Just link it: *The Importance of Using Linked Scores in a period when state summative assessments are not administered and beyond*

Introduction

The COVID-19 pandemic has brought about many changes in assessment including lack of summative assessment data throughout the United States. Most states had to cancel their summative spring assessments in both 2020 and 2021. In some instances, these assessments were optional. For many states, the main driver for the decision to administer the yearly summative state assessments was the mode of delivery of assessments – remote versus in school testing. States had to make tough considerations whether it was necessary to ask students who for the most part had been learning remotely, to come to school and take these assessments. For some people, that conflicted with the rationale for having students learn from home for the whole year. In addition, other schools of thought pointed out that remote administration of state summative assessment was going to threaten the validity of the scores. Third and most importantly, most states wanted to administer state summative assessments especially in spring 2021 to monitor students' progress and not for accountability. However, they wondered how they would control for any unintended use by other stakeholders if they used state assessments for accountability.

Cancellation of summative state assessments for both spring 2020 and 2021 meant that there was no data to gauge how the students were performing relative to state standards. These assessments are main driver of school improvement, evaluation and accountability for all states. Concerned with the need to monitor school improvement and as an alternative, many states were administering interim benchmark assessments to assess students' progress throughout the year. Unfortunately, the use of these assessments brought about new challenges through score interpretation and utility. Most of the interim assessments used have norm based inferences – students are ranked by comparing them to

other students, also called academic peers. However, problems of norm-referenced score interpretations are well documented in assessment literature – while norm-referenced scores provide the ranking that show how students perform in relation to other test-takers, they lack the ability to gauge if the performance is enough to meet standards that define masterly of specific skills at grade level. For example, knowledge of how many students a student has outperformed on a test does not translate to what skills a student has mastered

Because many states used interim assessment and not the familiar state summative assessments, educators and policy makers in schools had to find ways to interpret and use scores from interim benchmark assessments in a similar manner to state summative assessments. However, the ranks that are associated with normative inferences lacked the familiarity of standard based inferences that most summative assessments have. It is important to translate the ranks into a form of interpretation which most state assessment users are familiar with, which is standard based and aligned to state standards. To this end, the use of *linked scores* between interim benchmark assessments and state summative assessments became increasingly necessary in order to improve score interpretation and utility. Linking assessments creates a continuity in interpretation and utility of assessments even when summative assessments are not administered. Assessment users will therefore still be equipped with the same translation that they have used in the past and will be able to influence policy in more meaningful ways by placing students in programs that will put them back to track to achieving state standards. However, these results should only be used to get students the help they need and should never be used for accountability purposes. The utility of linking assessments is highlighted not only when summative assessments are not administered but also at different times of the school year – the high frequencies of administration of interim assessments allows for their use throughout the school year, in line with summative assessments. At the beginning or in the middle of the school year, policy makers can use the ability of interim assessments to predict what is going to happen at the end of the school year by using

linking studies, thereby allowing them to adjust policies way ahead of time, before summative assessments are administered.

Interim assessment results from Northwest Evaluation Association (NWEA) and Renaissance Learning have showed a clear message in 2020-2021 school year. Kuhfeld, Tarasawa, Johnson, Ruzek, and Lewis (2020) and Renaissance Learning (2020) found out that in the fall of 2020, student performance in mathematics was worse than it was before the pandemic. In reading however, performance was similar to pre- pandemic period. Store (2020) further investigated if the gaps in achievement widened for economically disadvantaged students and students with different ethnic backgrounds. Sullivan (2020) also used multi-level modeling to carry out a similar investigation. The two studies concluded that both economic and ethnic backgrounds were affecting student performance. Unfortunately, the ray of hope that we had in reading fall scores vanished by spring 2021 – student performance in both mathematics and reading declined (Lewis, Langi and Jensen, 2021). Therefore, the stakes to examine the gap in performance among different demographic groups of students have never been higher than what we are currently facing.

Betebenner and Wenning (2021) raise three issues about the results from interim assessments during the COVID-19 pandemic. First, the representativeness of the data used to examine learning loss was different, particularly with disadvantaged students. It is possible that the picture we have right now is a best case scenario on student learning loss assuming that those who missed out on testing would have been negatively impacted more than those who tested. Second, Betebenner, 2020 questions the amount of bias the results have due to the allowance for both at-home and in-class test administration when the norms themselves were derived from in-class test administration. Finally, and most importantly, Betebenner, 2020 highlights that interim benchmark assessments show, at best, modest coverage and alignment with state content standards. Betebenner wonders if learning loss would look

the same as learning loss derived from assessments having broad coverage and alignment with state standards.

It is important to highlight that the score linking that is mentioned here is not the same as test equating. The term score linking is used to describe the transformation from a score on one test to a score on another test (Dorans et al., 2010) or the process of relating scores across different forms of tests (Pommerich et al., 2004). Equating, a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004), is considered to be the strongest form of linking (Pommerich et al., 2004). Users of linked assessment results will need to be familiar with the statement that no amount of linking will adjust for differences in content. However, the accuracy of predicting the performance on state assessment based on interim benchmark test results is highlighted by the amount of convergent validity evidence collected through classification consistency rates and the level of correlation between scores from the two assessments.

NWEA has conducted a linking study based on data from the spring 2019 administration of the MAP Growth and Michigan's state assessments in grades 3-8. The equipercentile linking method (Kolen & Brennan, 2004) was used to identify the spring MAP Growth scores that correspond to the spring Michigan performance level cut scores. According to Mislevy (1992), this form of linking is called projection. Spring cuts for grade 2 were derived based on grade 3 and the 2020 NWEA growth norms (NWEA, 2020). Fall scores and winter cut scores that predict proficiency on the spring state assessment were then projected using the 2020 growth norms (Thum & Kuhfeld, 2020). With this linking, not all of the five requirements that are necessary for a linking to be an equating (Holland & Dorans, 2006) are met – equal construct requirement, equal reliability requirement, equal symmetry requirement, equity requirement, and the population invariance requirement. To be more specific, the Michigan state assessments and MAP Growth are designed for different purposes and measure slightly different constructs even within the same content area (NWEA, 2020). Although the high correlations between

performance on MAP Growth assessments and Michigan state assessments (0.77 – 0.92) and high classification rates (0.77 - 0.90) are good indicators of convergent validity between the two assessments, scores on the two tests cannot be assumed to be interchangeable. The high correlations and classification results are only a good pointer for educators and policy makers to use the results of the linking for classification purposes (likely to be proficient or not likely to be proficient) and target remediation by placing students in remedial programs that will put them back on track to achieving state standards.

Using NWEA's linking study between MAP Growth reading and mathematics tests with Michigan's summative state assessments, the fall achievement percentile on NWEA MAP Growth mathematics that project students to be proficient on Michigan's summative state assessments administered in the following spring in grades 3-8 range from 59^{th} to 70^{th} in mathematics and 57^{th} to 65^{th} in reading. These percentiles show how student scores compare to their peers nationwide. In addition, the percentiles highlight the high rigor of Michigan's performance level designations for its summative assessments – typical performance, which is a trademark of many norm based assessments, is not enough to make students proficient at the end of the school year. Whenever the state rigor of performance level designations is higher than typical performance on interim assessments, using interim assessments to monitor the performance of the typical student presents a rosier picture of student performance and fails to capture the full picture of the amount of ground students need to cover to get to the required state standards. The availability of linked scores will help solve the problem. Fortunately, NWEA has conducted linking studies between MAP Growth and many summative state assessments for many states that can be utilized. Other testing companies have conducted similar analyses.

We have seen gaps in administration of summative state assessments from one year to another. However, interim assessments are regularly administered two or three times a year. For example, in Michigan, state summative assessments were not administered in 2015, 2020 and were optional in

2021. The goal of this paper is threefold. First, the paper highlights the utility of linking interim assessment results to state assessment results to be used when state assessments are not administered so that there is no gap in interpretation of assessment results. Second, the paper highlights students' performance during the pandemic and the call for putting in place remedial programs that will help put students on track to meeting state standards. Finally, the study points out the need to use these interim assessments at the beginning of the school year in order to flag students early to provide the help they need.

Data and methodology

Data used in this study were collected from 49 schools in Michigan for 2017-18 to 2020-21 school years. For 2017-18 and 2018-19, the sample has students who were assessed in mathematics and reading on both Michigan Student Test of Educational Progress and fall sessions on NWEA MAP growth assessments. This data will be used to report classification consistency rates. In this case, the classification consistency rates are the number of times when NWEA Growth assessments correctly predicted the classification proficiency (proficient/not proficient) of students on the state summative assessments. The study uses 9558 students for 2017-18 and 10285 students for 2018-19 in mathematics. For reading, the study uses 9531 students for 2017-18 and 10247 students for 2018-19. For 2017-18 school year, grades 3-8 were used whereas for 2018-19 school year, the grade levels only included 3-7 because the state changed its policy and exempted 8th graders from the Michigan Student Test of Educational Progress. In addition, mathematics data for MAP Growth assessments from 13480 students in 2019-20 and 12713 students in 2020-21 were used to predict the classification proficiency on the state summative assessments – the school years when state assessments were not administered. Similarly, reading data for MAP Growth assessments from 13478 students in 2019-20 and 12787 students in 2020-21 was used to predict the classification proficiency on the state summative assessments. The goal was to show that even when summative assessments were not administered in

these two years, the data could be translated and utilized as if state summative assessments were administered.

In utilizing the linking study between the Michigan Student Test of Educational Progress and NWEA MAP growth assessments, this study seeks to explore the following questions:

1) What is the relationship in student performance between the Michigan Student Test of Educational Progress (M-STEP) and NWEA MAP growth assessments?
2) How has student performance changed between 2018-19 and 2020-21 school years?
3) How do learning losses, if any, compare with learning losses than have been reported from interim assessments results e.g., NWEA and Renaissance assessments?
4) How have Free/reduced lunch students and non-Free/reduced lunch students been affected by the COVID-19 pandemic?

By seeking answers to these questions using interim assessments data, the study seeks to demonstrate that interim assessments can fill the void that is left by summative assessments and where possible, complement each other e.g., when interim assessments are used at the beginning of the year to flag students that are in danger of not meeting state standards.

Results

Table 1 shows classification consistency rates for correctly predicting the proficiency level of a student on M-STEP using NWEA MAP Growth assessments. In mathematics, the classification consistency rates range from 84.3% to 91.1%. Stated differently, NWEA Growth MAP mathematics scores can consistently classify students' proficiency status on M_STEP mathematics test 84 -91% of the time. Similarly, NWEA Growth MAP reading scores can consistently classify students' proficiency status on M_STEP reading test 82-85% of the time. These consistency rates are high indicating stronger congruence between M-STEP

and NWEA MAP Growth cut scores. Both NWEA Growth MAP mathematics and reading tests are very good predictors of student's proficiency status on M-STEP.

Table 1

Classification Consistency Rates between MAP Growth and M-STEP Mathematics and Reading Tests

2017-2018 Mathematics				**2017-2018 Reading**			
Grade	Consistency Rate	False Positive	False Negative	Grade	Consistency Rate	False Positive	False Negative
3	86.7%	7.3%	6.0%	3	82.6%	9.3%	8.1%
4	84.3%	10.9%	4.8%	4	82.5%	11.2%	6.2%
5	90.8%	4.1%	5.1%	5	84.4%	11.4%	4.2%
6	89.8%	6.3%	4.0%	6	83.9%	9.0%	7.2%
7	90.5%	4.5%	5.0%	7	84.4%	9.8%	5.7%
8	91.1%	3.6%	5.3%	8	84.2%	10.0%	5.8%
2018-2019 Mathematics				**2018-2019 Reading**			
Grade	Consistency Rate	False Positive	False Negative	Grade	Consistency Rate	False Positive	False Negative
3	86.4%	5.5%	8.1%	3	84.1%	9.8%	6.0%
4	87.1%	8.9%	4.0%	4	84.2%	9.8%	6.0%
5	89.8%	4.5%	5.7%	5	85.2%	10.9%	3.9%
6	90.5%	4.2%	5.3%	6	84.8%	8.4%	6.9%
7	89.9%	3.2%	6.9%	7	84.8%	9.1%	6.1%
8	N/A	N/A	N/A	8	N/A	N/A	N/A

To investigate the relationship between M-STEP and NWEA MAP Growth assessments, more statistics are provided in table 2 below. The correlation coefficient between M-STEP and NWEA MAP Growth mathematics tests range from 0.79 to 0.88. These results are similar to those from NWEA sample from the linking study and indicate that there is a strong relationship between M-STEP and NWEA MAP Growth test scores. In addition, the percentage of students projected to be proficient on M-STEP, using NWEA Growth assessments in fall, aligns with percentages of students who actually were proficient on M-STEP in the following spring. As has already been discussed, consistency rates are almost 90% and for the most part, the projections seem to overestimate students' actual performance thereby ending up

with more false positives than negatives. The percentages of students that are projected to be proficient are about the same for 2017-18 and 2018-19 at grade level.

Table 2

Descriptive Statistics for 2017-18 and 2018-19 Mathematics Data

School Year	Grade	N	r	M-STEP			MAP Growth		
				Mean Score	SD	% Prof	% Projected Proficient	Mean Score	SD
2017-18	3	1826	0.82	1288.57	26.05	32.9	34.2	185.97	13.32
2017-18	4	1851	0.83	1387.54	25.01	31.3	37.4	197.55	13.91
2017-18	5	1861	0.83	1480.12	26.34	24.6	23.6	206.13	14.91
2017-18	6	1390	0.79	1580.58	25.73	20.7	23.0	209.99	14.74
2017-18	7	1343	0.79	1678.90	25.36	19.3	18.8	214.44	16.16
2017-18	8	1287	0.80	1776.53	24.00	16.7	15.0	220.16	17.09
2018-19	3	2169	0.84	1291.31	27.40	36.7	34.1	187.83	13.51
2017-19	4	2034	0.85	1387.64	25.09	30.5	35.3	198.52	13.72
2018-19	5	2083	0.86	1481.22	27.71	25.4	24.2	208.14	15.48
2018-19	6	2076	0.88	1581.98	26.07	25.9	24.8	211.68	15.44
2018-19	7	1923	0.88	1684.60	27.57	29.8	26.2	218.99	17.50

Table 3 shows a similar relationship between M-STEP and NWEA MAP Growth assessments in reading. The correlation coefficient between M-STEP and NWEA MAP Growth reading tests range from 0.73 to 0.82. Although the correlations coefficients are not as high as those for mathematics, they are still high and align with the correlation coefficients from NWEA sample from the linking study. The high level of correlation indicates that there is a strong relationship between M-STEP and NWEA MAP Growth reading test scores. In addition, the percentage of students projected to be proficient on M-STEP, using NWEA Growth assessments in fall, aligns with percentages of students who actually were proficient on M-STEP in the following spring. The high classification consistency rates that were discussed earlier (lower to mid-80s) and the projected and actual percentages of students who are proficient show that, as was the case with mathematics, the projections seem to overestimate students' actual performance – there are more false positives than false negatives. Also, the percentages of students that are projected to be proficient are about the same for 2017-18 and 2018-19 at grade level.

Table 3

Descriptive Statistics for 2017-18 and 2018-19 Reading Data

School Year	Grade	N	r	M-STEP			MAP Growth		
				Mean Score	SD	% Prof	% Projected Proficient	Mean Score	SD
2017-18	3	1821	0.77	1287.77	25.72	32.8	34.1	184.03	16.20
2017-18	4	1843	0.78	1390.90	26.78	36.0	41.0	195.38	16.32
2017-18	5	1858	0.79	1490.35	28.09	35.6	42.8	202.27	15.67
2017-18	6	1383	0.73	1586.57	27.33	31.5	33.3	207.12	15.96
2017-18	7	1341	0.77	1686.84	26.61	30.2	34.3	210.22	16.45
2017-18	8	1285	0.80	1785.70	26.45	30.6	34.7	215.15	15.50
2018-19	3	2163	0.80	1288.35	26.83	32.7	36.5	186.25	16.43
2017-19	4	2032	0.80	1390.12	26.43	36.1	40.0	196.46	15.79
2018-19	5	2079	0.82	1491..84	29.50	39.0	46.0	204.83	15.79
2018-19	6	2066	0.81	1589.31	26.58	35.1	36.6	209.97	14.64
2017-19	7	1907	0.81	1692.38	27.18	39.5	42.5	214.22	15.68

In order to show how students' performance can be interpreted when interim benchmark assessments are used in a year that state summative assessments are not administered e.g., as was the case with the COVID-19 pandemic or during a time when state assessments are not administered for other reasons, results of student performance on NWEA Growth mathematics tests are displayed in table 4. These results show that there was no differences in students' performance between 2018-19 and 2019-20 school years in mathematics. The small differences that are observed between these two school years appear to be random and very minimal. However, for the fall of 2020-21 school year, students' performance was lower for all grade levels compared to the two prior years by at least 6 percentage points. This shows that by the fall of the 2020-21 school year, students had already started feeling the impacts of the pandemic.

Table 4

Descriptive Statistics for 2018-19, 2019-20 and 2020-21 Mathematics Data

School Year	Grade	N	M-STEP	MAP Growth Performance		
			% Proficient	% Projected proficient	Mean Score	SD
2018-19	3	2169	36.7	34.1	187.83	13.51
2018-19	4	2034	30.5	35.3	198.52	13.72
2018-19	5	2083	25.4	24.2	208.14	15.48
2018-19	6	2076	25.9	24.8	211.68	15.44
2018-19	7	1923	29.8	26.2	218.99	17.50
2019-20	3	2455	N/A	32.5	186.80	13.56
2019-20	4	2372	N/A	37.1	197.79	14.43
2019-20	5	2228	N/A	24.8	206.60	15.66
2019-20	6	2242	N/A	24.5	210.85	15.73
2019-20	7	2144	N/A	22.5	216.90	17.08
2020-21	3	2232	N/A	28.0	184.24	14.67
2020-21	4	2295	N/A	27.5	194.11	14.79
2020-21	5	2145	N/A	19.3	203.49	15.81
2020-21	6	2076	N/A	19.3	208.31	15.34
2020-21	7	2000	N/A	20.1	215.59	16.88

While the differences in student performance are clear in mathematics between 2018-19 and 2019-20 versus 2020-21, in reading there was no differences in students' performance between 2018-19, 2019-20 and 2020-21 school years. The small differences that are observed between these three school years appear to be random and very minimal. These findings in both mathematics and reading are consistent with NWEA and Renaissance's fall reports on student performance during the pandemic.

Table 5

Descriptive Statistics for 2018-19, 2019-20 and 2020-21 Reading Data

School Year	Grade	N	M-STEP	MAP Growth Performance		
			% Proficient	% Projected proficient	Mean Score	SD
2018-19	3	2163	32.7	36.5	186.25	16.43
2018-19	4	2032	36.1	40.0	196.46	15.79
2018-19	5	2079	39.0	46.0	204.83	15.79
2018-19	6	2066	35.1	36.6	209.97	14.64
2018-19	7	1907	39.5	42.5	214.22	15.68
2019-20	3	2453	N/A	35.4	185.57	16.12
2019-20	4	2369	N/A	38.0	195.15	16.64
2019-20	5	2227	N/A	42.9	202.93	15.75
2019-20	6	2251	N/A	35.8	209.23	15.38
2019-20	7	2146	N/A	38.2	213.23	15.04
2020-21	3	2243	N/A	39.4	185.76	17.95
2020-21	4	2327	N/A	40.2	194.84	16.64
2020-21	5	2158	N/A	41.5	201.72	16.89
2020-21	6	2083	N/A	35.8	207.77	15.98
2020-21	7	2008	N/A	39.3	212.99	16.12

Finally, in line with the theme, *cultivating equitable education systems for the 21st century*, examination of the performance of students from different socio-economic backgrounds is necessary. Table 6 shows the percentage of students who were projected to be proficient in mathematics for the school years 2017-18, 2018-19, 2019-20 and 2020-21. From the results, it is clear that non-Free and reduced lunch students performed better than Free and reduced lunch students for all four schools years. However, the biggest gap in performance is observed between free and reduced lunch students for the 2020-21 school year (14.3% vs. 39.9%). This shows that there was a growing gap in student performance in mathematics during the COVID-19 pandemic.

Table 6

Percent Projected to be Proficient on M-STEP for Free and Reduced Lunch and Non-Free and Reduced Lunch students in Mathematics

School Year	FRL (N)	Non-FRL (N)	% Projected Prof. FRL	% Projected Prof. Non-FRL
2017-18	6181	3219	18.6	42.4
2018-19	6190	4068	18.9	44.4
2019-20	8480	4889	19.2	42.6
2020-21	8432	4088	14.3	39.9

Unfortunately, similar findings are observed with reading data. Table 6 shows that non-Free and reduced lunch students performed better than Free and reduced lunch students for all four schools years. However, just like mathematics data, the biggest gap in performance is observed between free and reduced lunch and non-free and reduced lunch students for the 2020-21 school year (30.2% vs. 60.6%). This shows that there is a growing gap in student performance in reading too during the COVID-19 pandemic despite earlier results that examined all students.

Table 7

Percent Projected to be Proficient on M-STEP for Free and Reduced Lunch and Non-Free and Reduced Lunch students in Reading

School Year	FRL (N)	Non-FRL (N)	% Projected Prof. FRL	% Projected Prof. Non-FRL
2017-18	6164	3210	28.7	54.7
2018-19	6164	4058	30.0	55.7
2019-20	8473	4885	29.8	54.3
2020-21	8501	4091	30.2	60.6

Conclusions

Interim benchmark assessments can provide similar information that test users get when they use state summative assessments. Although they may not be used for high stakes evaluation, the information that interim benchmark assessments provide may be very helpful especially at the beginning of the school year to put students on track to meet state standards early in the year before state summative

assessments are administered. Fortunately, there are many interim assessments that are linked to state assessments. The pause in the administration of state summative assessments in 2019-20 and 2020-21 due to the COVID-19 pandemic caused a growing need to use interim assessments. The two forms of assessments can coexist and work together in giving information that will help put students on successful paths to meeting state proficiency standards.

The data has shown that when the process of linking is done successfully, results on interim assessments can be very useful. The high classification consistency rates and high levels of correlation between scores on the two assessments indicated high levels of agreement between the two assessments discussed in this paper. Policy makers using the findings presented in this study need to understand the challenges that both false positives and false negatives present. If resource allocation is an issue, having more false negatives will prevent resources from getting where they are most needed.

Finally, as we hopefully move beyond the COVID-19 pandemic, it is important to focus our attention on different subgroups of students to ensure that all students are put on a recovery path. The data has shown that there are concerning gaps that are emerging between students of different socio-economic backgrounds that need to be addressed sooner rather than later.

References

Betebenner, D. W., & Wenning, R. J. (2021). Understanding Pandemic Learning Loss and Learning Recovery: The Role of Student Growth & Statewide Testing. *National Center for the Improvement of Educational Assessment*.

Dorans N.J., Moses, T.P., & Eignor, D.R. (2010). *Principles and Practices of Test Score Equating*. (ETS Research Report. RR-10-29). ETS, Princeton, New Jersey.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking. New York: Springer.

Kuhfeld, M, Tarasawa, B., Johnson, A., Ruzek, I, & Lewis K. (2020). *Learning during COVID-19: Initial findings on students' reading and math achievement and growth.* NWEA.

Lewis, K., Langi, M., Jensen, N. (2021). Learning During COVID-19: Summer 2021 Update. *Achievement and growth in the 2020-21 school year*. NWEA

Mislevy, R. J. (1992). *Linking educational assessments*: *Concepts, issues, methods, and prospects.* Princeton, NJ: ETS.

NWEA. (2020). Linking study report: predicting performance on the Michigan state assessment system in grades 3-8 based on NWEA MAP Growth scores.

Pommerich, M., Hanson, B., Harris, D., & Sconing, J. (2004). Issues in conducting linkage between distinct tests. Applied Psychological Measurement, 28(4), 247–273.

Renaissance Learning (2020). *How Kids are Performing: Tracking the impact of COVID-19 on Reading and Mathematics Achievement.* Special Report Series, Fall 2020 Edition (https://www.renaissance.com/how-kids-are-performing/)

Store, D. (2021). Analysis of Fall 2020 NWEA Assessment Data. The Governor John Engler Center for Charter Schools at Central Michigan University.

Sullivan, W. (2021). COVID-19 Impact on Education Outcomes. The Governor John Engler Center for Charter Schools at Central Michigan University

Thum, Y. M., & Kuhfeld, M. (2020). NWEA 2020 MAP Growth achievement status and growth norms for students and schools. NWEA Research Report. Portland, OR: NWEA.