%
A student's results on the MAP Growth assessment is an accurate indicator of whether that student is performing on, below, or above grade level.	57.9%	26.3%	0%	10.5%	5.3%	0%

Focus Group Findings

Two focus groups were also conducted by the study leads. The first focus group was comprised of seven selected SME reviewers, each of whom had substantive experience with MAP Growth assessments. The second focus group was comprised of six division-level leaders, each of whom held some degree of responsibility for directing the implementation of MAP Growth assessments in their respective school divisions. Participants' experience with MAP Growth assessments ranged from 1 to 5 years, with an average of 2.64 years. Although the focus groups were conducted separately, they were provided the same written questions in advance, and these questions served to structure the focus group discussion. (See Appendix B.)

The focus groups were designed to examine the perceived accuracy and utility of the MAP Growth reports in drawing inferences about student learning. MAP Growth assessments generate 19 different reports (two of which distinguish multiple "views" that convey different information and analyses). The MAP Growth reports are listed below, and it was the perceived accuracy and utility of these reports about which focus group participants were asked:

1. Class Report
2. Grade Report
3. Class Breakdown By RIT
4. Class Breakdown by Goal
5. Grade Breakdown Report
6. Class Breakdown by Projected Proficiency
7. Projected Proficiency Summary
8. District Summary: Aggregate by School
9. District Summary: Aggregate by District
10. Learning Continuum: Display Options
    a. Test View, Grouped by Standard
    b. Test View, Grouped by Topic
    c. Class View, Grouped by Standard
    d. Class View, Grouped by Topic
11. Achievement Status and Growth Projection
12. Achievement Status and Growth Summary
13. Achievement Status and Growth Summary with Quadrant Chart
14. Student Growth Summary
15. Student Goal Setting Worksheet
16. Student Progress Report
17. Student Profile Report
    a. Comparisons
    b. Instructional Areas
    c. Growth Goals
18. Family Report
19. Retest Recommended: Rapid Guessing Report

The study leads took detailed notes during the discussion, and then participants provided their own individual written responses to the questions following the focus groups. The study leads

reviewed the notes and written responses to identify clear themes. One initial finding of note was the similarity in responses both within and between the two focus groups. Thus, the presentation of findings does not distinguish between the two focus groups. It is important to note, though, that focus group responses do not necessarily represent consensus among participants; rather, they represent contributions to the discussion that were stated, explained, (in some cases) discussed by multiple participants, and (importantly) not contested by another participant.

A second finding of practical significance is that all focus group participants indicated that they deem MAP Growth reports to be both accurate and useful. While there were varying insights, experiences, and questions shared by participants, the perception was consistently positive. However, it is important to reiterate that the purpose of the focus group was to gauge end-users' perceptions of the accuracy and utility of MAP Growth reports. Therefore, potential misuses or misunderstandings were not necessarily corrected during the sessions nor in the findings presented here.

What report(s) do you/would you use to determine whether a student is on, below, or above grade level?

Focus group participants identified a number of reports they prefer to use to determine student performance relative to grade level. They also expressed several cautions in using the reports. The reports are enumerated below, each followed by brief statements gleaned from respondents' comments about the accuracy and/or utility of each report. Note: Statements in parentheses indicate a proviso, caution, or limitation of a report, as shared by one or more participants.

1. Class Report

    - Provides class breakdown by RIT
    - Provides "a good band breakout"
    - Can rank order by RIT or percentile
    - Shows performance relative to classroom peers and to norms
    - Provides goal and projected proficiency
    - Shows low, average, or high performance for teachers (which are interpreted as below, at, or above grade level)
    - Facilitates grouping for instruction
    - Particularly useful for teachers (but may be less useful for other "audiences")
    - RIT score can be used to identify students for summer school
    - "Color-coded display makes this a particularly easy data point to interpret for end users as it groups students according to achievement, from low to high"
    - At the building level, can help teachers "see students' strengths and weaknesses in an at-a-glance manner"

2. Class Breakdown By RIT

    - Can be used to determine whether students are “on grade level”

3. Class Breakdown by Projected Proficiency

    - Can be used to make quick determinations of progress relative to benchmark expectations

4. Student Profile Report

    - Provides clear and helpful visualization of data for quick and accurate interpretation
    - Shows student growth or progress over time (assuming multiple administrations)
    - Curriculum coordinators deem it helpful with respect to alignment to curriculum pacing guide
    - Students can use to evaluate their own growth
    - Useful in professional learning community (PLC) meetings to discuss grouping for enrichment and intervention
    - Helpful to classroom teachers when designing personalized learning for students
    - Effective tool to assist with planning for parent conferences
    - Helpful in RtI (response to intervention) meetings “to support assertions that students are exhibiting difficulties in specific areas, as specific SOLs are highlighted as relative strengths and weaknesses”

5. Student Progress Report

    - Provides detailed information that can be used to identify gaps in skills
    - Shows student performance relative to division and national norms, as well as project RIT score
    - The option to drill down to standards with this report makes it even more useful

6. Achievement Status and Growth Summary with Quadrant Chart

    - Provides an overview of class and individual student growth
    - Growth data can help building administrators “see” strengths and weaknesses of teachers from a “bird’s eye view” (explicitly using the Quadrant Report, especially to see the relationship between growth and achievement)
    - Determining class growth facilitates determination of instructional needs of students and decision-making about remediation and enrichment

- May be helpful to building administrators in grouping students into classes for next instructional year

As a follow-up, participants were asked, "How does information based on national norms help to gauge on, below, or above grade level?" The ensuing discussion led to some participants articulating that they exercise caution in interpreting in terms of "grade level" and "performance level" because MAP Growth does not definitively determine *at, above,* or *below* grade level. Some respondents shared a concern that using the language of "grade level" can be a misnomer. Some stated that interpreting results using "percentiles" and "growth" was more useful to them.

What report(s) do you/would you use to determine specific gaps in student learning?

Participants identified several MAP Growth reports for determining specific gaps in student learning. Importantly, many participants explained that they used these three reports in tandem, rather than using only one or another of them.

1. Class Report

    - Can be used to identify "glows & grows" for students
    - Provides an overview of student performance
    - Provides breakdown of performance by instructional areas (which are similar to SOL "strands") relative to national norms, facilitating instructional decisions regarding remediation and enrichment
    - Reveals weaknesses in content areas' reporting categories for the entire class
    - Can serve as "a teacher's springboard for delving into the pool of information on specific gaps in student learning"

2. Student Profile

    - Can be used to identify "what is lacking" in a student's performance
    - Breaks down RIT score into instructional areas and associated standards
    - Can identify areas of strength and areas of need, and lists relevant standards within each instructional area
    - Helpful for working with teachers to identify areas of need and to determine student grouping for enrichment or remediation
    - "Interactive format is user-friendly and helps instructors examine a myriad of factors that combine to create this learning litmus test of sorts"

3. Student Goal Setting Worksheet

    - Facilitates identifying specific needs and developing individualized student learning plans

- Can be used with students and parents for identifying and setting goals
- Can be used to work with students "individually about whether they made their goal or not, and what their weaknesses and strengths are"

4. Learning Continuum (multiple views)

    - Provides analysis by instructional areas and lists relevant standards to the respective instructional areas (but may not identify specific standards of need and therefore should be used in conjunction with other commercially produced or division-made formative assessments)
    - Provides for depth of analysis (although "teachers can go too far into the weeds with their analysis," thus limiting utility)
    - Identifies "specific gaps" by instructional areas
    - Useful in setting goals
    - Useful in grouping students for remediation and intervention
    - Provides "standard-specific information" about each student's performance and displays it in the aggregate
    - (MAP Growth Assessments only report on skills when a certain number of students in a population have been tested on that skill,so analysis may be limited in smaller schools/divisions or early in implementation)

Notably, several participants explained that a number of the reports are useful to analyze "vertically and horizontally," meaning student performance, needs, and goals both across a grade level, as well as between grade levels. The latter can aid long-range curricular and instructional planning for a school division.

What report(s) do you/would you use to make near-term instructional decisions?

The utility of assessment results is premised in the judgements and actions that they facilitate. Focus group participants identified a number of reports that support near-term instructional decisions, such as those made with an eye towards days, weeks, or a marking period.

1. Class Report

    - Review of RIT ranges and "descriptors" is useful in planning "Tier 2 instruction"
    - Provides a "quick snapshot" to show low, average, and high in the instructional areas
    - Allows for decisions about grouping for remediation and enrichment
    - Provides an "an overview of the current status of students performance and areas of strength and weakness"
    - "Helps organize students by RIT for immediate information as to how students performed on the assessment"

2. Student Profile Report

    - Allows focus of student needs in terms of specific instructional areas
    - Provides a "clear picture of what is lacking" in a student's performance
    - Interactive display features allow for more targeted analysis of student performance
    - Can be used to gauge whether students "rushed on the test"

3. Learning Continuum (multiple views)

    - Helpful "on a student-by-student basis"
    - Provides "information specific to which standards in each instructional area that the students are ready to learn"
    - Helpful in targeting specific skills that can be focused on for small group or whole group instruction
    - "Very detailed information regarding standards with which a student is excelling or struggling"—provides information specific to "actual content and skills based on the standards"

4. Achievement Status and Growth Summary with Quadrant Chart

    - Provides specific information relative to low and high achievement, as well as low and high growth (four "quadrants")
    - Visual display is intuitive for teachers
    - Distinguishes "growth" from "achievement"
    - Can use scores to make scheduling decisions, such as accelerating underrepresented students
    - Allows for decisions about pullout or intervention, when used in conjunction when other, division-made common assessments
    - Useful in identifying and addressing weaknesses for small-group, Tier 1 instruction
    - Particularly helpful this year with virtual learning, information about both achievement and growth facilitated instructional decision-making by teachers
    - (Shares a great deal of useful information, but some teachers report that the quadrants are "off" for some students—either too low or too high)

As a follow-up, participants were asked who makes use of these reports for instructional decision-making. Responses included teachers, instructional specialists, and curriculum leaders. Participants also indicated that review of reports and instructional decision-

making was undertaken by individuals, but perhaps more frequently undertaken by PLCs, grade-level teams, departments, or other formal groups or teams at the building level.

Respondents were also asked, "In instances when analysis of MAP Growth results indicate the need for remediation, does this create a time conflict with the local pacing guide? If so, how do you address that?" Several participants responded to this, and two key points emerged:

- One school division representative explained that their pacing guide has time for remediation "built in"
- A representative from another school division noted that analysis of MAP Growth results have prompted them to review their pacing guide to ensure that student have had adequate opportunity to learn, which may result in a longer-term revision of the pacing guide itself

What report(s) do you/would you use to project students' performance on end-of-year SOL tests?

Participants indicted two reports upon which they rely to gauge students' projected performance on end-of-year SOL tests.

1. Student Profile Report

    - Includes projected performance for each student on the upcoming SOL assessment

2. Class Breakdown by Projected Proficiency Report

    - Provides projection to both the SOL and the ACT/SAT
    - Indicates whether "students are or are not on track to pass, or if they are on track to pass advanced"
    - Given recent revisions to the Reading SOLs, projected proficiencies were not yet available for reading, although two respondents from two different decisions indicated different judgements based on their own analysis
        - One division through their own analysis found MAP Growth to be "reasonably predictive"
        - A representative of a different division found the MAP Growth "not as aligned to our actual results" as in previous years

As a follow up, participants were asked, "Since MAP Growth is set to national norms, are there any disconnects in your analysis at the local level in Virginia?" As a general consensus, respondents did not view the use of national norms as problematic in their analysis and use of MAP Growth results.

What, if any, other assessments do you use in place of or in addition to MAP Growth for any of the above purposes, and what do these assessments provide to complement or confirm information garnered from MAP Growth assessments?

As a general point of consensus, participants indicated that their school divisions use a number of other assessments to complement MAP Growth results. Such assessments include commercially produced, standardized assessments (e.g., Phonological Awareness Literacy Screening [PALS], Developmental Reading Assessment [DRA]); locally developed division-level benchmark assessments; and "common assessments" (such as those developed by grade-level and departmental teams of teachers). A common conveyed perception was that MAP Growth assessments do not assess the full scope of the curriculum and also that it is poor practice to base high-stakes instructional decisions on a single assessment. Despite this, it was also evident that there was broad-based consensus that MAP Growth assessments provided particularly accurate, useful information about student achievement and growth.

Other Relevant Points Regarding the Accuracy and Utility of Reports

Several additional points arose during the focus groups that are relevant to the consideration of the accuracy and utility of MAP Growth reports. While there were not necessarily consensus opinions, others in the groups did not dispute them and, in some instances, the points generated some discussion as focus group participants occasional attempted to learn from others' insights.

1. Several respondents indicated that the Family Report is particularly helpful in conveying important information about student achievement and growth to students and to their family members.
2. To reiterate a point made at the introduction to the Phase 3 findings, respondents accentuated the need to use multiple reports to serve different purposes. Of note, no respondent indicated that they make use of all 19 reports. Several respondents, though, indicated that this could be from lack of experience and the considerable amount of information provided in many of the most used reports.
3. Several respondents indicated a desire to have access to report information that would "drill down to" the level of specific standards of learning (since MAP Growth reports present results at the instructional area level).
4. Several respondents noted that the instructional areas of MAP Growth do not directly align to either the SOL "Strands" or to SOL "Reporting Categories." These the differences, though, are not severe and do not present a significant issue to those aware of the differences in conceptual organization and wording. Nonetheless, several participants were not previously aware of the issue and appreciated learning about it during the focus group so that they can address it in practice.
5. Some participants noted that the estimates of the time range of testing events ranged from 20 to more than 160 minutes, the latter greatly exceeding estimates reported by NWEA.

6. All respondents indicated that training in the content, delimitations, and uses of the various MAP Growth reports is essential for teachers, specialists, instructional leaders, building-level leaders, and division level leaders.
7. Several respondents indicated modest concern that MAP Growth reports are based on an algorithm that may draw an inference about a student's achievement and growth relative to an SOL when, in fact, the student was not presented an item on that SOL during a testing event. This was of concern, but once understood, was not deemed problematic by participants.
8. NWEA does not advocate the use of MAP Growth results for purposes of teacher evaluation. However, all respondents indicated that, as a matter of practice, MAP Growth is used by in their respective school divisions to create SMART goals for purposes of teacher supervision and evaluation. Notably, many respondents questioned the appropriateness of this division-directed practice given the stated purposes of MAP Growth assessments.

Discussion and Conclusions

The primary purposes of this investigation, to determine the alignment of MAP Growth assessments to Virginia SOLs and the use of scores to estimate performance relative to grade level and learning gaps, was essentially a validity investigation that focused on gathering evidence for the appropriate use of MAP Growth data in Virginia. The sources of evidence for the validity argument included a conceptual component, consisting of a review of relevant literature and NWEA documents, a quantitative analysis of alignment based on reviewer judgements, and a qualitative component that gathered user perspectives. The information and data gathered provided for a comprehensive analysis from different sources, resulting in triangulation that supports the major findings and recommendations.

It was clear from NWEA documents and other literature that the MAP Growth assessments are technically sound measures that provide reliable scores with some degree of predictability from one testing time to another. The data reported are most reliable at the total score and instructional area levels, not at the instructional sub-area levels. This suggests that the validity argument is strongest for total and instructional area scores. Furthermore, the nature of MAP Growth assessments is to use items and item response theory (IRT) modeling to estimate proficiency based on national norms. Predicted probabilities of level of performance are used for standards on which an indivicual student may not have been directly assessed. This procedure for estimating performance results in some degree of error, which, when combined with error associated with internal consistency and stability over time, a lack of inclusion of all SOLs, and an imperfect match between curriculum taught and the tests, suggests that caution is needed in interpreting the results.

As a general conclusion, MAP Growth items have strong alignment to Virginia's SOLs. However, clearly, MAP Growth assessments are not substitutes for SOL tests. MAP scores may be considered proxies for SOLs, but an important finding and recommendation of this study is that other evidence of student performance must be used in conjunction with MAP results to verify levels of proficiency. MAP Growth assessments are more like traditional standardized achievement tests than competency-based assessments. While both math and reading MAP assessments have very good alignment with a majority of SOLs, both in relation to content and depth of knowledge as verified through the systematic review of thousands of MAP items, it is essentially an external test that is best utilized to provide one source of information about student proficiency, not an absolute indicator of student performance relative to the SOLs.

User feedback about MAP assessments suggests that the scores and various reports that are available are viewed as informative and helpful in identifying possible learning gaps and instructional needs. However, this finding is limited to those participating in the focus groups and completing opinion surveys and may not reflect perceptions of a wider population of teachers, building-level leaders, and division-level administrators. Because MAP reports use unique terminology, difficult to comprehend RIT scores, and many different types of reports, it is important to align targeted use of results with the appropriate report and to provide adequate training and support to assure appropriate use of the results. NWEA-provided support and training appear to be adequate to good, at least for the four school divisions in this study.

The evidence gathered in this study supports the use of MAP Growth assessments as a valid indicator of growth over time as long as it is recognized that this is primarily a measure of ability in targeted areas defined by NWEA, in part aligned to Virginia standards. MAP Growth assessments may provide some indication of proficiency on SOLs, learning gaps, strengths and weaknesses, and areas that need further or improved instruction, but such conclusions must be verified with other assessment results, particularly for instructional sub-areas, with interpretations that include standard errors of measurement.

It is problematic, based on the nature of MAP Growth assessment results, to classify performance in terms of "at," "below," or "above" grade level. In particular, when applied to a specific division, school, class, or student, such determination that uses national normative data from NWEA may not be matched well with local curriculum, instruction, and student populations. This conclusion was verified by those participating in the study's focus groups.

This study is limited to the personnel and experiences of four Virginia school divisions, the use of NWEA-determined matches of items to SOLs and depth of knowledge, and to the items provided by NWEA. Results using a different methodology (e.g., asking raters to generate SOL matches rather than verify what was determined by NWEA) could result in different findings.

The validity of MAP Growth assessments—that is, essentially, the inferences and uses of the scores—depends on a thorough understanding of the nature and limitations of the tests and alignment of reports with local needs and purposes. MAP Growth assessments cannot replace what is measured by SOL tests. Nevertheless, MAP Growth tests are soundly constructed, well-aligned assessments that provide reliable scores, which, when used with appropriate cautions and limitations, can be useful in identifying different levels of performance and gaps in learning in Virginia.

Works Cited

Burns, M. K., & Young, H. (2019). MAP test review. *Journal of Psychoeducational Assessment, 37*(5). 665-668. https://doi.org/10.1177/0734282918783509

Hess, K. (2013). *A guide for using Webb's depth of knowledge with Common Core State Standards.* Common Core Institute. Retrieved from https://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/How-to-Design-and-Select-Quality-Assessments/Webbs-DOK-Flip-Chart.pdf.aspx.

Jones, S. L. (2015). *Predicting performance on the Georgia criterion referenced competency test based on the Measures of Academic Progress in mathematics and reading.* [Unpublished doctoral dissertation]. Capella University.

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. A. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, *52*(5), 500-514. https://doi.org/10.1002/pits.21839.

Mitchell, M. (2019). *Academic screening in middle school: How well do AIMSweb Measures of Oral Reading Fluency, and NWEA Measures of Academic Progress, predict future performance on state exams.* [Unpublished educational specialist degree thesis]. University of Wisconsin—Eau Claire.

NWEA. (2016, March). *Linking the Virginia SOL assessments to NWEA MAP Growth tests.*

NWEA. (2019a, March). *MAP Growth technical report*.

NWEA. (2019b, March). *MAP Growth instructional areas: Standard alignment—Virginia reading (Reading 2-5 VA 2017; Reading 6+ VA 2017).*

NWEA. (2020a). *Linking study report: Predicting performance on the Virginia Standards of Learning (SOL) mathematics assessments based on NWEA MAP Growth scores.*

NWEA. (2020b, February). *MAP Growth instructional areas: Standard alignment—Virginia mathematics (Math 2-5 VA 2016).*

NWEA. (2020c, February). *MAP Growth instructional areas: Standard alignment—Virginia mathematics (Math 6+ VA 2016).*

NWEA. (2021). *MAP Growth reports portfolio V 2.0.*

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Measurement in Education, 20*(1), 7–25. https://doi.org/10.1080/08957340709336728

Additional References

Case, B. J., Jorgensen, M. A., & Zucker, S. (2004, December). *Alignment in educational assessment.* Pearson.

Christopherson, S. (2019, July). *Comparative alignment analysis of four interim assessment programs with Oklahoma State ELA, mathematics, and science academic standards.* University of Wisconsin–Madison Wisconsin Center for Education Research.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use.* Routledge.

Egan, K. L., & Davidson, A. H. (2018, December). *Alignment of the NWEA MAP Growth & MAP Growth K-2 to the 2018 Oklahoma standards—English language arts & mathematics.* [Unpublished report]. EMetric.

Gareis, C. R., & Grant, L. W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (2nd ed.). Routledge.

McMillan, J. H. (2017). *Classroom assessment: Principles and practice that enhance student learning and motivation.* Pearson.

Webb, N. L. (2009). *Design of content alignment studies in mathematics and reading for 12th grade NAEP and other assessments to be used in preparedness research studies.* National Assessment Governing Board.

Wise, S. L., Kingsbury, G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practices, 34*(4), 41-48.

.

Appendix A:

Item Review Protocol Guide

Item Review Protocol Guide

Expectations for Handling Proprietary Materials

Proprietary materials are materials that are owned by someone, namely NWEA. As a reviewer in the study, you have been granted permission to access a sample of MAP Growth items and metadata, but it would be a violation of your executed non-disclosure agreement to use the actual materials or information about them for any purpose other than the intended purpose of the study. With that in mind, you must comply with the following expectations:

1. Do not copy any item in whole or in part.
2. Do not distribute any item in whole or in part.
3. Do not allow anyone access to the Review Portal, except for lead members of the study team.
4. Do not discuss specific items outside of our Study Team.

Keep in mind that the published report will not refer to specific items but will analyze and draw inferences from our data collection in the aggregate. Discussion, presentation, and use of findings in the final report will be acceptable.

Anticipated Questions About the Item Review Protocol

1. **Are we analyzing alignment to the Standards themselves or to all elements of the Curriculum Framework?** MAP Growth items are pegged to Standards, not to discrete knowledge, understandings, skills, or dispositions articulated in the Curriculum Framework. Therefore, we are analyzing the match to the Standards statements. Nevertheless, analyzing match requires you to exercise your subject matter expertise, which includes your knowledge of the Curriculum Framework and the vertical articulation of the curriculum.
2. **Which Standards of Learning documents are we using?** We are using 2016 Mathematics SOLs and 2017 Reading (English Language Arts) SOLs.
3. **What constitutes an item for us?** As a computer-adaptive test (CAT), MAP Growth assessment items are technology-enhanced, allowing the use of different types of formats, including traditional multiple-choice, binary-choice in reading passages, multiple-correct choice, drag-and-drop, “click-and-pop,” and text-entry (constructed-response). Regardless of format, items have a prompt, which might be in the form of a question or a statement. All questions have a means of response, such as select-response or constructed-response. Many items have an extended prompt, a stand-alone reading passage, or stimulus material (e.g., graph, table, diagram). All elements together--prompt, stimulus material, and response format--constitute an item. The prompt is the core element of an item, but the other two elements should be considered.
4. **How many items will I review?** Reviewers will have 100, 125, or 150 items total to review. Our total sample is 1,200 per subject area, for a grand total of 2,400 items.
5. **How long should I spend on each item?** We suggest spending no more than 5 minutes per item.

6. **What if I experience a situation that significantly affects my ability and/or availability to complete the review?** For any immediate or pressing concern or question, contact a Study Lead, Lead SME for your subject area, or Steering Committee member.
7. **What if I have a question about an individual item, the spreadsheet, or some other step of the review process?** For questions about specific items, we suggest waiting until you have reviewed a subset of 20-25 items and then contacting your Lead SME for an inter-rater agreement check, which is also a good time to discuss a specific item. When in doubt, contact your Lead SME.
8. **When am I to do the work?** Outside of your regular "contract" hours.
9. **What is the deadline for the review?** The end of the day Monday, April 12.

Item Review Questions and Operational Definitions

The item review is driven by the following three questions. You will evaluate the item with respect to each question and indicate your judgement by selecting a choice from the dropdown menu. Following each question below are descriptors to help operationalize the response.

1. **"To what degree does the item MATCH the NWEA-tagged SOL?"**

   See table that follows.

	4 Ways to Operationalize the MATCH Scale			
LEVEL	**Inferential**	**Dialogic**	**Level of Confidence**	**Archery Analogy**
CLEAR MATCH	I could draw a **definite** inference about a student's mastery of this SOL based on this item.	Yes	I'm **90-100%** confident that there is a match.	Bull's Eye
LIKELY MATCH	I could draw a **tentative** inference about a student's mastery of this SOL based on this item.	Yes, **but…**	I'm **70-89%** confident that there is a match.	On Target
UNLIKELY MATCH	I would **not depend upon** this item to draw an inference about a student's mastery of this SOL.	No, **unless...**	I'm only **40-69%** confident that there is a match.	Aiming in the Right Direction
NO MATCH	I would **reject** an inference about student mastery of the SOL based on this item.	No.	I'm **less than 40%** confident that there is a match.	Potentially harmful to those around you!

2. **To what degree does the item clearly MATCH another SOL at the same grade level?**
    a. The intent of this question is to gauge the preponderance of items that address an SOL other than the SOL tagged to the item by NWEA.
    b. We are *not* seeking to identify more than one additional SOL--only one, if any.
    c. Your response will essentially be "yes/no," as indicated by these two possible responses:
        i. "CLEAR OR LIKELY MATCH" = yes
        ii. "UNLIKELY OR NO MATCH" = no

3. **What is the DOK level of the item?**

See table that follows

DOK Level	Description	Examples of Associated Cognitive Activity
Level 1: Recall & Reproduction	Curricular elements that fall into this category involve basic tasks that require students to recall or reproduce knowledge and/or skills. The subject matter content at this level usually involves working with facts, terms, details, calculations, principles, and/or properties. It may also involve use of simple procedures or formulas. There is little or no transformation of the target knowledge or skill required by the tasks that fall into this category. A student answering a Level l item either knows the answer or does not; that is, the answer does not need to be figured out" or "solved.	Locate, calculate, define, identify, list, label, match, measure, copy, memorize, repeat, report, recall, recite, recognize, state, tell, tabulate, use rules, answer who, what, when, where, why, how
Level 2: Working with Skills & Concepts	Level 2 includes the engagement of mental processing beyond recalling, reproducing, or locating an answer. This level generally requires students to compare or differentiate among people, places, events, objects, text types, etc.; apply multiple concepts when responding; classify or sort items into meaningful categories; describe or explain relationships, such as cause and effect, character relationships; and provide and explain examples and non-examples. A Level 2 "describe or explain" task requires students to go beyond a basic description or definition to predict a possible result or explain "why" something might happen. The learner makes use of information provided in context to determine intended word meanings, which tools or approach is appropriate to find a	Infer, categorize, organize and display, compare-contrast, modify, predict, interpret, distinguish, estimate, extend patterns, interpret, use context clues, make observations, summarize, translate from table to graph, classify, show cause/effect, relate, edit for clarity, make basic inferences

	solution (e.g., in a math word problem), or what characteristics to pay attention to when making observations. At this level, students are asked to transform/process target knowledge before responding.	
Level 3: Short-Term Strategic Thinking	Tasks and classroom discourse falling into this category demand the use of planning, reasoning, and higher order thinking processes, such as analysis and evaluation, to solve real-world problems or explore questions with multiple possible outcomes. Stating one's reasoning and providing relevant supporting evidence are key markers of DOK 3 tasks. The expectation established for tasks at this level require an in-depth integration of conceptual knowledge and multiple skills to reach a solution or produce a final product. DOK 3 tasks and classroom discourse focus on in-depth understanding of one text, one data set, one investigation, or one key source, whereas DOK 4 tasks expand the breadth of the task using multiple texts or sources, or multiple concepts/disciplines to reach a solution or create a final product *Note: On CATs that do not allow for extended constructed responses (e.g., short-answer, essay, model creation), Level 3 thinking must be **inferred** from the nature of the mental operations necessary to engage with the prompt, stimulus material, and response format/choices.*	Critique, appraise, revise for meaning,assess, investigate, cite evidence, test hypothesis, develop a logical argument, use concepts to solve non-routine problems, explain phenomena in terms of concepts, draw conclusions based on data
Level 4: Extended Strategic Thinking	Curricular elements assigned to this level demand extended and integrated use of higher order thinking processes such as critical and creative-productive thinking, reflection, and adjustment of plans over time. Students are engaged in conducting multi-faceted investigations to solve real-world problems with unpredictable solutions. Employing and sustaining strategic thinking processes over a longer period of time to solve the problem or produce an authentic product is a key feature of curricular objectives assigned to DOK 4. Key aspects that denote this particular level typically include authentic problems and audiences,	Initiate, design and conduct, collaborate, research, synthesize, self-monitor, critique, produce/present

	and collaboration within a project-based setting. *NOTE: An inherent limit of the current state-of-the-art of standardized assessment is that they cannot tap Level 4 thinking. Therefore, there are* ***no Level 4-tagged items*** *on the MAP Growth tests.*	

Adapted from Hess, K. (2013). A Guide for Using Webb's Depth of Knowledge with Common Core State Standards. Common Core Institute. Retrieved from https://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/How-to-Design-and-Select-Quality-Assessments/Webbs-DOK-Flip-Chart.pdf.aspx.

Step-by-Step of Item Review

To review items, you will need to be logged into the NWEA Review Portal and you will need access to your Item Review Spreadsheet. Note: You will not enter any information in the NWEA Review Portal; rather, all information will be entered into your Item Review Spreadsheet.

1. Locate the item in the Review Portal >>> Enter the **NWEA Item Number** into your spreadsheet

2. Locate the NWEA-tagged SOL number >>> Enter **NWEA-tagged SOL** into the spreadsheet.

    1. If the NWEA identifies a sub-element of the SOL with an alphabetic identifier, include that as well (e.g., 3.4.c).

3. Making use of the Curriculum Framework and your subject matter expertise (SME), determine the li**kelihood of the MATCH of item to the NWEA-tagged SOL** >>> Select from the dropdown menu to indicate your evaluation of alignment in the spreadsheet.

    a. If you indicate either "Likely Match" or "Unlikely Match," then include a brief **comment or make notations** using keywords or short phrases that give some indication of your reasoning.

    b. Comments do *not* need to consist of complete sentences and should *not* be lengthy explanations. The most important consideration is that comments or notations have sufficient detail that *you* could use them to verbally explain your reasoning to a Study or SME Lead at some later date, if asked.

4. Making use of the Curriculum Framework and your subject matter expertise (SME), determine the likelihood that the item is a **CLEAR MATCH OR LIKELY MATCH to one other SOL at the same grade level** >>> Select either "CLEAR OR LIKELY MATCH" or "UNLIKELY OR NO MATCH" from the dropdown menu.

a. **A useful way to judge this** is to ask yourself, "Is success on this item clearly dependent upon skills or understanding indicative of a different SOL at the same grade level?"

b. **Possible instances** in which a CLEAR OR LIKELY MATCH might be evident include:

    1. The item matches an SOL that is *altogether different from* the NWEA-tagged.

    2. The item *overlaps* a second SOL with the NWEA-tagged SOL.

c. If you select "CLEAR OR LIKELY MATCH" >>> Enter the SOL number

    i. Include a brief comment or notation to indicate your reasoning

d. If you select "UNLIKELY OR NO MATCH," then no other entries are needed in this section.

Making use of the Depth of Knowledge (DOK) framework and your subject-matter expertise, **determine the DOK of the item** >>> Select the DOK level from the dropdown menu.

a. Indicate comments or notations, if they would be helpful at a later time.

b. Open to **"Metadata"** in the NWEA Review Portal and note the NWEA-tagged DOK >>> Enter the NWEA-tagged DOK in the spreadsheet.

    i. **NOTE:** Your judgement of DOK does *not* need to match the NWEA tag.

    ii. If DOK levels are different, you might find it useful to include a comment or notation to reflect your reasoning.

    iii. If the levels are different this causes you to pause to re-think your determination, then you may change it. BUT, ***only change your judgement if your reasoning has changed.***

The final column is *optional*. Use it to add any additional comments or concerns. For example, if you have a concern about developmental appropriateness, lack of clarity, or implicit bias in the item, you might note that.

Thank you for contributing your time, energy, and expertise to this important study!

Appendix B:

Focus Group Questions

Phase 3: Accuracy and Utility of MAP Growth Reports

Focus Group Questions for SME Reviewers & Steering Committee Members

DIRECTIONS: (1) Make your own copy of this document to use in capturing your notes. (2) Be forthright and specific in your judgments. You need *not* write in narrative. You are welcome to share your thinking in bulleted statements, but please provide complete thoughts. (3) Save your document titled by your first and last name (e.g., "Jasmine Foley"), and upload it to the "Phase 3 SME Reviewers' Folder" in Google Drive.

1. How many years of experience do you have using MAP Growth assessments?

2. List the MAP Growth reports with which you have prior experience:

3. What report(s) do you/would you use **to determine whether a student is *on, below,* or *above* grade level**?

    a. Specifically, what information does (do) the report(s) provide you for making this determination?

    b. When, how, and with whom do you make use of the report(s)?

4. What report(s) do you/would you use **to determine specific *gaps in student learning***?

    a. Specifically, what information does (do) the report(s) provide you for making this determination?

    b. When, how, and with whom do you make use of the report(s)?

5. What report(s) do you/would you use **to make near-term instructional decisions?**

    a. Specifically, what information does (do) the report(s) provide you for making this determination?

    b. When, how, and with whom do you make use of the report(s)?

6. What report(s) do you/would you use **to project students' performance on end-of-year SOL tests?**

    a. Specifically, what information does (do) the report(s) provide you for making this determination?

    b. When, how, and with whom do you make use of the report(s)?

7. **What, if any, other assessments do you use** in place of or in addition to MAP Growth for any of the above purposes, and **what do these assessments provide to complement or confirm** information garnered from MAP Growth assessments?

8. Overall, which MAP Growth report(s) is (are) **most accurate and useful** to you?