Appendix B.3 M-STEP Student Data File Format

M-STEP Student Data File Format

The downloaded file containing student test scores is a Comma Delimited File (CSV) with the following fields in order:

Please note: fields containing "Reporting Level" information are referring to Claims for ELA/Math and Disciplines for Science/Social Studies.

Excel Column	Field	Descriptor	Field Type (length)	Format
A	TestCycleID	M-STEP test period and fiscal year	text(20)	
B	ISDCode	ISD code number	varchar(5)	99999
C	DistrictCode	District code number	varchar(5)	99999
D	SchoolCode	School code number	varchar(5)	99999
E	Grade	Student grade	varchar(2)	99
F	LastName	Student last name	varchar(25)	
G	FirstName	Student first name	varchar(25)	
H	MiddleInitial	Student middle initial	char(1)	9
I	Gender	Student's gender M = Male, F = Female	char(1)	9
J	Ethnicity	Student's ethnic code 0 = Native Hawaiian or Other Pacific Islander 1 = American Indian or Alaska Native 3 = Black or African American 4 = Hispanic or Latino 5 = White 6 = Two or more Races 9 = Asian	int(1)	9
K	UIC	Student UIC	char(10)	9999999999
L	StudentNumber	Student number from local school district	varchar(20)	
M	BirthDate	Student's date of birth	datetime(8)	mm/dd/yyyy
N	Barcode	Student's barcode number	varchar(10)	9999999999
O	ED	Economically disadvantaged Y = Yes, N = No	char(1)	9
P	EL	English learner Y = Yes, N = No	char(1)	9
Q	FEL	Former English learner Y = Yes, N = No	char(1)	9
R	FosterCare	Student is in foster care Y = Yes, N = No	char(1)	9
S	Homeless	Homeless student Y = Yes, N = No	char(1)	9

Excel Column	Field	Descriptor	Field Type (length)	Format
T	HomeSchooled	Homeschooled Y = Yes, N = No *Note: If a student is homeschooled in any subject then a student is considered homeschooled.*	char(1)	9
U	MS	Migrant status Y = Yes, N = No	char(1)	9
V	MilitaryConnected	MilitaryConnected Y = Yes, N = No	char(1)	9
W	SE	Special education Y = Yes, N = No	char(1)	9
X	ContentArea	Content Area MA = Math SC = Science *(not reported for 2018)* SS = Social studies EL = ELA	varchar(2)	99
Y	Online	Online Y = Yes, N = No	char(1)	9
Z	Valid	Valid Y = Yes, N = No	char(1)	9
AA	OutOfLevel	Student was tested out of level Y = Yes, N = No	char(1)	9
AB	Attemptedness	Attempted Y = Yes, N = No	char(1)	9
AC	ProhibitedBehavior	Prohibited behavior Y = Yes, N = No	char(1)	9
AD	Misadministration	Assessment misadministration Y = Yes, N = No	char(1)	9
AE	NonstandardAccommodation	Nonstandard accommodation used Y = Yes, N = No	char(1)	9
AF	InvalidReasonCode	1= Not found in MSDS 5 = Multiple Answer Documents Returned 6 = Late Return 7 = Misadministration 8 = Did not meet attemptedness	varchar(1)	99
AG	StandardAccommodation	Standard accommodation used Y = Yes, N = No	char(1)	9

Excel Column	Field	Descriptor	Field Type (length)	Format
AH	ReportingCode	Reporting code if provided by school	varchar(4)	9999
AI	ReportingCodeLabel	Reporting code label if provided by school	varchar(25)	
AJ	ResearchCode1	Research use code 1 Values 01-10 if provided by school	varchar(2)	99
AK	ResearchCode2	Research use code 2 Values 01-10 if provided by school	varchar(2)	99
AL	FormFixed	Fixed test form *Note: Says "CAT" for CAT tests*	varchar(4)	9999
AM	SS	Scale score	int(4)	9999
AN	SSSE	Scale score standard error	int(3)	999
AO	PL	Performance level 4 = Advanced proficient 3 = Proficient 2 = Partially proficient 1 = Not proficient	varchar(1)	9
AP	GrowthScore	Growth score *Note: NA possible*	varchar(2)	99
AQ	GrowthTarget	Growth target *Note: NA possible*	varchar(2)	99
AR	TargetTimeframe	Target timeframe (years) *Note: NA possible*	varchar(2)	99
AS	TotalPts	Raw-score total points *Note: NA for ELA and Math*	varchar(3)	999
AT	ReportingLevel1Code	Reporting level 1 code	varchar(10)	
AU	ReportingLevel1Description	Reporting level 1 description	varchar(100)	
AV	ReportingLevel1SS	Reporting level 1 scale score *Note: NA for social studies*	varchar(4)	9999
AW	ReportingLevel1SSSE	Reporting level 1 scale score standard error *Note: NA for social studies*	varchar(3)	999
AX	ReportingLevel1PerfIndicator	Reporting level 1 performance indicator: 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99

Excel Column	Field	Descriptor	Field Type (length)	Format
AY	ReportingLevel1Pts	Reporting level 1 raw score points *Note: NA for ELA and Math*	varchar(2)	99
AZ	ReportingLevel1PtsPossible	Reporting level 1 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
BA	ReportingLevel1PctCorrect	Reporting level 1 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
BB	ReportingLevel2Code	Reporting level 2 code	varchar(10)	
BC	ReportingLevel2Description	Reporting level 2 description	varchar(100)	
BD	ReportingLevel2SS	Reporting level 2 scale score *Note: NA for social studies*	varchar(4)	9999
BE	ReportingLevel2SSSE	Reporting level 2 scale score standard error *Note: NA for social studies*	varchar(3)	999
BF	ReportingLevel2PerfIndicator	Reporting level 2 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	
BG	ReportingLevel2Pts	Reporting level 2 raw score points *Note: NA for ELA and Math*	varchar(2)	99
BH	ReportingLevel2PtsPossible	Reporting level 2 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
BI	ReportingLevel2PctCorrect	Reporting level 2 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
BJ	ReportingLevel3Code	Reporting level 3 code	varchar(10)	
BK	ReportingLevel3Description	Reporting level 3 description	varchar(100)	
BL	ReportingLevel3SS	Reporting level 3 scale score *Note: NA for social studies*	varchar(4)	9999
BM	ReportingLevel3SSSE	Reporting level 3 scale score standard error *Note: NA for social studies*	varchar(3)	999

Excel Column	Field	Descriptor	Field Type (length)	Format
BN	ReportingLevel3PerfIndicator	Reporting level 3 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99
BO	ReportingLevel3Pts	Reporting level 3 raw score points *Note: NA for ELA and Math*	varchar(2)	99
BP	ReportingLevel3PtsPossible	Reporting level 3 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
BQ	ReportingLevel3PctCorrect	Reporting level 3 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
BR	ReportingLevel4Code	Reporting level 4 code	varchar(10)	
BS	ReportingLevel4Description	Reporting level 4 description	varchar(100)	
BT	ReportingLevel4SS	Reporting level 4 scale score *Note: NA for social studies*	varchar(4)	9999
BU	ReportingLevel4SSSE	Reporting level 4 scale score standard error *Note: NA for social studies*	varchar(3)	999
BV	ReportingLevel4PerfIndicator	Reporting level 4 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99
BW	ReportingLevel4Pts	Reporting level 4 raw score points *Note: NA for ELA and Math*	varchar(2)	99
BX	ReportingLevel4PtsPossible	Reporting level 4 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
BY	ReportingLevel4PctCorrect	Reporting level 4 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
BZ	ReportingLevel5Code	Reporting level 5 code	varchar(10)	
CA	ReportingLevel5Description	Reporting level 5 description	varchar(100)	

Excel Column	Field	Descriptor	Field Type (length)	Format
CB	ReportingLevel5SS	Reporting level 5 scale score *Note: NA for social studies*	varchar(4)	9999
CC	ReportingLevel5SSSE	Reporting level 5 scale score standard error *Note: NA for social studies*	varchar(3)	99
CD	ReportingLevel5PerfIndicator	Reporting level 5 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99
CE	ReportingLevel5Pts	Reporting level 5 raw score points *Note: NA for ELA and Math*	varchar(2)	99
CF	ReportingLevel5PtsPossible	Reporting level 5 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
CG	ReportingLevel5PctCorrect	Reporting level 5 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
CH	ReportingLevel6Code	Reporting level 6 code	varchar(10)	
CI	ReportingLevel6Description	Reporting level 6 description	varchar(100)	
CJ	ReportingLevel6SS	Reporting level 6 scale score *Note: NA for social studies*	varchar(4)	9999
CK	ReportingLevel6SSSE	Reporting level 6 scale score standard error *Note: NA for social studies*	varchar(3)	99
CL	ReportingLevel6PerfIndicator	Reporting level 6 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99
CM	ReportingLevel6Pts	Reporting level 6 raw score points *Note: NA for ELA and Math*	varchar(2)	99
CN	ReportingLevel6PtsPossible	Reporting level 6 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99

Excel Column	Field	Descriptor	Field Type (length)	Format
CO	ReportingLevel6PctCorrect	Reporting level 6 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
CP	ReportingLevel7Code	Reporting level 7 code	varchar(10)	
CQ	ReportingLevel7Description	Reporting level 7 description	varchar(100)	
CR	ReportingLevel7SS	Reporting level 7 scale score *Note: NA for social studies*	varchar(4)	9999
CS	ReportingLevel7SSSE	Reporting level 7 scale score standard error *Note: NA for social studies*	varchar(3)	999
CT	ReportingLevel7PerfIndicator	Reporting level 7 performance indicator 3 = Adequate progress 2 = Attention may be indicated 1 = Most at risk of falling behind *Note: NA for social studies*	varchar(2)	99
CU	ReportingLevel7Pts	Reporting level 7 raw score points *Note: NA for ELA and Math*	varchar(2)	99
CV	ReportingLevel7PtsPossible	Reporting level 7 raw score points possible *Note: NA for ELA and Math*	varchar(2)	99
CW	ReportingLevel7PctCorrect	Reporting level 7 percent correct *Note: NA for ELA and Math*	varchar(5)	999.9
CX	EssayPtsEarned	Essay raw score points Condition Codes: B = Blank I = Insufficient L = Nonscorable Language T = Off Topic M = Off Purpose	varchar(2)	99
CY	EssayPtsPossible	Essay points possible	int(2)	99
CZ	CreatedDate	Student data file creation date	datetime(16)	mm/dd/yyyy hh:mm

Modification Log:

1. Initial version created 7/25/16
2. Formatting updated 8/1/16
3. Changed column R, LEP (Limited English proficient), to EL (English learner) 6/19/17

4. Field Type (length) updated to varchar(100) for the following Reporting Level Description fields: AS, BA, BI, BQ, BY, CG, CO; 6/19/17
5. Removed “Spring 2016” from document title 6/19/17
6. Changed document title to “Spring 2018 M-STEP Student Data File Format” 8/2/18
7. Removed the following fields 8/2/18
    a. FeederSchoolCode
    b. Unused
    c. FormPTask
    d. PTPts
    e. EssayTypeCode
    f. EssayTypeDescription
    g. EssayRubric1Code
    h. EssayRubric1Name
    i. EssayRubric1Pts
    j. EssayRubric1PtsPoss
    k. EssayRubric2Code
    l. EssayRubric2Name
    m. EssayRubric2Pts
    n. EssayRubric2PtsPoss
    o. EssayRubric3Code
    p. EssayRubric3Name
    q. EssayRubric3Pts
    r. EssayRubric3PtsPoss
8. Added the following fields 8/2/18
    a. FEL (Q)
    b. FosterCare (R)
    c. MilitaryConnected (V)
    d. ReportingCodeLabel (AI)
9. Demographics fields are now in alphabetical order 8/2/18
10. Field title of EssayTotalPts changed to EssayPtsEarned 8/2/18
11. Field title of EssayTotalPtsPoss changed to EssayPtsPossible 8/2/18
12. The Descriptor for the following fields have changed 8/2/18
    a. EL
    b. FEL
    c. Homeless
    d. HomeSchooled
    e. MS
    f. SE
    g. Valid
    h. OutOfLevel
    i. Attemptedness
    j. ProhibitedBehavior
    k. Misadministration
    l. NonstandardAccommodation

m. StandardAccommodation
n. EssayPtsEarned

13. Changed column AP, SGP (Student Growth percentile), to GrowthScore 7/11/19
14. Added following fields 7/11/19
    a. GrowthTarget (AQ)
    b. TargetTimeframe (AR)

Appendix B.4 M-STEP Aggregate Data File Format

M-STEP Aggregate Data File Format

The downloaded file containing M-STEP aggregate data is a Comma Delimited File (CSV) with the following fields in order:

Excel Column	Field	Descriptor	Field Type and Length
A	TestCycle	Test name and year	text(20)
B	ISDCode	ISD code	varchar(05)
C	ISDName	ISD name	varchar(50)
D	DistrictCode	District code	varchar(05)
E	DistrictName	District name	varchar(50)
F	SchoolCode	School code	varchar(05)
G	SchoolName	School name	varchar(50)
H	Grade	Tested grade	varchar(02)
I	Subject	English Language Arts Mathematics Science *(NA for 2018)* Social Studies	varchar(20)
J	SubGroupType	All Students Economically disadvantaged (ED) English learner (EL) Ethnicity Former English learner (FEL) Foster Care Gender Homeless Migrant (MS) Military Connected Standard Standard - EL	varchar(20)
K	DemographicSubGroup	All Students Female Male Students With Disabilities American Indian or Alaska Native Asian Black or African American Hispanic or Latino Native Hawaiian or Other Pacific Islander Two or More Races White No *(not used for All Students, Ethnicity, Gender)* Yes *(not used for All Students, Ethnicity, Gender)*	char(37)
L	AvgSS	Average scale score of selected group	integer
M	StdDev	Standard deviation of selected group	integer

Excel Column	Field	Descriptor	Field Type and Length
N	NotProficientN	Number of students not proficient in selected group	integer
O	NotProficientPct	Percent of students not proficient in selected group	decimal(8,1)
P	PartiallyProficientN	Number of students partially proficient in selected group	integer
Q	PartiallyProficientPct	Percent of students partially proficient in selected group	decimal(8,1)
R	ProficientN	Number of students proficient in selected group	integer
S	ProficientPct	Percent of students proficient in selected group	decimal(8,1)
T	AdvancedN	Number of students advanced in selected group	integer
U	AdvancedPct	Percent of students advanced in selected group	decimal(8,1)
V	MetStandardsN	Number of students who met standards in selected group	integer
W	MetStandardsPct	Percent of students who met standards in selected group	decimal(8,1)
X	DidNotMeetStandardsN	Number of students who did not meet standards in selected group	integer
Y	DisNotMeetStandardsPct	Percent of students who did not meet standards in selected group	decimal(8,1)
Z	NumberTestedN	Number of students tested in selected group	integer
AA	NumberIncludedN	Number of students included in selected group	integer

Modification Log:

1. Initial version created in December, 2015
2. Formatting revisions made 8/10/16
3. SubGroupType of Foster Care added 8/2018
4. SubGroupType of Military Connected added 8/2018

Appendix C: Target Score Report

Psychometric Analysis Report for the
Michigan 3-8 English Language Arts (ELA) and Mathematics Assessment Target Reporting

June 2019

DRC Psychometric Services Michigan Project Team
Data Recognition Corporation

Table of Contents

List of Tables

Introduction

The assessment target score report is designed to report a group of students' (e.g., at the grade, school, teacher, and/or district levels) relative strength and weakness at the assessment target level. It is for aggregate level reports only.

Unlike the performance categories provided at the total test and claim levels, these strengths and weakness do not imply proficiency. Instead, they show how a group of students' performance is distributed across the content target relatively to their overall performance. For example, a group of students may have performed very well on a subject, but performed lower on a target. Thus, performance level code of C not necessarily imply a lack of proficiency, but that these students' performance on that target was lower than their performance across other targets put together. It can be concluded that the students performed lower than expected on that target.

Assessment target score report should serve as a starting point in an overall investigation of students' strengths and weaknesses and constitutes only one of many sources of evidence that should be used in evaluating student performance.

This was conducted for the English Language Arts (ELA) and Mathematics M-STEP assessments.

Methodology

Item response theory (IRT) based residual analysis can be used to conduct analyses for the assessment target score report. The residual is the difference between the observed score and expected score at the item level. The observed score is the score (e.g., 0 to 3) a student submitted for each item. The expected score is derived using the 2 parameter logistic (2PL) model for dichotomously scored items and generalized partial credit model (GPCM) for polytomously scored items.

The expected score for a multiple-choice item (MC, one point item) was computed using the two-parameter logistic (2PL) model as shown below in equation 1.

$$P_i\,(x_i = 1|\theta, a_i, b_i) = \frac{\exp Da_i(\theta - b_i)}{1+\exp Da_i(\theta - b_i)}, \qquad (1)$$

where a_i is item discrimination parameter and b_i is item difficulty for item *i*, and p_i is the probability of the item getting correct given the observed overall ability estimate, θ, and *D* is 1.7. The expected score

for a constructed response (CR) item, the observed overall ability estimate, θ, was computed with generalized partial credit mode (equation 2).

$$P_{ik}\ (x_i|\theta, a_i, b_{i0}, b_{i1}, \dots, b_{iK_i}) = P_{ik}\ (\theta) = \frac{\exp Da_i \sum_{k=0}^{x_i}(\theta - b_{ik})}{\sum_{r=0}^{K_i}\left[\exp Da_i \sum_{k=0}^{r}(\theta - b_{ik})\right]}\ , \qquad (2)$$

where $\sum_{k=0}^{0}(\theta - b_{ik}) \equiv 0$.

Equation (2) computes the probability of obtaining the score of $0 \leq x_i \leq K_i$ on CR item i. The item discrimination parameter is a_i, and b_{ik} is the category intersection parameter (in SBAC scoring specification, it is referred to step parameters). Equation (1) is a special case of equation (2) with $K_i = 1$. This means that the computation of probability can be completed for both 2PL MC and CR items using equation (2).

For all items, the residual, R_i, is found by using equation (3),

$$R_i = O_i - P_i\ (\theta) \qquad (3)$$

where O_i is the observed score for item *i* and $P_i\ (\theta)$ is the expected score for item *i.*

Once the individual residuals were calculated, the weighted average of the residuals were calculated for each assessment target meeting the reporting criteria (see the Reporting Criteria Section below for more details) criteria using equation (4).

$$\bar{R}_{target} = \frac{1}{\sum_{i=1} w_i}\sum_{i=1} R_i \qquad (4)$$

where R_i is the residual for item *i* and w_i is the weight associated with item *i* that accounts for the number of score points for that item.

Reporting Criteria

Target assessment results were reported for both ELA and Mathematics. Table 1 provides the claim and target level for which target assessment results were reported. Once the average residual for each assessment target was computed, a flagging criterion of +/- 0.05 was used to indicate the assessment target level performance. Table 2 provides a description of the performance levels.

Table 1: Assessment targets are listed below by content area.

ELA		Math	
Claim	**Assessment Target**	**Claim**	**Assessment Target**
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	1	A, B, C, D, E, F, G, H, I, J, K, L
2	1, 2, 3, 4, 5, 6, 7, 8, 9	2	A, B, C, D
3	4	3	A, B, C, D, E, F
4	2, 3, 4	4	A, B, C, D, E, F

Table 2: Performance level descriptions.

PL Code	Target Level	Description
A	Better than performance on the test as a whole	This target is a relative strength. The group of students performed better on items from this target than they did on the rest of the test, as a whole.
B	Similar to performance on the test as a whole	This target is neither a relative strength nor a relative weakness. The group of students performed about as well on items from this target as they did on the rest of the test, as a whole.
C	Worse than performance on the test as a whole	This target is a relative weakness. The group of students did not perform as well on items from this target as they did on the rest of the test, as a whole.
--	Insufficient Information	Not enough information is available to determine whether this target is a relative strength or weakness.

Additionally, since the M-Step administration was a CAT, the number of items presented in each assessment target varied for each administration. Thus, reporting criteria was used to ensure that a specified number of unique items were presented in order the assessment target results to be provided. The criteria used is listed below:

- Number of unique students per target: n=15
- Number of unique items per target: n=3
- Number of responses per target: n=25
- Use 0.05 criterion on the rescaled residual scale

Exclusions

It should be noted that some students were excluded from the target reporting analysis. Students who were Force Submit or scored at the lowest and highest obtainable scale score (LOSS and HOSS) were excluded from the analysis. Additionally, students with invalid tests and home schooled students were excluded.

Results

Aggregate results were provided to MDE by State, ISD, District, and Building. Private school students were only included in the building level aggregate results. Table 3 shows N count used at the state level and Tables 4 and 5 provide the state level results for ELA and Mathematics. Note that the PL codes in Tables 3 and 4 correspond to those found in Table 2.

Table 3: Valid student counts at the state level by content area and grade.

Content Area	Grade	N
ELA	**3**	101,611
	4	103,932
	5	108,186
	6	107,819
	7	106,545
	8	109,525
Math	**3**	101,580
	4	104,171
	5	108,237
	6	107,629
	7	106,699
	8	108,329

Table 4: State level aggregate results for ELA.

Claim	Target	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
		Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code
1	**1**	72,440	B	100,577	B	78,166	B	92,623	B	105,700	B	94,777	B
	2	99,332	B	90,766	B	86,900	B	84,927	C	52,489	B	58,382	B
	3	91,170	B	87,448	B	92,632	B	100,241	B	88,280	B	84,525	B
	4	81,303	B	73,114	B	97,029	B	81,141	B	81,954	D	77,553	B
	5	50,550	B	23,344	B	20,171	B	103,318	C	67,099	D	65,505	C
	6	24,174	B	68,407	B	88,476	B	83,833	B	70,526	B	92,374	B
	7	89,932	B	87,879	B	73,857	B	77,205	B	75,802	B	99,151	B
	8	85,588	B	95,987	B	96,149	B	95,913	B	94,665	B	108,488	B
	9	84,711	B	96,492	B	63,289	B	82,502	B	104,823	B	103,187	B
	10	88,257	B	97,266	A	95,529	B	82,441	B	83,427	A	90,208	B
	11	73,457	B	90,941	B	92,020	B	107,703	B	98,424	B	84,060	B
	12	70,689	B	49,542	B	69,226	B	13,020	B	64,351	A	28,154	B
	13	82,670	B	56,231	B	82,560	B	61,042	B	66,454	B	38,487	B
	14	72,369	B	69,309	B	95,599	B	49,928	B	62,483	A	26,391	A
2	**1**	99,126	B	90,606	B	104,329	B	105,256	B	73,185	B	98,635	B
	2	91,210	C	97,301	B	99,288	B	93,575	C				
	3	101,611	A	103,932	B	108,186	B	107,819	B	102,791	C	85,496	B
	4	74,037	B	75,471	B	81,901	C	81,955	B	106,545	B	109,525	C
	6	99,126	B	100,577	B	78,166	B	107,819	B	101,773	C	106,544	B
	8	101,611	B	103,932	B	108,186	B	107,819	B	106,545	B	109,525	B
	9	101,611	B	103,932	B	108,186	B	107,819	B	106,545	B	109,525	B
3	**4**	101,611	B	103,932	B	108,186	B	102,850	B	106,545	B	109,525	B
4	**2**	101,609	B	103,928	B	107,324	B	107,599	B	96,516	A	102,902	B
	3	76,865	B	78,258	B	105,603	B	98,530	B	105,514	B	101,878	B
	4	85,848	C	99,764	B	103,643	B	92,623	B	105,054	B	74,084	B

Table 5: State level aggregate results for Mathematics.

Claim	Target	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
		Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code	Valid N	PL Code
1	A	101,580	B	80,042	B	25,419	B	107,629	B	106,330	B	84,985	B
	B	35,624	B	59,209	C	1,270	A	80,583	B	104,889	B	59,213	B
	C	87,179	B	22,444	B	104,525	B	95,500	B	105,522	B	106,687	B
	D	91,865	B	104,171	B	80,650	B	107,629	A	106,217	B	103,755	B
	E	99,657	B	57,118	B	107,877	B	74,738	B	97,812	B	108,078	B
	F	86,428	B	98,572	B	108,237	B	107,474	A	80,822	B	85,801	B
	G	85,563	B	104,171	A	74,296	B	102,067	B	66,906	B	70,058	B
	H	101,580	B	104,171	B	68,534	B	84,517	B	37,165	B	67,439	B
	I	40,416	A	102,313	B	86,303	B	25,516	B	71,332	B	48,746	B
	J	32,129	B	22,518	B	106,809	B	52,411	B			84,986	B
	K	26,459	A	51,002	A	37,392	B						
	L			104,162	B								
2	A	101,580	B	104,171	B	108,237	B	107,629	B	106,699	B	108,329	B
	B	79,098	B	16,944	B	43,006	B	46,925	B	53,536	B	40,117	B
	C	53,871	B	97,868	B	51,396	B	57,138	B	62,980	B	81,610	B
	D	18,616	B	23,750	B	81,909	B	72,059	B	64,152	B	59,175	B
3	A	93,407	B	99,308	B	97,790	B	106,368	B	101,335	B	87,806	B
	B	94,659	B	89,651	B	62,131	B	47,333	B	22,686	D	37,355	B
	C	65,198	B	73,329	B	78,718	B	75,279	B	73,411	B	55,437	B
	D	88,367	B	84,778	B	94,780	B	68,115	B	88,822	B	104,274	B
	E	87,098	B	96,868	B	104,609	B	107,256	B	106,699	B	108,313	B
	F	82,304	B	85,654	B	83,955	B	77,591	B	62,622	B	39,267	B
	G							21,529	B	45,166	B	85,050	B
4	A	63,514	B	94,418	B	83,793	B	100,491	B	85,201	B	63,053	B
	B	1,295	B	22,131	B	47,161	C	14,987	B	13,439	B	32,060	B
	C	58,571	B	62,005	B	54,181	B	30,272	B	49,120	B	46,269	B
	D	89,071	B	56,549	B	82,625	B	52,554	B	76,988	B	94,709	B
	E	100,285	B	82,040	B	61,076	B	92,646	B	94,028	B	76,269	B
	F	42,788	B	42,175	B	54,056	B	77,357	B	57,579	B	62,060	B

Considerations and Cautions

Unlike the performance levels provided at the total test and claim levels, these strengths and weakness do not imply proficiency. Instead, they show how a group of students' performance is distributed across the content target relatively to their overall performance. For example, a group of students may have performed very well on a subject, but performed lower on a target. Thus, a target performance code of C a target does not necessarily imply a lack of proficiency, but that these students' performance on that target was lower than their performance across other targets put together. In other words, the students performed lower than expected on that target. Although the students are doing well, the educators may still want to focus instruction on the targets with performance code C.

Assessment target score report should serve as a starting point in an overall investigation of students' strengths and weaknesses and constitutes only one of many sources of evidence that should be used in evaluating student performance.

Appendix D: M-STEP SGP and AGP Report

Psychometric Analysis Report for the
Michigan English Language Arts (ELA), Mathematics, Science, and SAT
Student Growth Percentile and Adequate Growth Percentile Reporting

March 2019

DRC Psychometric Services Michigan Project Team
Data Recognition Corporation

Table of Contents

List of Tables

List of Figures

Introduction

The use of student growth models is common in K-12 testing. The most commonly used approaches by states are conditional growth percentile models, which include student growth percentiles (SGPs, Betebenner, 2008; 2009; 2011) or an alternative known as percentile rank residuals (Castellano & Ho, 2013). Both models attempt to describe individual student growth relative to other students who are academically similar by using prior test scores as predictors. Adequate growth percentiles (AGPs, Betebenner, 2008; 2009; 2011) which use quantile regression models, provide the likelihood students are on track to reaching or maintaining proficiency at some time point in the future. Individual level results from these models can be aggregated at a group level.

SGP analyses were conducted for the M-STEP, SAT, and WIDA, and PRR analysis was conducted for MI-Access assessments. AGP analyses were conducted for M-STEP.

Methodology

Student Growth Percentiles (SGP)

For assessments with a sufficient sample size (M-STEP, SAT, and WIDA Access) student growth percentiles (SGPs) were calculated using the R SGP package (Betebenner et. al., 2015) version 1.8-3.17 as compiled from the master branch of the SGP GitHub repository. SGPs defined this way take a normative approach.

Specially, let Y_t denote an assessment score at time $\boldsymbol{t}$, the expected value of Y_t at the $\boldsymbol{\tau}$-th quantile, $Q_{Y_t}(\tau|Y_{t-1},\ldots,Y_1)$ based on prior assessment scores $Y_{t-1},\ldots,Y_1$, is then given by (Betebenner, 2011, p17)

$$Q_{Y_t}(\tau|Y_{t-1},\ldots,Y_1)=\sum_{j=1}^{t-1}\sum_{i=1}^{3}\phi_{ij}(Y_j)\beta_{ij}(\tau) \qquad (1)$$

Where ϕ_{ij}, $i=1,2,3$ and $j=1,\ldots,t-1$ denote the B-spline basis functions for quantile τ. For instance, for τ=.5, Q_{Y_t} returns the estimated median expectation of Y_t for any combination of $Y_{t-1},\ldots,Y_1$. This analysis used the default parameters of the SGP package which generates 1+7*(number of pretest) parameters per quantile. For example, for a 3-pretest model we have 1+7*3 = 22 parameters per quantile and we estimate 100 quantiles independently (from 0.005 to 0.995 in 0.01 increments).

Calculating a SGP from equation 1 requires prior test score information to determine predicted scores. The SGP for a student is defined as the midpoint of the (ranked) two quantiles between which the student's score falls.

$$SGP_i=\left(\max\left\{\tau_i:\hat{Q}_\tau(Y|X=x_i)<y_i\right\}+\min\left(\left\{\tau_i:\hat{Q}_\tau(Y|X=x_i)>y_i\right\}\right)\right)*\frac{100}{2} \qquad (2)$$

Where x_i is the student *i's* vector of prior test scores.

Adequate Growth Percentiles (AGP)

Using the same methodology as described above for calculating SGPs, to calculate a projection or the trajectory a student needs to meet a certain target. An adequate growth percentile, AGP, is the SGP that a student needs to have to meet or exceed the proficient cut score (or any pre-determined achievement target) within a specified time frame (number of academic years).

Betebenner (2011) contextualizes AGPs in terms of "catch-up", "keep-up", or "move-up." Suppose that an AGP is calculated for a given students Y years away. The following would apply:

> **Catch-Up** is used for students currently not proficient who are expected to reach proficient within *Y* years or by the time they have finished their education, whichever comes first
>
> **Keep-Up** is used for students currently at or above proficient who are expected to remain at or above proficient for all *Y* years or by the time they have finished their education, whichever comes first.
>
> **Move-Up** is used for students currently proficient who are expected to advance beyond proficient within *Y* years or by the time they have finished their education, whichever comes first.

Additionally, a lagged AGP target is also calculated and this value is similar to the AGP. But in this case the current year AGP (i.e. 2018) using the quantile regression model. This gives information to determine if students are on track to reaching proficiency or if they will maintain proficiency over a specified number of years.

Percentile Rank Residuals (PRR)

For assessments with small sample sizes (MI-Access), the PRR method (Castellano & Ho, 2013) was used to estimate the conditional student growth percentiles. This method uses an ordinary least squares (OLS) model, where the predictors consist of past student achievement data.

$$Y_{it} = \beta_0 + \beta_1 y_{i(t-1)} + \beta_2 y_{i(t-2)} + \varepsilon_{it} \quad (5)$$

where Y_{it} is the observed score on the assessment at time *t* for student *i*, $Y_{i,\,t-1}$ is the observed score at prior time 1 and $Y_{i,\,t-2}$ is the observed score at prior time 2. The βs are the regression coefficients, and ε_{it} is a residual error.

After estimating Equation 5, the residuals are calculated using Equation 6:

$$\hat{\varepsilon}_{it} = y_{it} - \hat{y}_{it} \quad (6)$$

where $\hat{\varepsilon}_{it}$ is the residual for student *i* at time *t*, $\hat{y}_{it}$ the predicted score from equation 5.

Next, the residuals are rank ordered (Castellano & Ho, 2013, p. 195).

$$PRR_{it} = F(\hat{\varepsilon}_{it}) \times 100 = \frac{\#residuals \leq \hat{\varepsilon}_{it}}{n} \times 100 \quad (7)$$

where $\hat{\varepsilon}_{it}$ is the residual for student *i* at time *t* and *n* is the total sample size for all students with MI-Access FI results for a given posttest in 2017-18.

A standard error of measurement can be obtained by simulation for this method. Specifically, for a given posttest, y_{it}, and $CSEM(y_{it})$ 100 posttest were simulated such that they follow a normal distribution given by Equation 8:

$$y_{its} \sim N(mean = y_{it}, sd = CSEM(y_{it})) \quad (8)$$

For each simulated y_{its}, calculate the corresponding PRR using equations 5-7 while holding all other student data constant. Repeat this for each student.

Reporting Results

Results were reported at both the student and aggregate levels. This section provides a brief overview of the results provided to MDE.

For each assessment, results were reported for different content areas. Table 1 provides a list of the assessment and content areas combinations for which SGPs or PRRs were provided. Table 1 provides a list of the grades and domains for which results were reported. Content areas for which AGPs are calculated are also noted in Table 1.

Table 1: Applicable assessments by grade

Grade	M-STEP	SAT	MI-Access	WIDA
K				Overall Composite
1				Overall Composite
2				Overall Composite
3	ELA, Math		ELA, Math	Overall Composite
4	ELA, Math		ELA, Math, Science	Overall Composite
5	ELA, Math, Social Studies		ELA, Math, Social Studies	Overall Composite
6	ELA, Math		ELA, Math	Overall Composite
7	ELA, Math		ELA, Math, Science	Overall Composite
8	ELA, Math, Social Studies		ELA, Math Social Studies	Overall Composite
11	Social Studies	ELA, Math	ELA, Math, Social Studies	Overall Composite
12				Overall Composite

AGP Projections

For ELA and Math grades 4 through 8, AGP targets and/or lagged targets were computed for 1 to 4 years from 2018 or 8th grade, whichever comes first. For example, a grade 4 student had AGPs to grades 5, 6, 7, and 8. While a grade 7 student had an AGP to 8th grade. Lagged AGP targets are calculated for Grades 4 through 8. Tables 2 and 3 show the grade progressions for AGP and AGP lagged targets.

Table 2: M-STEP Math and ELA AGP targets by grade, projection year, and grade projected to

	Projected AGP Target Year			
Grade 2018	**1 Year 2019**	**2 Year 2020**	**3 Year 2021**	**4 Year 2022**
4	5th grade	6th grade	7th grade	8th grade
5	6th grade	7th grade	8th grade	
6	7th grade	8th grade		
7	8th grade			
8				

Table 3: M-STEP Math and ELA AGP lagged targets by grade and projection year

	Projected AGP Lagged Target Year			
Grade 2017	**Current Year 2018**	**Current +1 Year 2019**	**Current +2 Year 2020**	**Current +3 Year 2021**
3	4th grade	5th grade	6th grade	7th grade
4	5th grade	6th grade	7th grade	8th grade
5	6th grade	7th grade	8th grade	
6	7th grade	8th grade		
7	8th grade			

Categorization of Individual (Level) Growth Percentiles

Individual (level) growth percentiles (either SGP or PRR) will also be assigned one of three categorical descriptors based on MDE reporting policies, which are defined as:

- Low: SGP 1-29
- Medium: SGP 30-69
- High: SGP 70-99

Additionally, individual (level) growth percentiles (either SGP or PRR) will also be assigned one of five categorical descriptors based on historical MDE accountability policies. These five categorical descriptors are no longer used in MDE accountability processes but were still calculated for analysis purposes. The five categorical descriptors are defined as:

- Significant Decline (SGP 0-19)
- Decline (SGP 20-39)
- Maintain (SGP 40-59)
- Improvement (SGP 60-79)
- Significant Improvement (SGP 80-99)

Valid Test Sequence Rules

Identified suitable pathways and their information can be found in Table 4 for the SGP method (M-STEP/SAT), the PRR approach (MI-Access FI), and the SGP method (WIDA Access).

Table 4: M-STEP Testing Program Valid Sequence for SGP/AGP calculations

Program	Grade 2018	Prior Year 1	Prior Year 2
M-STEP ELA & Math	4	M-STEP 3rd grade Spring 2017	
	5	M-STEP 4th grade Spring 2017	M-STEP 3rd grade Spring 2016
	6	M-STEP 5th grade Spring 2017	M-STEP 4th grade Spring 2016
	7	M-STEP 6th grade Spring 2017	M-STEP 5th grade Spring 2016
	8	M-STEP 7th grade Spring 2017	M-STEP 6th grade Spring 2016
SAT	11	M-STEP 8th grade Spring 2015	MEAP 7th grade Fall 2013
M-STEP Social Studies	8	M-STEP 5th grade Spring 2015	
	11	M-STEP 8th grade Spring 2015	MEAP 6th grade Fall 2012
MI-Access ELA & Math	4	MI-Access 3rd grade Spring 2017	
	5	MI-Access 4th grade Spring 2017	MI-Access 3rd grade Spring 2016
	6	MI-Access 5th grade Spring 2017	MI-Access 4th grade Spring 2016
	7	MI-Access 6th grade Spring 2017	MI-Access 5th grade Spring 2016
	8	MI-Access 7th grade Spring 2017	MI-Access 6th grade Spring 2016
	11	MI-Access 8th grade Spring 2015	MI-Access 7th grade Fall 2013
MI-Access Science	7	MI-Access 4th grade Spring 2015	
MI-Access Social Studies	8	MI-Access 5th grade Spring 2015	
	11	MI-Access 8th grade Spring 2015	
WIDA	1	WIDA Kindergarten Spring 2017	
	2	WIDA 1st grade Spring 2017	WIDA Kindergarten Spring 2016
	3	WIDA 2nd grade Spring 2017	WIDA 1st grade Spring 2016
	4	WIDA 3rd grade Spring 2017	WIDA 2nd grade Spring 2016
	5	WIDA 4th grade Spring 2017	WIDA 3rd grade Spring 2016
	6	WIDA 5th grade Spring 2017	WIDA 4th grade Spring 2016
	7	WIDA 6th grade Spring 2017	WIDA 5th grade Spring 2016

Program	Grade 2018	Prior Year 1	Prior Year 2
	8	WIDA 7th grade Spring 2017	WIDA 6th grade Spring 2016
	9	WIDA 8th grade Spring 2017	WIDA 7th grade Spring 2016
	10	WIDA 9th grade Spring 2017	WIDA 8th grade Spring 2016
	11	WIDA 10th grade Spring 2017	WIDA 9th grade Spring 2016
	12	WIDA 11th grade Spring 2017	WIDA 10th grade Spring 2016

Minimum Number of Students

A minimum of 5,000 students will be required for the SGP M-STEP & SAT run.
A minimum of 1,000 students is preferred for the MI-Access FI PRR run.
A minimum of 2,000 students will be required for the SGP WIDA Access for ELLs 2.0 run.

Repeat Test Takers

Students who repeated the grade immediately before the posttest will not be included in either the SGP or the PRR analysis, thus the SGPs were not calculated for these students. For instance, if posttest score (Y_t) and prior 1 year score (Y_{t-1}) are with the same grade, the student is not included in the analysis and does not receive an SGP.

Skipped Grades

Students who skipped the grade immediately prior to the posttest will not be included in the analysis (i.e. 5th grade posttest following skipping 4th grade in the previous example.) In addition, if a student has a test sequence with a skipped grade, only the grade prior will be used to calculate the SGP.

Gaps in Test Sequence

Some students in the dataset are missing certain years of test scores. This may be due to student mobility, missed test windows, or other factors (e.g., Grade 3 M-STEP ELA in Spring 2016, followed by Grade 5 M-STEP ELA in Spring 2018). Students with a gap will not be included unless they have a recent, valid sequence leading up to the posttest.

Home School and Private School Exclusion

All home schooled and private school test records will be excluded from computing SGP. MDE will ensure that students who were previously tested as home schooled or at a private school are also excluded from the data pull.

Student Level Results for SGPs and PRRs

Student level results provided to MDE for SGPs and PRRs included:

1. Demographic and assessment information
2. SGPs
3. SGP standard errors
4. SGP Growth Level Code
5. SGP Norm Group
6. Estimation Method
7. Prior achievement information used

Student Level Results for AGPs

Student level results provided to MDE for AGPs included:

1. Demographic and assessment information
2. AGP Years Projected (1-4)
3. AGP Target
4. AGP Lagged Target
5. AGP Stay/Move Up Target
6. AGP Lagged Stay/Move Up Target

Aggregation

Results were aggregated by assessment and accountability at the state, district, and school level using a variety of subgroups specified by MDE. Aggregation results included:

1. Count of students included
2. Average (arithmetic mean) of the SGPs
3. Standard deviation of SGPs
4. Count of students at each of five growth levels (Significant Improvement, Improvement, Maintain, Decline, Significant Decline)
5. Percentage of students at each of these five levels as a percentage of total students with SGPs
6. Count of students at each of three growth levels (Low, Medium, High)
7. Percentage of students at each of these three levels as a percentage of total students with SGPs.
8. Building z-score

Quality Control

DRC's psychometric team verified the data coming from MDE followed the rules, structure, and specifications agreed upon by both DRC and MDE. Any issues around unexpected data or missing fields were addressed by MDE.

To ensure that the proper growth model was used, base R code was written by the psychometrician and verified by a consultant and a statistical analyst. The code for each subject was reviewed and SGP, PRR, or AGP values were internally checked for reasonability. Two staff members from the psychometric services team verified aggregate results by independent replication, and MDE reviewed the reasonability of the aggregate and individual SGP, PRR, or AGP results. Results went through several iterations of independent replication and MDE review until all discrepancies were resolved.

Summary of Results

Tables 5 through 9 provide a summary of the number of students and median growth SGPs or PRR values by aggregate levels. Tables 5 and 6 provide the summary of number of students and median growth (SGP or PRR) by testing program, calculation method, content area, and grade. Table 7 provides the results by calculation method, content area, and grade. Table 8 provides the results by content area and grade and Table 9 provides the results by grade. As expected with these methods, the median values tend to be near 50.

Table 5: Number of cases and median SGP by testing program, content area, and grade.

Testing Program	Content Area	Grade	N	Median
M-STEP	English Language Arts	4	100,439	50
		5	104,348	50
		6	103,728	50
		7	103,092	50
		8	105,948	50
	Mathematics	4	100,786	50
		5	104,583	50
		6	104,088	50
		7	103,204	50
		8	105,981	50
	Social Studies	8	100,105	49
		11	93,541	50
SAT	English Language Arts	11	93,963	50
	Mathematics	11	93,984	49
WIDA	WIDA	1	8,264	50
		2	9,109	51
		3	9,142	51
		4	8,906	51
		5	6,681	51
		6	5,969	51
		7	5,600	52
		8	5,517	51
		9	5,346	51
		10	4,970	51
		11	3,667	50
		12	2,717	51

Table 6: Number of cases and median PRR by testing program, content area, and grade.

Testing Program	Content Area	Grade	N	Median
MI-Access	English Language Arts	4	959	50
		5	1,174	51
		6	1,237	51
		7	1,285	50
		8	1,302	50
		11	933	51
	Mathematics	4	1,013	50
		5	1,257	51
		6	1,337	51
		7	1,418	51.5
		8	1,425	50
		11	1,042	51
	Science	7	901	50
	Social Studies	8	960	50
		11	959	50

Table 7: Number of cases and median growth by method, content area, and grade.

Method	Content Area	Grade	N	Median
PRR	English Language Arts	4	959	50
		5	1,174	51
		6	1,237	51
		7	1,285	50
		8	1,302	50
		11	933	51
	Mathematics	4	1,013	50
		5	1,257	51
		6	1,337	51
		7	1,418	51.5
		8	1,425	50
		11	1,042	51
	Science	7	901	50
	Social Studies	8	960	50
		11	1,026	50
SGP	English Language Arts	4	100,439	50
		5	104,348	50
		6	103,728	50
		7	103,092	50
		8	105,948	50
		11	93,963	50
	Mathematics	4	100,786	50
		5	104,583	50
		6	104,088	50
		7	103,204	50
		8	105,981	50
		11	93,984	49
	Social Studies	8	100,105	49
		11	93,541	50

Table 8: Number of cases and median growth by content area and grade.

Content Area	Grade	N	Median
English Language Arts	4	101,398	50
	5	105,522	50
	6	104,965	50
	7	104,377	50
	8	107,250	50
	11	94,896	50
Mathematics	4	101,799	50
	5	105,840	50
	6	105,425	50
	7	104,622	50
	8	107,406	50
	11	95,026	49
Science	11	901	50
Social Studies	8	101,065	50
	11	94,567	50

Table 9: Number of cases and median growth by grade.

Grade	N	Median
1	8,264	50
2	9,109	51
3	9,142	51
4	212,103	50
5	218,043	50
6	216,359	50
7	215,500	50
8	321,238	50
9	5,346	51
10	4,970	51
11	288,156	50
12	2,717	51

Goodness of Fit

To examine the fit of the growth models, the correlations between the outcome score (2018) and the prior achievement score was calculated. Tables 10 and 11 provide the correlations by program, content area, and grade. All correlations are acceptable and within the moderate range. For the M-STEP program, all correlations are consistent within content area. In Mathematics and English Language Arts, correlations above 0.80, for Social Studies it is 0.73. With the SAT correlations similar with a correlation of 0.78 for English Language Arts and 0.80 for Mathematics. WIDA correlations are fairly consistent but lower, ranging from 0.65 to 0.81. Finally, the correlations for MI-Access are consistent within content

area but lower ranging from 0.54 to 0.66 for English Language Arts, from 0.48 to 0.62 for Mathematics, 0.51 for Science and 0.42 to 0.51 for Social Studies.

Table 10: Correlation between current SS and prior SS by testing program, content area, and grade for SGP models.

Testing Program	Content Area	Grade	N	Correlation
M-STEP	English Language Arts	4	100,439	0.82
		5	104,348	0.84
		6	103,728	0.83
		7	103,092	0.84
		8	105,947	0.84
	Mathematics	4	100,786	0.84
		5	104,583	0.86
		6	104,088	0.85
		7	103,204	0.87
		8	105,979	0.84
	Social Studies	8	100,105	0.73
		11	93,540	0.76
SAT	English Language Arts	11	93,962	0.78
	Mathematics	11	93,983	0.80
WIDA	WIDA	1	8,264	0.65
		2	9,109	0.76
		3	9,142	0.78
		4	8,906	0.77
		5	6,681	0.77
		6	5,969	0.74
		7	5,600	0.78
		8	5,517	0.81
		9	5,346	0.78
		10	4,970	0.80
		11	3,667	0.76
		12	2,717	0.68

Table 11: Correlation between current SS and prior SS by testing program, content area, and grade for PRR model.

Testing Program	Content Area	Grade	N	Correlation
MI-Access	English Language Arts	4	959	0.59
		5	1,174	0.64
		6	1,237	0.60
		7	1,285	0.66
		8	1,302	0.60
		11	933	0.54
	Mathematics	4	1,013	0.54
		5	1,257	0.62
		6	1,337	0.53
		7	1,418	0.48
		8	1,425	0.58
		11	1,042	0.58
	Science	7	901	0.51
	Social Studies	8	960	0.42
		11	1,026	0.51

Distributions of SGPs and PRRs

The distributions of SGPs and PRRs are provided in Figure 1 through Figure 3, which shows that SGPs tend to uniformly range from 1 to 99. While the PRRs also range from 1 to 99, they are a bit less stable due to the small sample sizes used in the calculations. It should be noted that the differences distributions of PRRs and SGPs across grade and content area tend to be relatively small given the scale of the density plots range from 0 to 0.012.

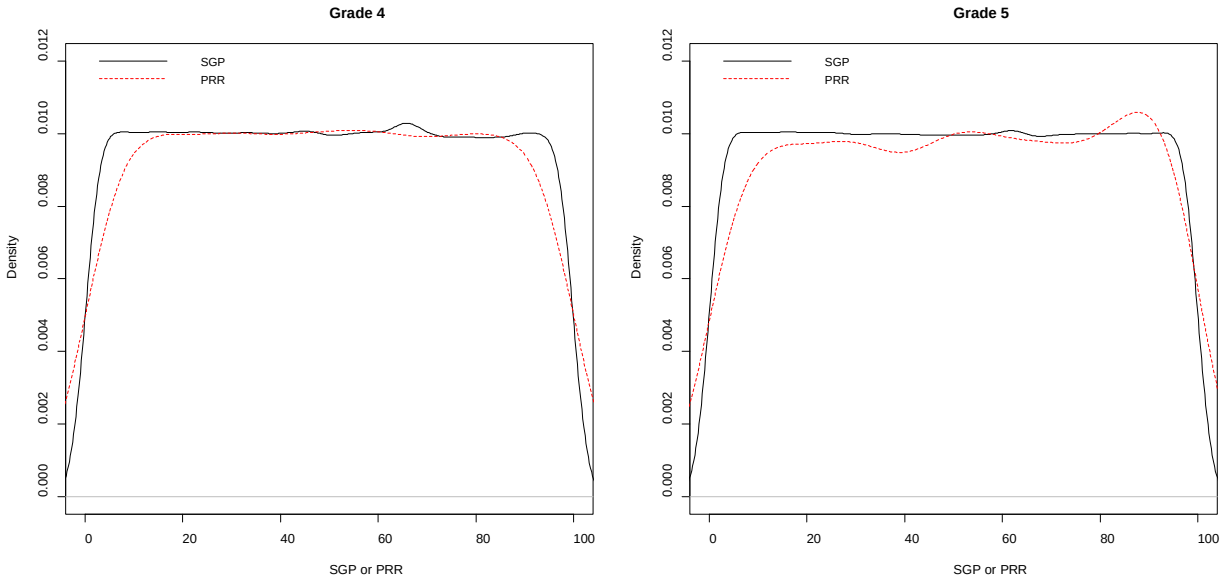


Figure 1. Distribution of SGP/PRR for Mathematics Grades, 4 and 5

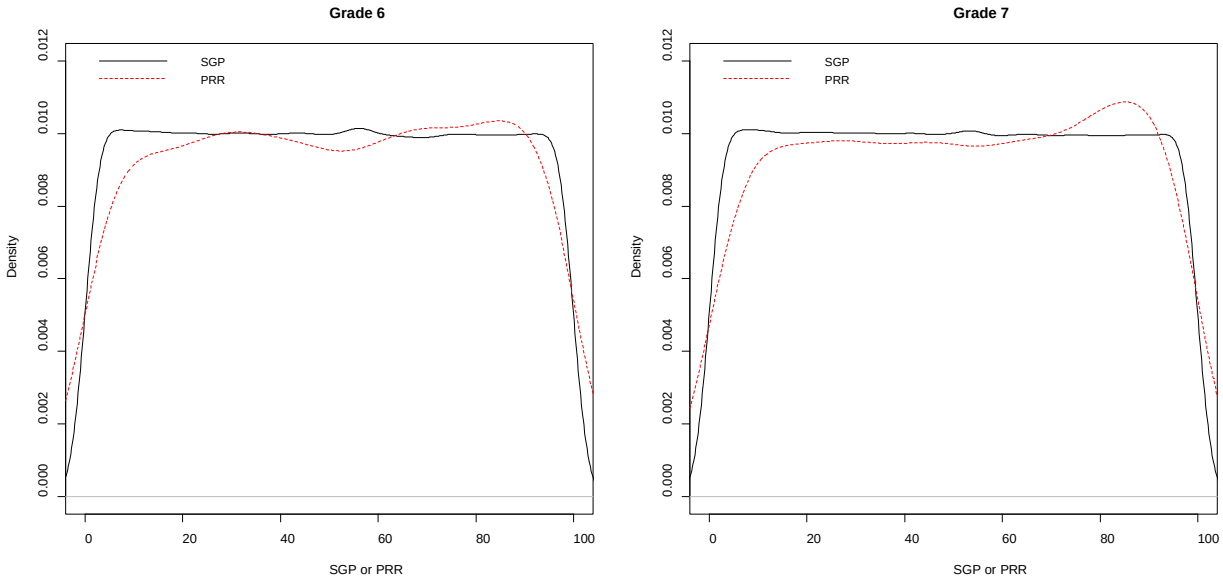


Figure 2. Distribution of SGP/PRR for Mathematics Grades, 6 and 7

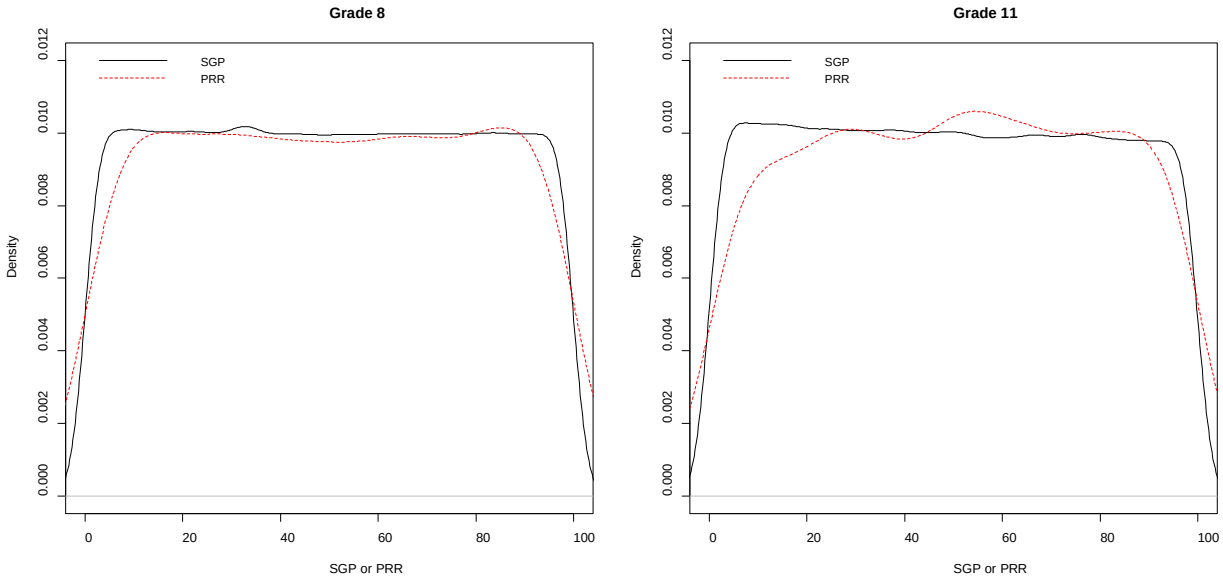


Figure 3. Distribution of SGP/PRR for Mathematics Grades, 8 and 11

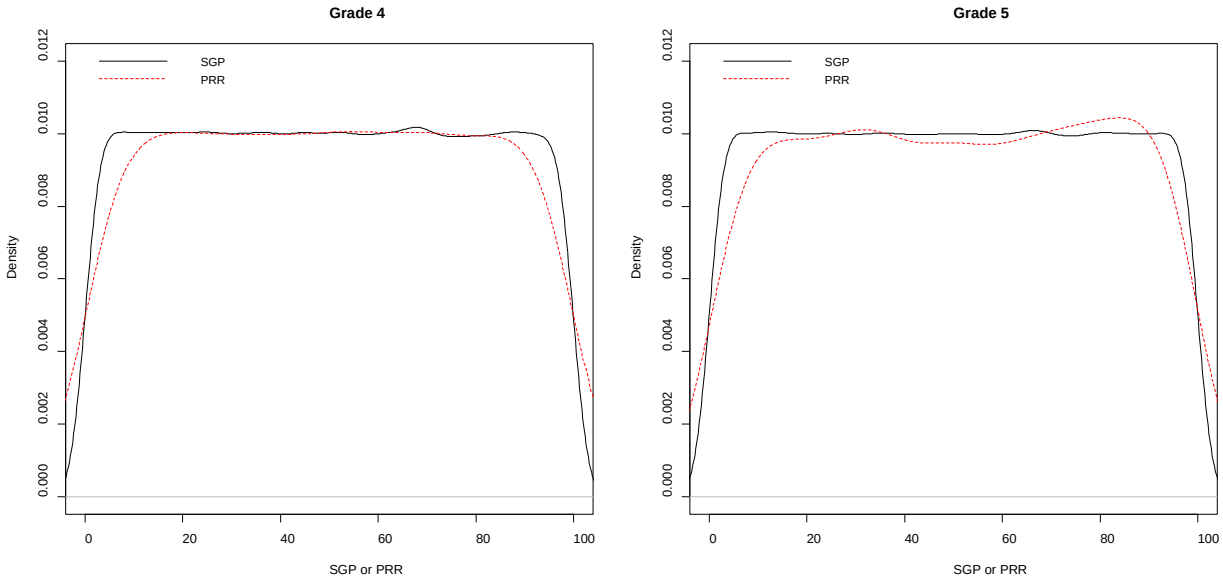


Figure 4. Distribution of SGP/PRR for English Language Arts Grades, 4 and 5

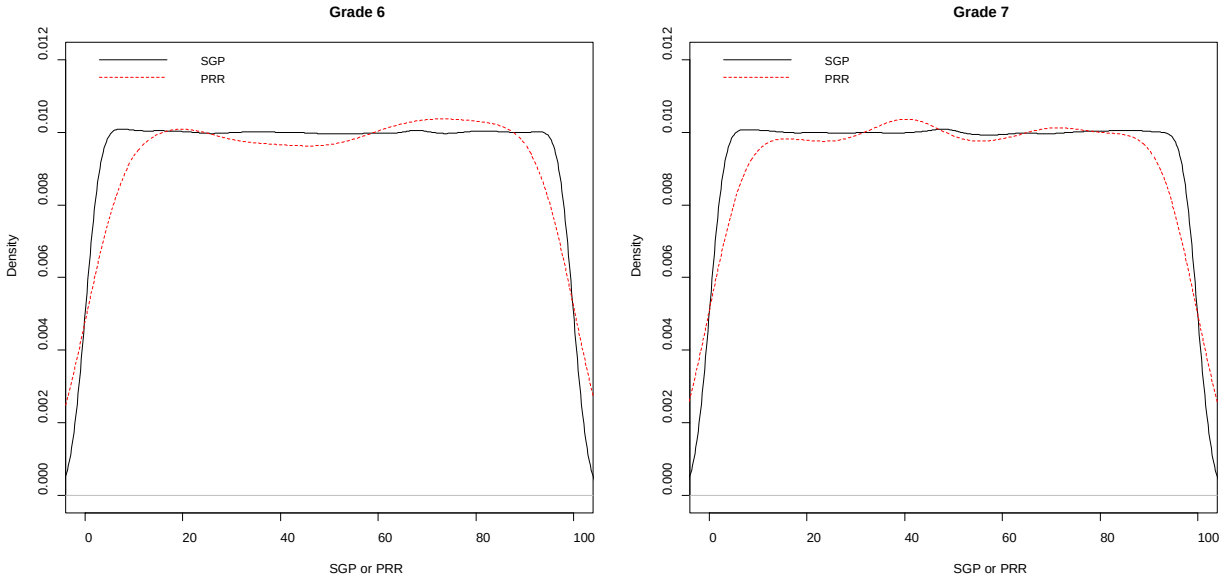


Figure 5. Distribution of SGP/PRR for English Language Arts Grades, 6 and 7

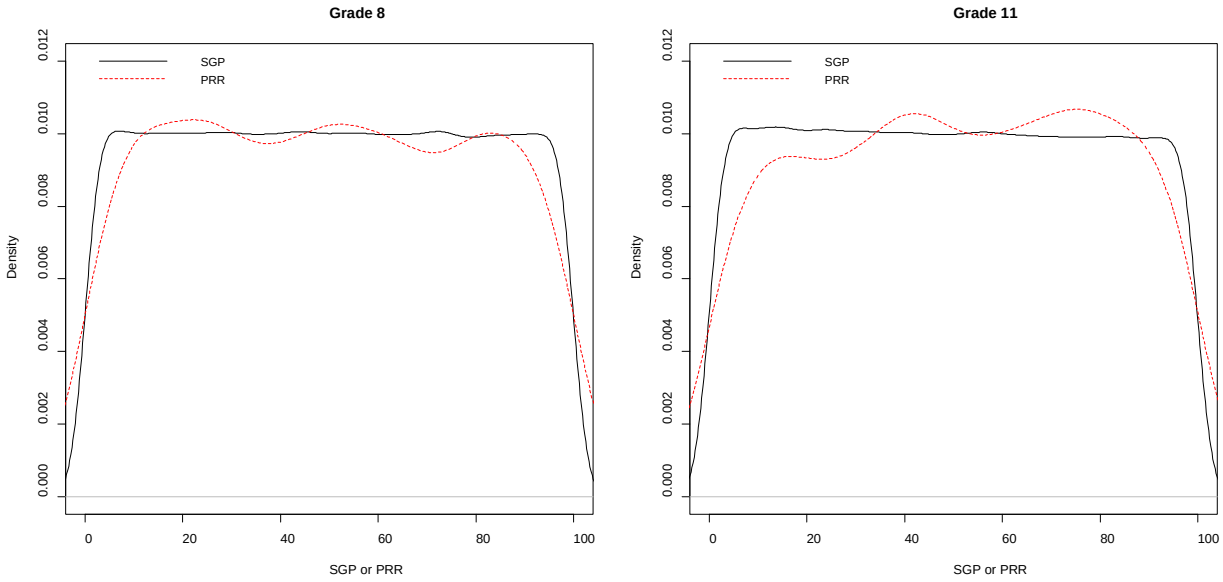


Figure 6. Distribution of SGP/PRR for English Language Arts Grades, 8 and 11

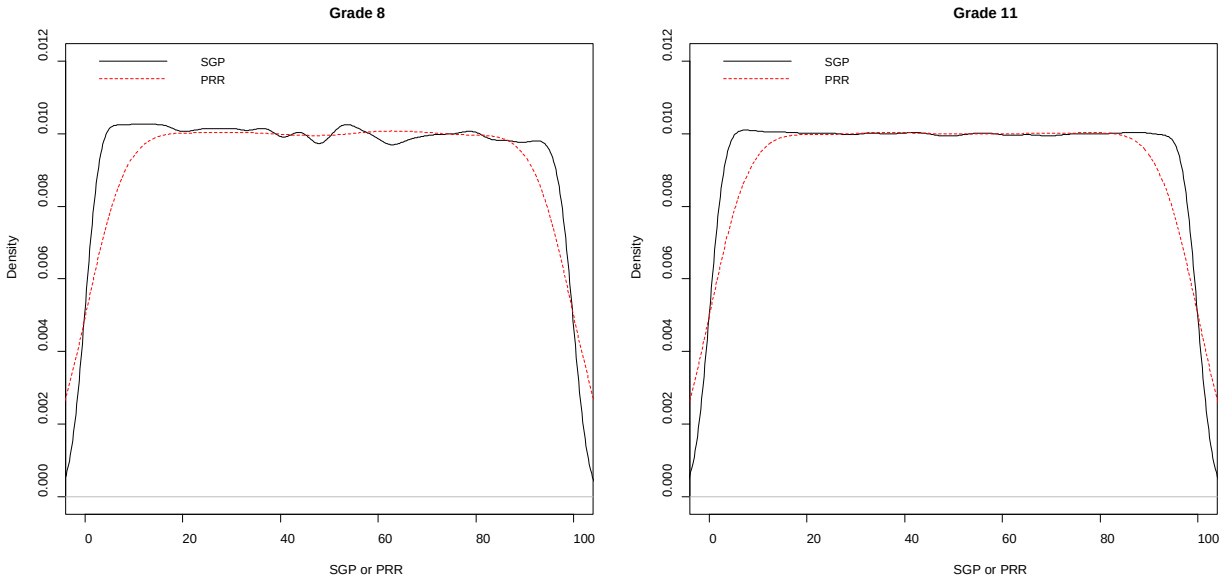


Figure 7. Distribution of SGP/PRR for Social Studies Grades, 8 and 11

Checks for Neutrality

Since the growth models used in this analysis do not control for demographic variables, particularly those that may have some impact on student growth rates and trajectories, it is unknown whether the results are biased, especially when aggregated at the school or district level (Education Analytics, 2015). Thus, it is important to look at the relationship between the aggregated growth measure, in this case median SGP and the variables of interest that were not controlled for in the growth models. It is important to note that it is unknown what the correlations "should be." Tables 12 and 13 provide the

correlations between the median SGP for a school or a district (with more than 20 students) related to the percentage of each demographic for that building or district. Graphs of these relationships can be found in the appendix.

Table 12: Correlations between Median SGP and Demographic at the school level.

Content Area	ED	SE	LEP	Non-White
English Language Arts	-0.37	-0.20	0.10	-0.18
Mathematics	-0.39	-0.20	0.04	-0.22
Social Studies	-0.38	-0.21	-0.06	-0.23
WIDA	-0.43	-0.12		-0.18

Table 13: Correlations between Median SGP and Demographic at the district level.

Content Area	ED	SE	LEP	Non-White
English Language Arts	-0.28	-0.23	0.12	-0.05
Mathematics	-0.35	-0.24	0.05	-0.15
Social Studies	-0.35	-0.20	0.00	-0.15
WIDA	-0.34	-0.09		-0.27

When aggregating growth model outcomes, it is also important to note that growth models, as with most regression models, have issues (more variability or less precision) when sample sizes are small. This is also true when aggregating growth model results at the school level. Figure 8 provides the relationship between the number of students and SGP. This shows that there is less variability in median SGP as the number of students increase.

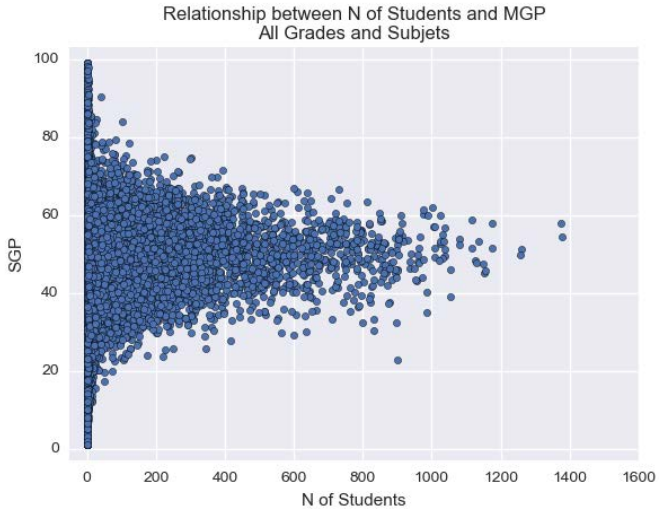


Figure 8. Number of Students versus SGP

AGP Outcomes

In 2018, AGPs and target AGPs were computed for M-STEP ELA and Mathematics, grades 4 through 8. The number of years projected in the model was varied between 1 and 4. Details can be found in Tables 2 and 3. One way to aggregate these results is to compare the percentage of students meeting targets by their 2018 performance level, grade, and years projected. Tables 14 and 15 do this by showing the percentage of students, by grade, who have a 2018 SGP greater than their 2018 lagged AGP, broken down by proficiency level, grade, and years projected. For example, in Grade 4 ELA, 62% of proficient students are on track to remain proficient (or reach advanced) in three years' time. These tables show that students who end in the highest performance level (Advanced) do so because they consistently grew at levels surpassing that which was necessary to achieve and maintain proficiency. Similarly, they also show that students who end in the lowest performance level (Not Proficient) do so because they consistently grew at levels well below what was necessary to reach proficiency.

Table 14: Percentage of students whose 2018 SGP exceeds their lagged by performance level and years projected for M-STEP ELA.

		Not Proficient		Partially Proficient		Proficient		Advanced	
Grade	Years Projected	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP
4	1	33,350	0%	21,282	17%	21,880	87%	23,927	100%
	2	33,350	0%	21,282	23%	21,880	65%	23,927	97%
	3	33,350	2%	21,282	30%	21,880	62%	23,927	94%
	4	33,350	4%	21,282	33%	21,880	59%	23,927	90%
5	1	32,832	0%	22,341	4%	30,314	76%	18,861	100%
	2	32,832	0%	22,341	18%	30,314	69%	18,861	98%
	3	32,832	2%	22,341	25%	30,314	64%	18,861	95%
	4	32,832	2%	22,341	25%	30,314	64%	18,861	95%
6	1	31,766	0%	28,509	11%	29,568	88%	13,885	100%
	2	31,766	0%	28,509	22%	29,568	75%	13,885	100%
	3	31,766	0%	28,509	22%	29,568	75%	13,885	100%
	4								
7	1	29,367	0%	28,366	8%	31,995	85%	13,364	100%
	2	29,367	0%	28,366	8%	31,995	85%	13,364	100%
	3								
	4								
8	1	30,927	0%	29,029	0%	33,376	96%	12,616	100%
	2								
	3								
	4								

Table 15: Percentage of students whose 2018 SGP exceeds their lagged by performance level and years projected for M-STEP Math.

		Not Proficient		Partially Proficient		Proficient		Advanced	
Grade	Years Projected	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP	N Total	% 2018 SGP Exceeds Lagged AGP
4	1	24,351	0%	33,674	2%	26,124	72%	16,637	100%
	2	24,351	0%	33,674	10%	26,124	66%	16,637	99%
	3	24,351	0%	33,674	18%	26,124	65%	16,637	98%
	4	24,351	1%	33,674	22%	26,124	60%	16,637	94%
5	1	38,194	0%	29,996	8%	18,818	80%	17,575	100%
	2	38,194	0%	29,996	22%	18,818	74%	17,575	99%
	3	38,194	2%	29,996	28%	18,818	64%	17,575	95%
	4	38,194	2%	29,996	28%	18,818	64%	17,575	95%
6	1	35,224	0%	32,486	11%	19,558	88%	16,820	100%
	2	35,224	0%	32,486	22%	19,558	71%	16,820	97%
	3	35,224	0%	32,486	22%	19,558	71%	16,820	97%
	4	35,224	0%	32,486	11%	19,558	88%	16,820	100%
7	1	36,724	0%	29,080	10%	20,304	74%	17,096	100%
	2	36,724	0%	29,080	10%	20,304	74%	17,096	100%
	3								
	4								
8	1	41,907	0%	27,855	1%	17,000	92%	19,219	100%
	2								
	3								
	4								

References

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories.* Downloaded March 9, 2018 from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf.

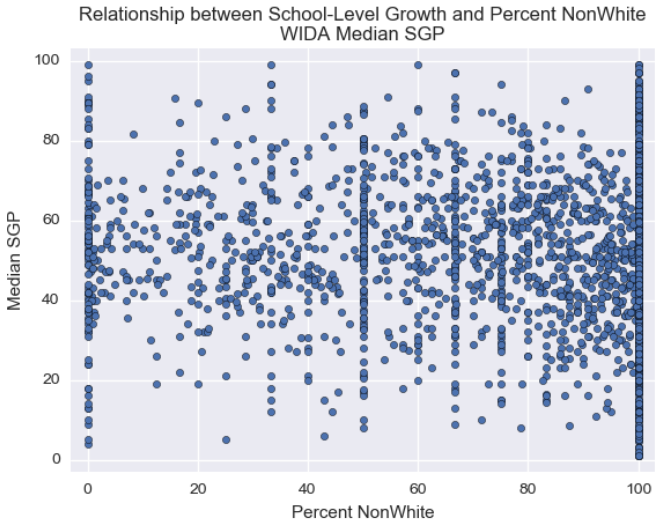
Betebenner, D. W., VanIwaarden, A., Domingue, B., and Shang, Y. (2016). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories. R package version 1.5-0.0. https://cran.r-project.org/web/packages/SGP/

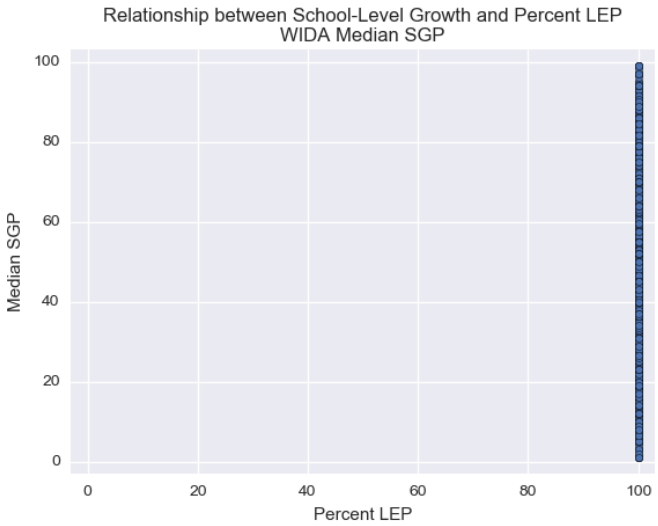
Castellano, K.E., and Ho, A.D. (2015). Practical Differences Among Aggregate-Level Conditional Status Metrics: From Median Student Growth Percentiles to Value-Added Models. Journal of Educational and Behavioral Statistics, 40(1), 35-68. doi:10.3102/107699861454848

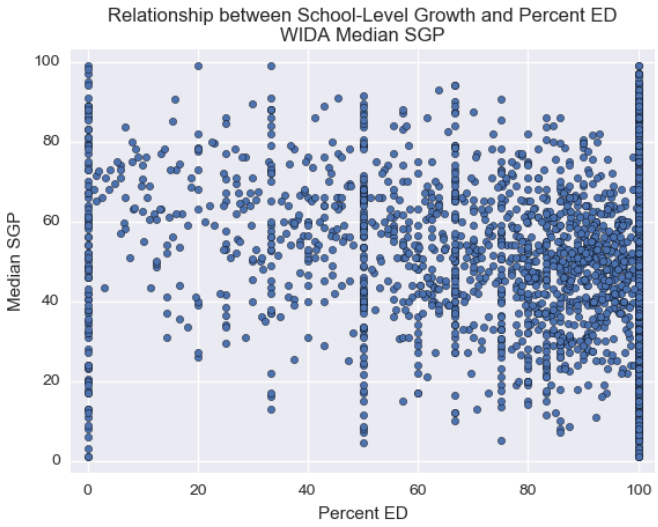
Castellano, K.E., and Ho, A.D. (2013). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles. Journal of Educational and Behavioral Statistics, 38(2), 190-215. doi:10.3102/1076998611435413

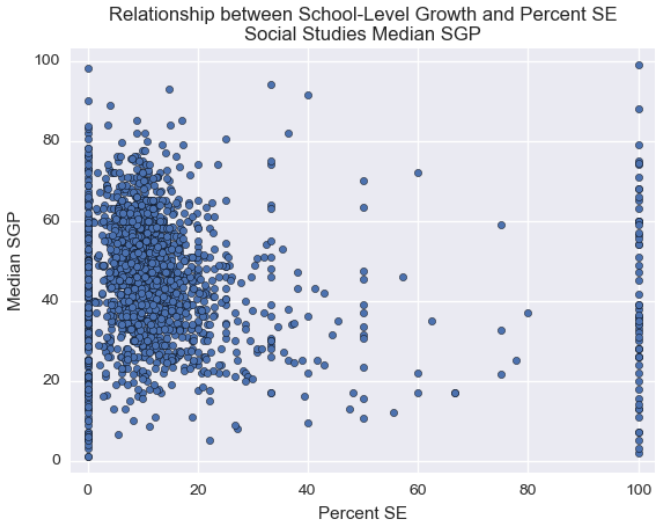
Education Analytics (2015). Michigan Department of Education Technical Report.

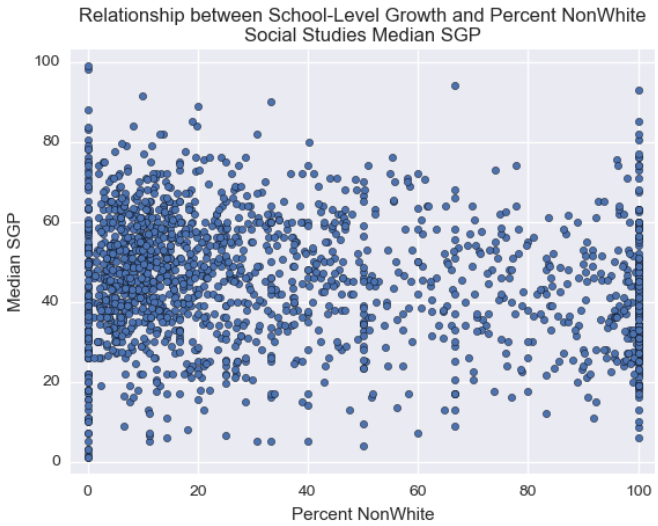
Appendix

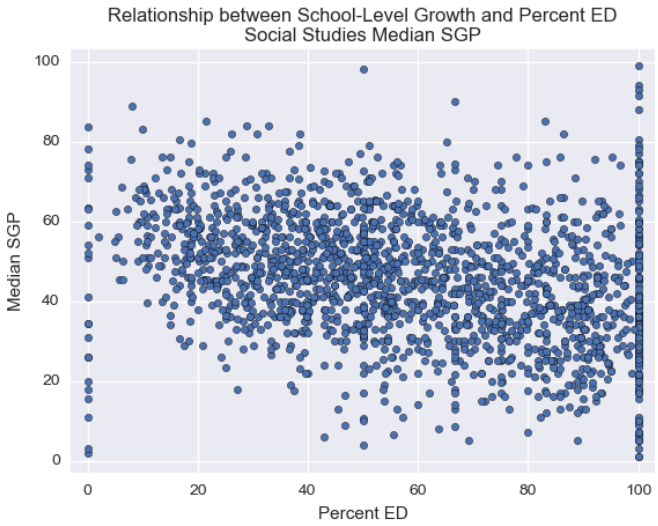

Relationship between School-Level Growth and Percent NonWhite
WIDA Median SGP
Median SGP
Percent NonWhite
0
20
40
60
80
100

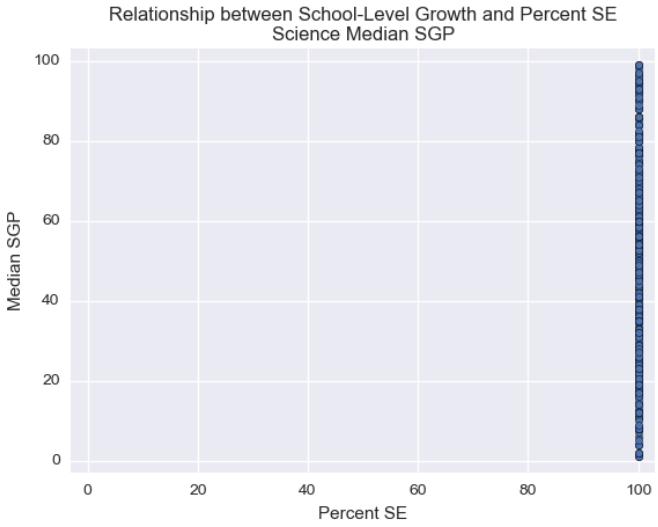

Relationship between School-Level Growth and Percent LEP
WIDA Median SGP
Median SGP
Percent LEP
0
20
40
60
80
100

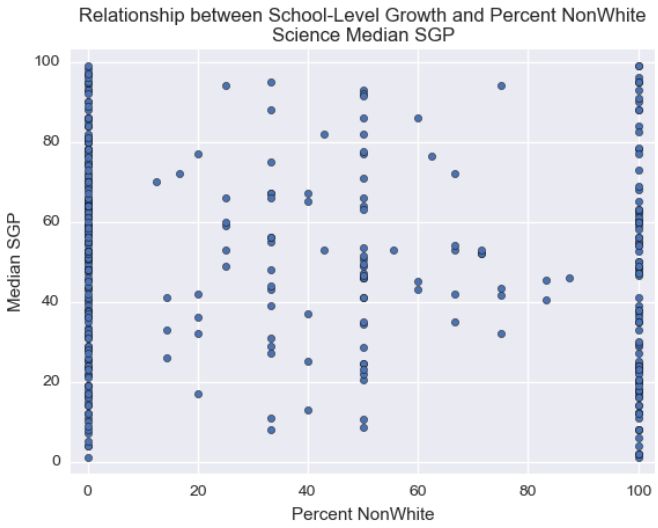

Relationship between School-Level Growth and Percent ED
WIDA Median SGP
Median SGP
Percent ED
0
20
40
60
80
100

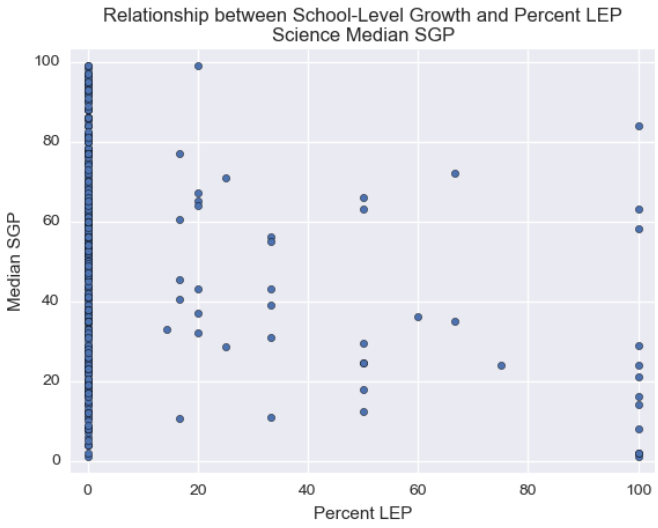

Relationship between School-Level Growth and Percent SE
Social Studies Median SGP
Median SGP
Percent SE
0
20
40
60
80
100

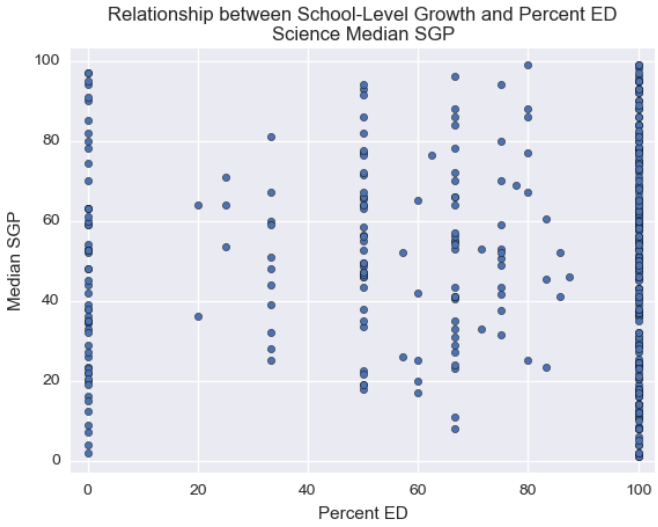

Relationship between School-Level Growth and Percent NonWhite
Social Studies Median SGP
Median SGP
Percent NonWhite
0
20
40
60
80
100

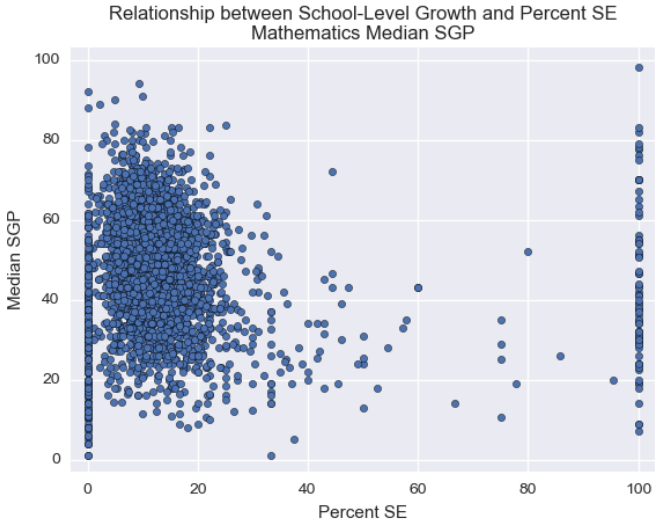

Relationship between School-Level Growth and Percent ED
Social Studies Median SGP
Median SGP
Percent ED
0
20
40
60
80
100

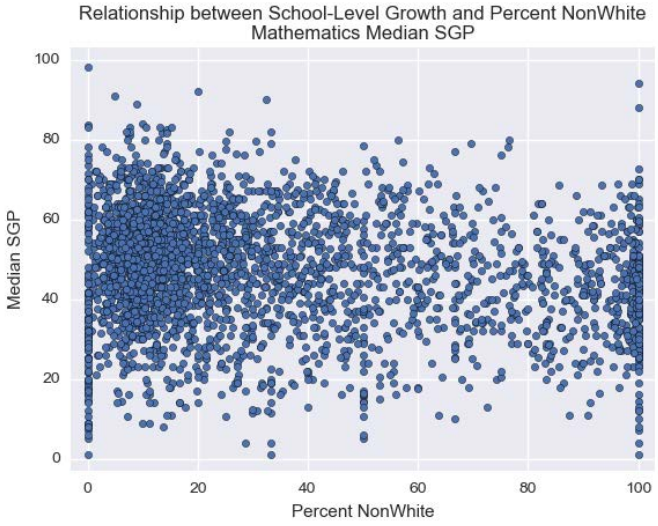

Relationship between School-Level Growth and Percent SE
Science Median SGP
Median SGP
Percent SE
0
20
40
60
80
100

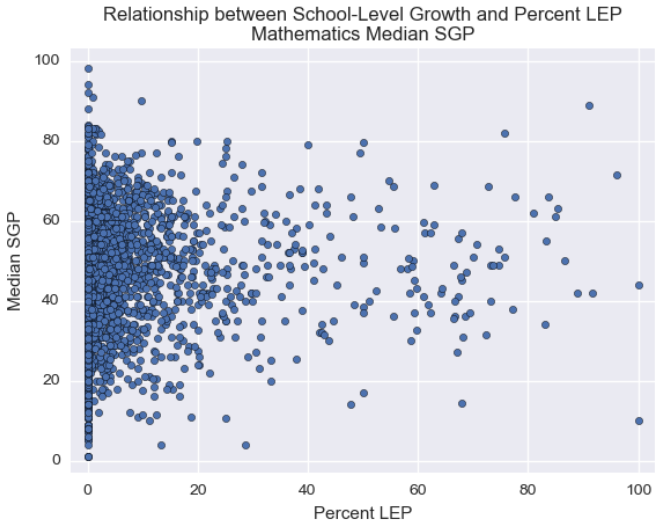

Relationship between School-Level Growth and Percent NonWhite
Science Median SGP
Median SGP
Percent NonWhite
0
20
40
60
80
100

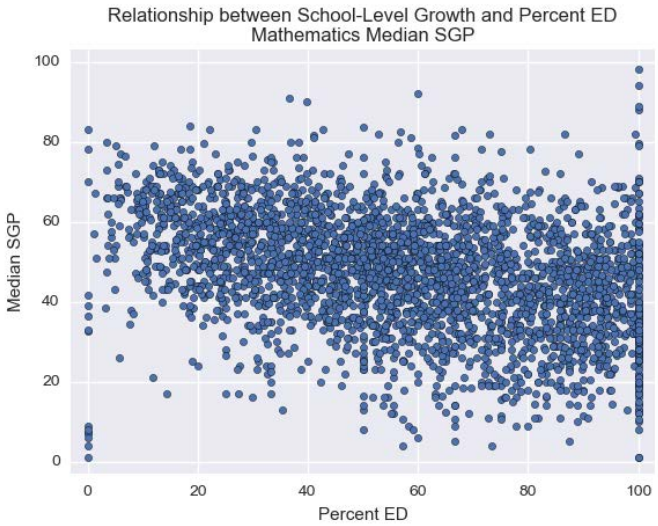

Relationship between School-Level Growth and Percent LEP
Science Median SGP
Median SGP
Percent LEP
0
20
40
60
80
100

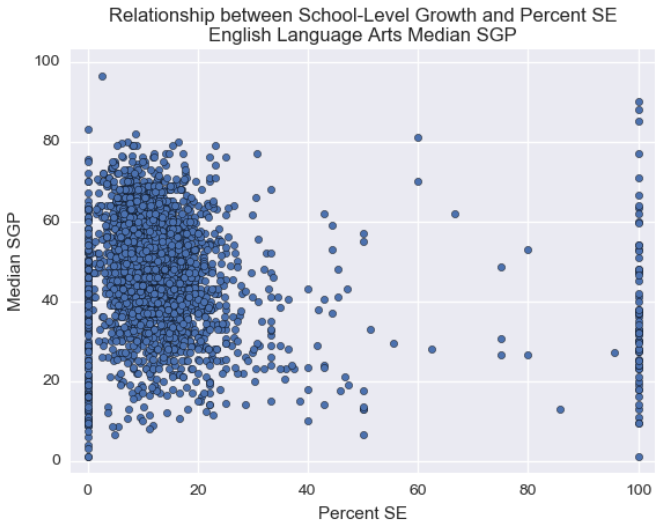

Relationship between School-Level Growth and Percent ED
Science Median SGP
Median SGP
Percent ED
0
20
40
60
80
100

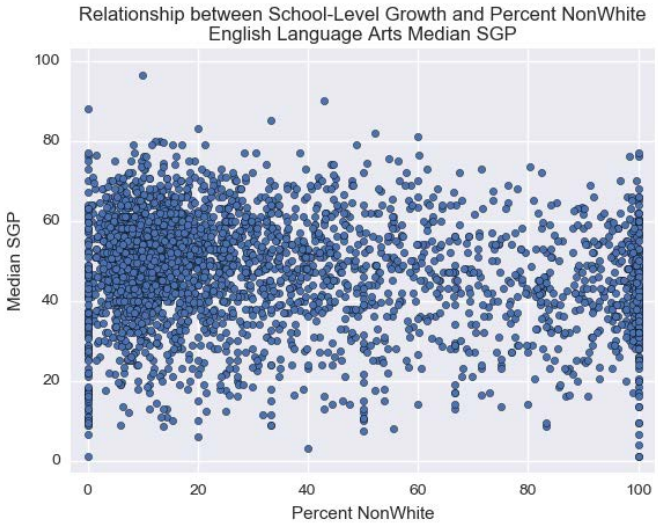

Relationship between School-Level Growth and Percent SE
Mathematics Median SGP
Median SGP
Percent SE
0
20
40
60
80
100

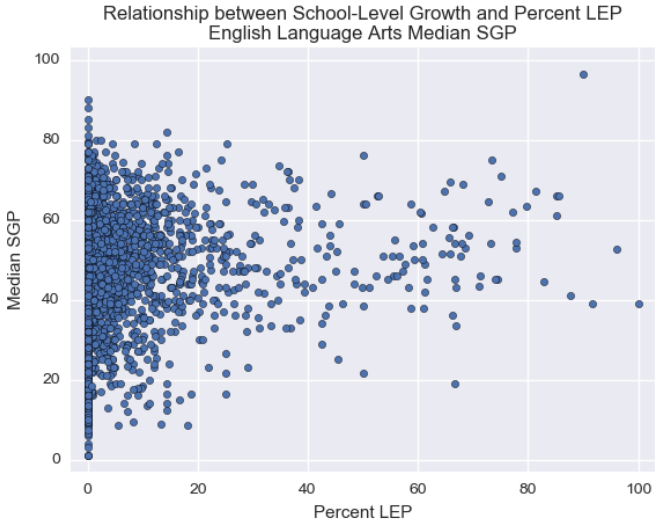

Relationship between School-Level Growth and Percent NonWhite
Mathematics Median SGP
Median SGP
Percent NonWhite
0
20
40
60
80
100

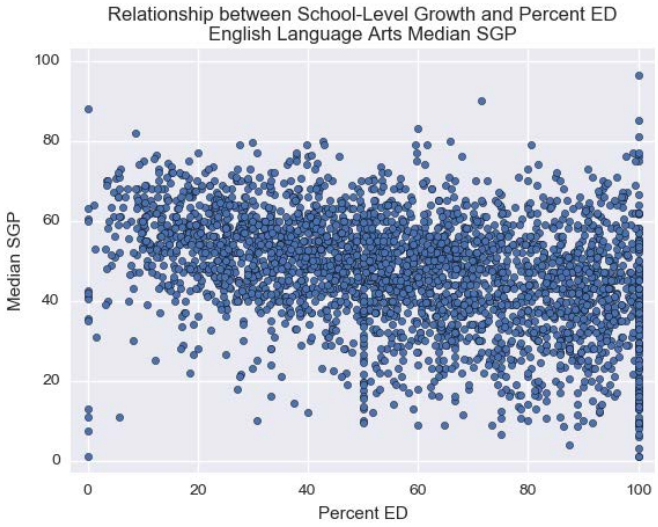

Relationship between School-Level Growth and Percent LEP
Mathematics Median SGP
Median SGP
Percent LEP
0
20
40
60
80
100

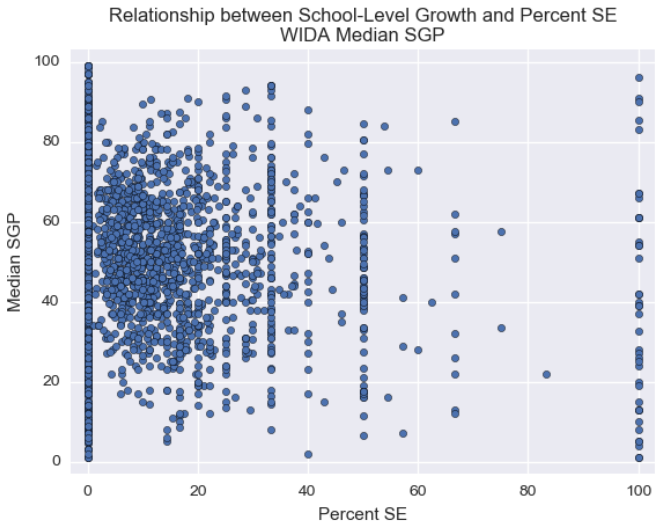

Relationship between School-Level Growth and Percent ED
Mathematics Median SGP
Median SGP
Percent ED
0
20
40
60
80
100


Relationship between School-Level Growth and Percent SE
English Language Arts Median SGP
Median SGP
Percent SE
0
20
40
60
80
100


Relationship between School-Level Growth and Percent NonWhite
English Language Arts Median SGP
Median SGP
Percent NonWhite
0
20
40
60
80
100


Relationship between School-Level Growth and Percent LEP
English Language Arts Median SGP
Median SGP
Percent LEP
0
20
40
60
80
100


Relationship between School-Level Growth and Percent ED
English Language Arts Median SGP
Median SGP
Percent ED
0
20
40
60
80
100


Relationship between School-Level Growth and Percent SE
WIDA Median SGP
Median SGP
Percent SE
0
20
40
60
80
100

Appendix E: M-STEP Standards Validation

Appendix E-1. Validity Evidence for English Language Arts and Mathematics Cut Scores

Ricardo Mercado, Jessalyn Smith, Sara Kendallen, Mayuko Simon, Alassane Savadogo, and Ben Sorenson *Data Recognition Corporation*

July 15, 2018

Appendix E-2 Summary

- On July 9–12, 2018, the Michigan Department of Education (MDE) partnered with Data Recognition Corporation (DRC) to conduct a standards validation for the Michigan Student Test of Educational Progress (M-STEP) tests of English language arts (ELA) and mathematics for grades 3–8.
- The *standards validation* was needed because of test-length reductions implemented in spring 2018. Specifically, proportional reductions in the number of items by reporting category were implemented for mathematics; and for ELA, new passage-based writing items replaced other performance tasks.
- The purpose of the standards validation workshop was to determine whether the existing M STEP cut scores were still valid for continued use on the updated tests.
- Participants' recommendations at the standards validation were consistent with the existing cut scores, providing evidence of their validity for continued use.

Appendix E-3 Background

The M-STEP is administered to assess Michigan students' mastery of the Michigan Academic *Standards*. The assessments began as an implementation of the Smarter Balanced Assessment Consortium's (SBAC) ELA and mathematics tests. The current cut scores for the tests are taken from the SBAC tests.

Over the course of several years, important changes have been made to the assessments to make them more meaningful to Michigan educators. These include the alignment of the test items to the Michigan Academic *Standards*, the implementation of a Michigan-specific test blueprint, and a reduction in the number of performance tasks used in ELA to reduce overall test time. These changes were made cautiously and deliberately with the active involvement of Michigan educators and stakeholders.

In school year 2017–18, the tests in grades 3–8 were shortened to reduce the time burden on students and schools. To do so, all performance tasks in ELA were replaced with passage-based writing items, a new item type for Michigan. The ELA test blueprints were adjusted to accommodate the new item type and the reduction in test length. In grades 3–8 mathematics, the test was also shortened to reduce overall testing time, but this change did not involve adding new test items or significantly altering the test blueprint.

Appendix E-4 Standards Validation Methodology

The purpose of the standards validation was to determine whether the current M-STEP cut scores for grades 3–8 ELA and mathematics were still valid for continued use, given the 2018 updates to the tests.

A total of 54 Michigan educators engaged in a modification of the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) to validate the cut scores. This method has been used on large-scale assessments in Michigan and across the nation, including for SBAC.

Participants studied the existing Michigan performance level descriptors (PLDs) and Michigan Learning *Standards* to review the knowledge, skills, and abilities expected of students in each performance level. The four performance levels on M-STEP are *Not Proficient*, *Partially Proficient*, *Proficient*, and *Advanced*. Each performance level is associated with a level of mastery of the Michigan Learning *Standards*. Participants then discussed the content-based expectations for students at the threshold of each performance level (e.g., a student who is just *Proficient*). To support their discussions of these threshold students, participants were provided with the SBAC achievement level descriptors (ALDs). These SBAC ALDs were used at the original standard setting where the cut scores were established.

Participants studied collections of test items that were ordered in terms of difficulty. The existing cut scores were presented as *benchmarks* for participants' consideration: participants were asked to consider the knowledge and skills that students would need to demonstrate on the updated ELA and mathematics tests, as based on the benchmarked (existing) cut scores. Then, participants compared these expectations against the content-based expectations for students at the thresholds of each performance level. Participants were instructed to recommend retaining the existing cut scores if there was good correspondence between the benchmarks and these content-based expectations, or to recommend alternative cut scores that reflect better correspondence. Participants engaged in two rounds of individual judgments and group discussion. (The grade 5 mathematics committee engaged in three rounds of judgments to accommodate additional discussion.) The committees' median judgments were taken as their final recommendations.

Before the workshop, it was hypothesized that participants would recommend cut scores which were similar to, but not exactly equal to, the existing cut scores. The rationale behind this hypothesis was that nearly any group of educators going through an iterative, judgmental process like the Bookmark Procedure will tend to arrive at slightly different judgments at the end of the process. Accordingly, it was not expected that standards validation participants would recommend cut scores exactly equal to the existing cut scores: slight differences in cut score recommendations could be attributed to random statistical errors. This hypothesis was later used to inform the interpretation of the workshop results, presented under the heading "Review of Recommendations Made at the *Standards* Validation."

Table E-1 shows the median recommended cut scores from the standards validation workshop plus the associated impact data for ELA and mathematics using Spring 2018 administration data. Impact data are the percentages of students who would be classified in each performance level if the cut scores were applied to students' scores. Note that the impact data presented in this document are based on the test data available at the time of the standards validation, so they should not be considered final; however, these impact data provide a reasonable estimate of the percentages of students that would be included in each performance level based on the cut scores shown.

Table E-1a. Cut Scores Associated with Participants' Median Recommendations

Content	Grade	Partially Proficient	Proficient	Advanced
ELA	3	1279	1299.5	1316
ELA	4	1382	1399.5	1417
ELA	5	1481	1499.5	1521
ELA	6	1578	1599.5	1624
ELA	7	1679	1699.5	1726
ELA	8	1775	1794	1828
Math	3	1281	1299.5	1321
Math	4	1376	1397	1417
Math	5	1475	1496	1515
Math	6	1579	1599.5	1614
Math	7	1679	1699.5	1715
Math	8	1777	1799.5	1815

Table E-1b. Impact Data Associated with Participants' Median Recommendations

Content	Grade	Not Proficient	Partially Proficient	Proficient	Advanced
ELA	3	29.70%	25.80%	21.10%	23.30%
ELA	4	32.60%	22.30%	21.50%	23.60%
ELA	5	32.10%	21.20%	25.30%	21.40%
ELA	6	31.30%	27.30%	28.20%	13.20%
ELA	7	29.20%	27.30%	30.70%	12.80%
ELA	8	27.50%	22.50%	38.20%	11.80%
Math	3	27.80%	26.40%	27.20%	18.60%
Math	4	24.70%	28.70%	26.90%	19.70%
Math	5	33.50%	26.60%	23.30%	16.60%
Math	6	34.40%	30.90%	18.60%	16.00%
Math	7	36.20%	28.00%	18.30%	17.50%
Math	8	36.50%	30.80%	14.90%	17.80%

Appendix E-5 Review of the Recommendations Made at the Standards Validation

As hypothesized, educators at the content-based standards validation workshop recommended cut scores that were similar to the existing cut scores. MDE and DRC evaluated the recommendations in context. Table E-2 shows the difference between the median cut score recommendations and the existing cut scores, expressed in multiples of the conditional standard error of measurement (CSEM). The CSEM quantifies the amount of statistical error associated with the test. If a student were tested many times, one would expect her scores to fall within a range of ±1.0 CSEM about 2/3 of the time.

Figures E-1 and E-2 show a graphical representation of the existing cut scores beside the recommended cut scores and their associated CSEM.

Table E-2a. Median Cut Score Recommendations from the Standards Validation, Existing ELA and Math Cut Scores, and Differences in Terms of Conditional Standard Error of Measurement (CSEM)

Content	Grade	Partially Proficient	Proficient	Advanced
ELA	3	1279	1299.5	1316
ELA	4	1382	1399.5	1417
ELA	5	1481	1499.5	1521
ELA	6	1578	1599.5	1624
ELA	7	1679	1699.5	1726
ELA	8	1775	1794	1828
Math	3	1281	1299.5	1321
Math	4	1376	1397	1417
Math	5	1475	1496	1515
Math	6	1579	1599.5	1614
Math	7	1679	1699.5	1715
Math	8	1777	1799.5	1815

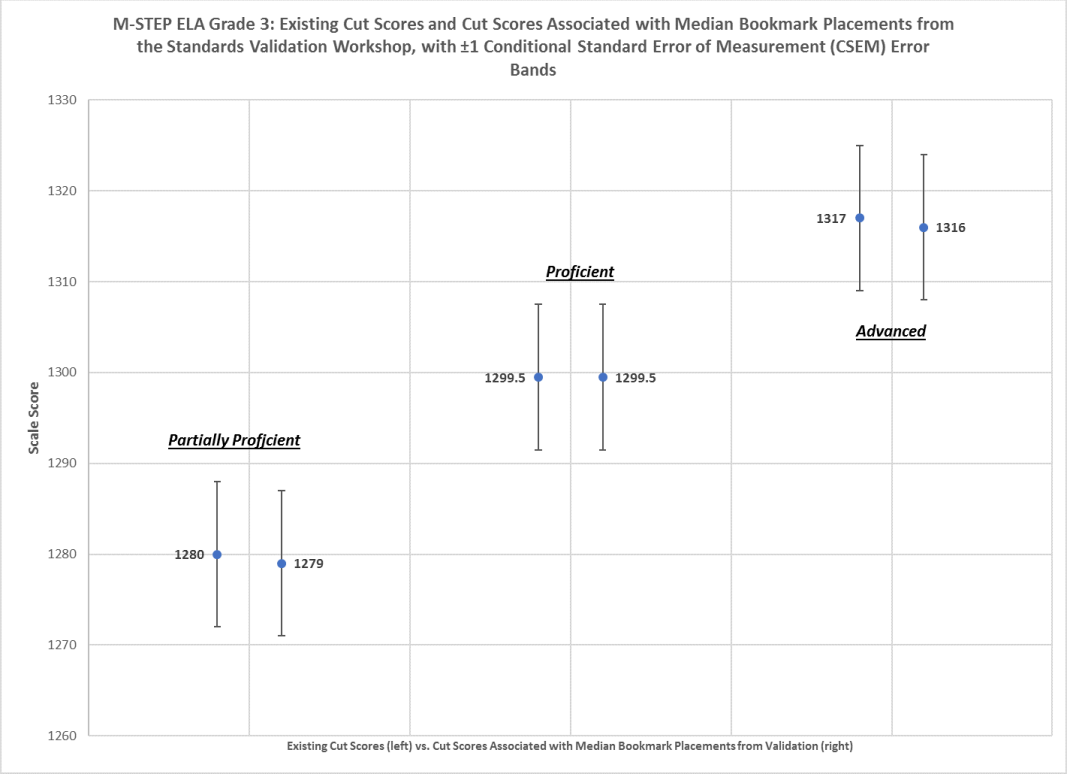
Table E-2b. Median ELA and Math Cut Scores

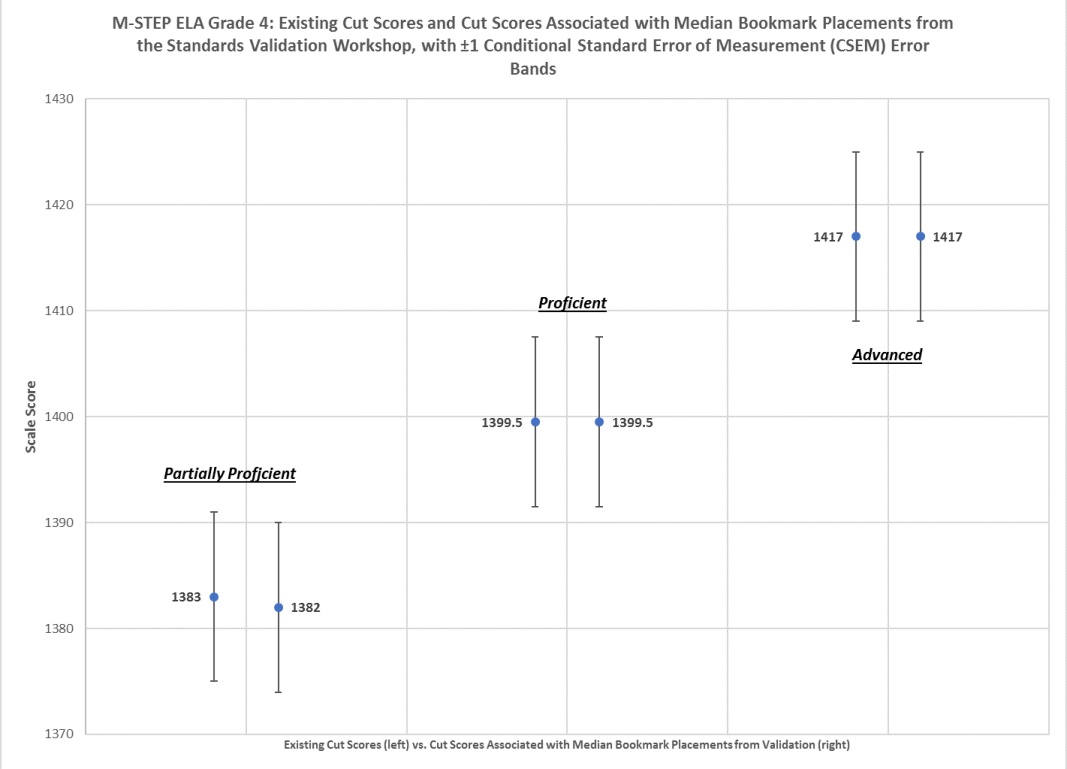
Content	Grade	Partially Proficient	Proficient	Advanced
ELA	3	1280	1299.5	1317
ELA	4	1383	1399.5	1417
ELA	5	1481	1499.5	1524
ELA	6	1578	1599.5	1624
ELA	7	1679	1699.5	1726
ELA	8	1777	1799.5	1828
Math	3	1281	1299.5	1321
Math	4	1376	1399.5	1420
Math	5	1478	1499.5	1515
Math	6	1579	1599.5	1614
Math	7	1679	1699.5	1716
Math	8	1780	1799.5	1815

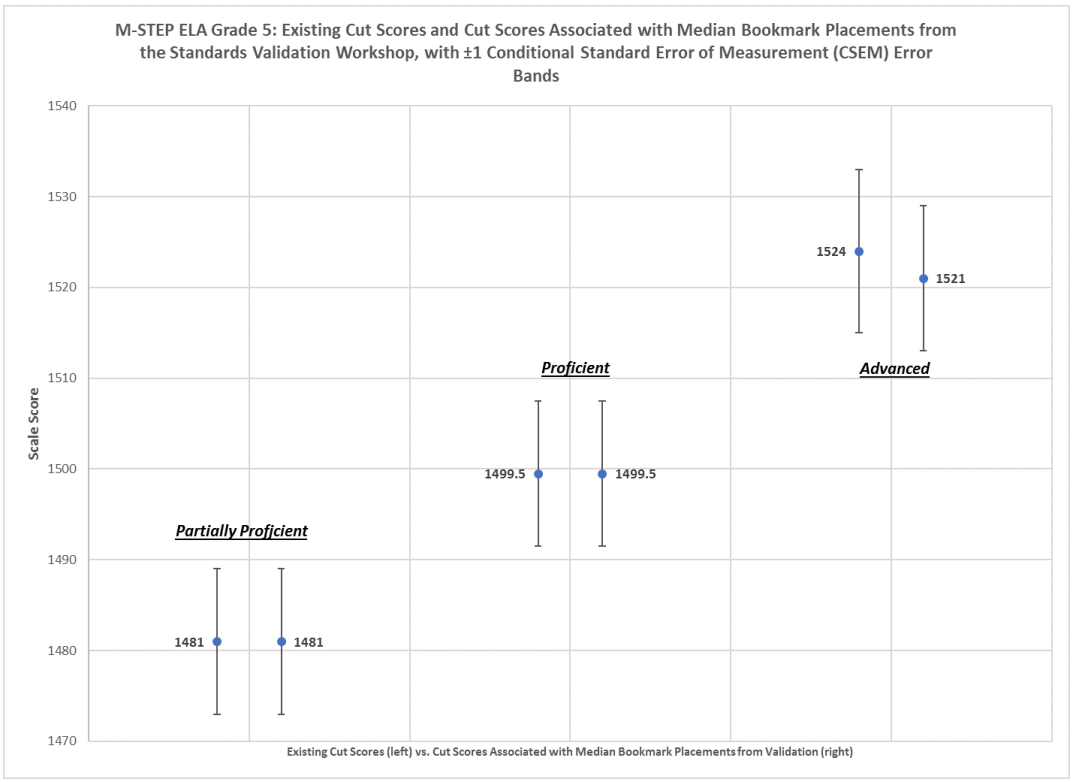
Table E-2c. Differences between Existing and Recommended Cut Scores in Terms of Conditional Standard Error of Measurement (CSEM)

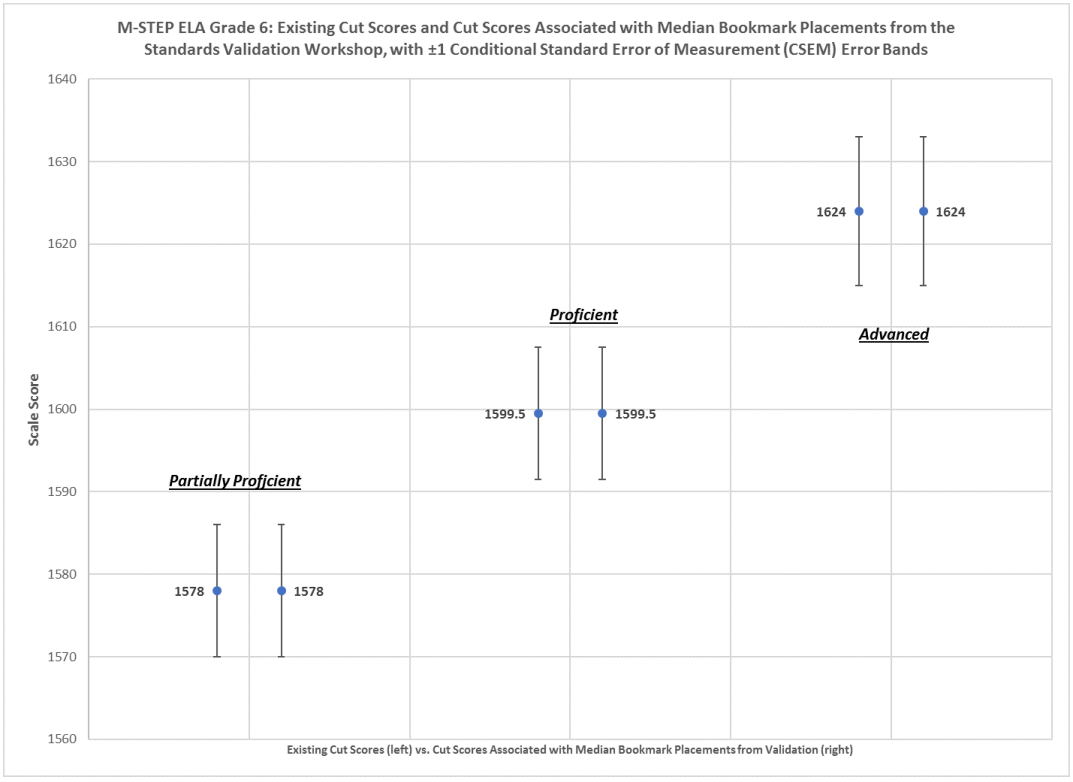
Content	Grade	Partially Proficient	Proficient	Advanced
ELA	3	-0.13	0	-0.13
ELA	4	-0.13	0	0
ELA	5	0	0	-0.38
ELA	6	0	0	0
ELA	7	0	0	0
ELA	8	-0.22	-0.69	0
Math	3	0	0	0
Math	4	0	-0.42	-0.43
Math	5	-0.33	-0.44	0
Math	6	0	0	0
Math	7	0	0	-0.17
Math	8	-0.33	0	0

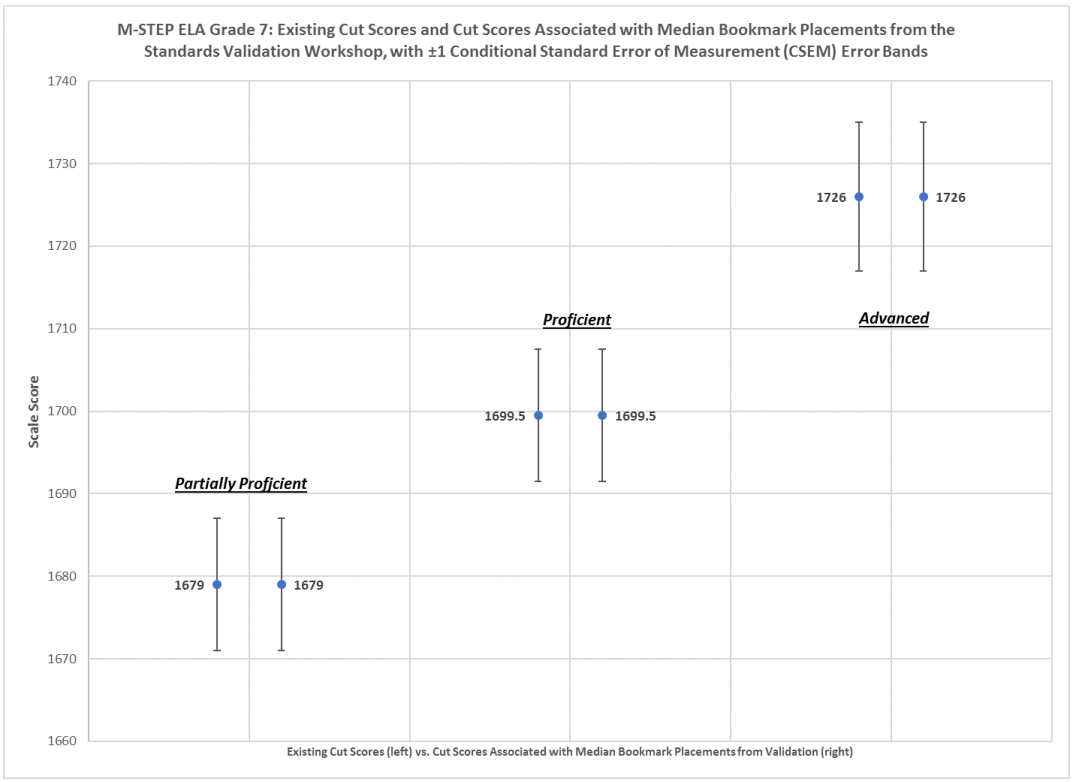
Figure E-1. ELA Comparison of Median Cut Score Recommendations and Existing Cut Scores, with Differences Expressed in Terms of Conditional Standard Error of Measurement (CSEM)






M-STEP ELA Grade 5: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1540
1530
1520
1510
1500
1490
1480
1470
Partially Proficient
Proficient
Advanced
1481
1481
1499.5
1499.5
1524
1521
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)


M-STEP ELA Grade 6: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1640
1630
1620
1610
1600
1590
1580
1570
1560
Partially Proficient
Proficient
Advanced
1578
1578
1599.5
1599.5
1624
1624
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)


M-STEP ELA Grade 7: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1740
1730
1720
1710
1700
1690
1680
1670
1660
Partially Proficient
Proficient
Advanced
1679
1679
1699.5
1699.5
1726
1726
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)

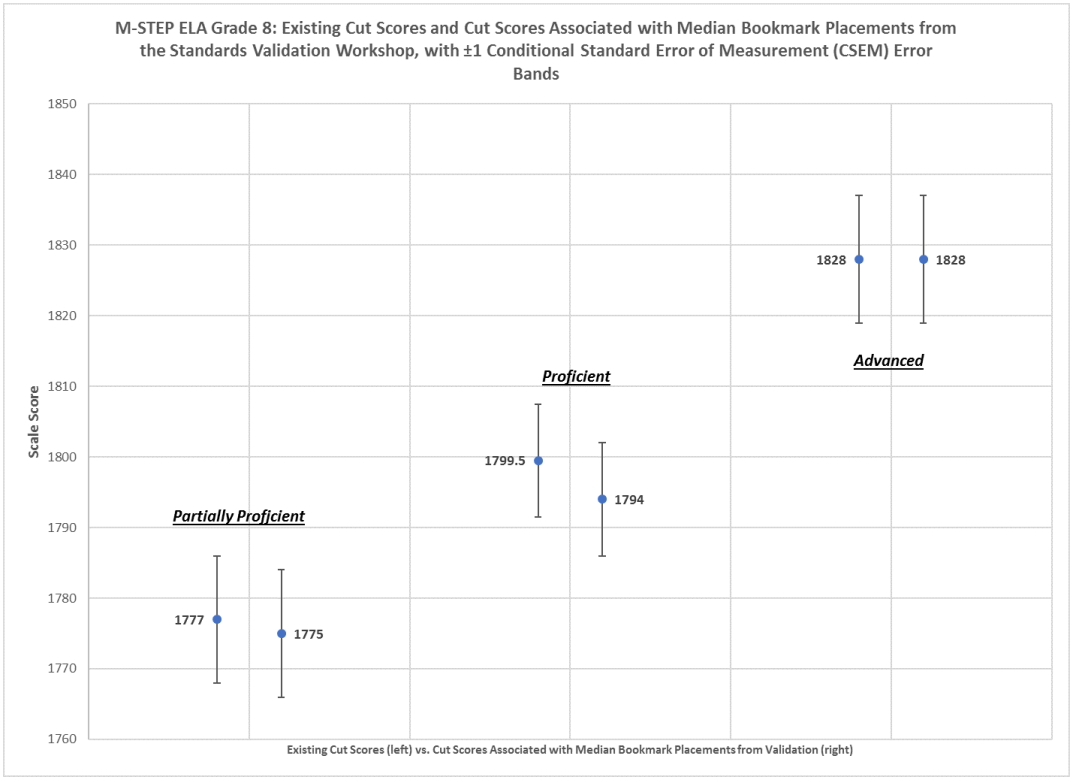
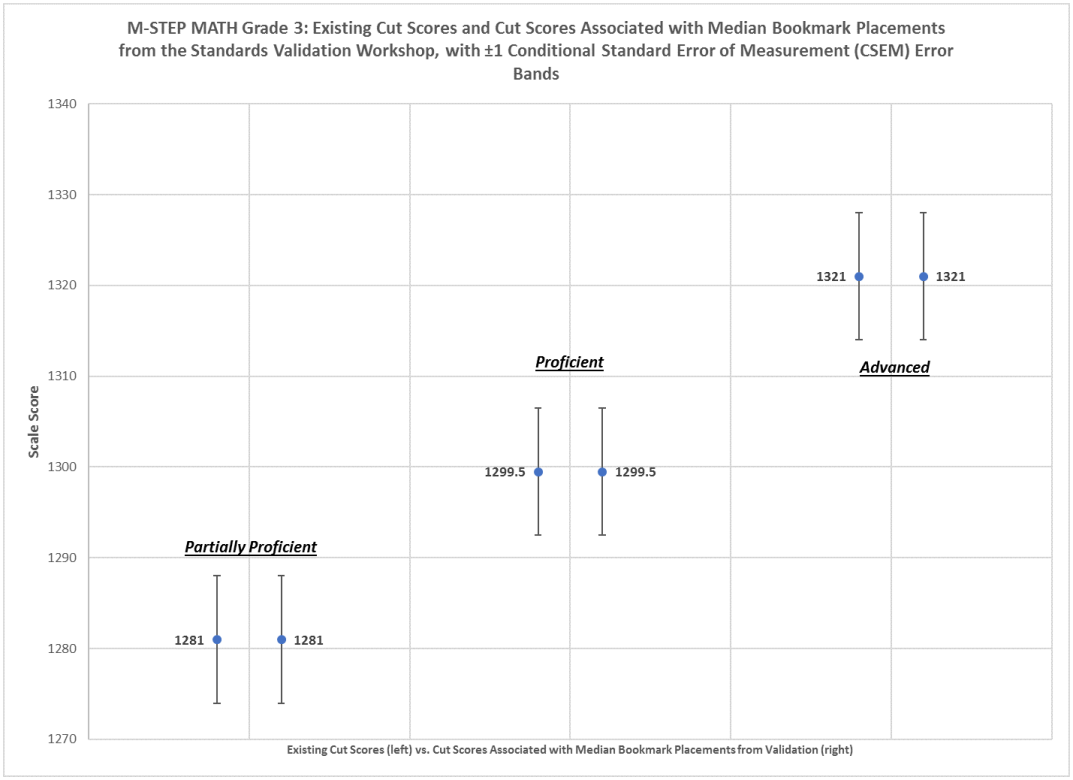
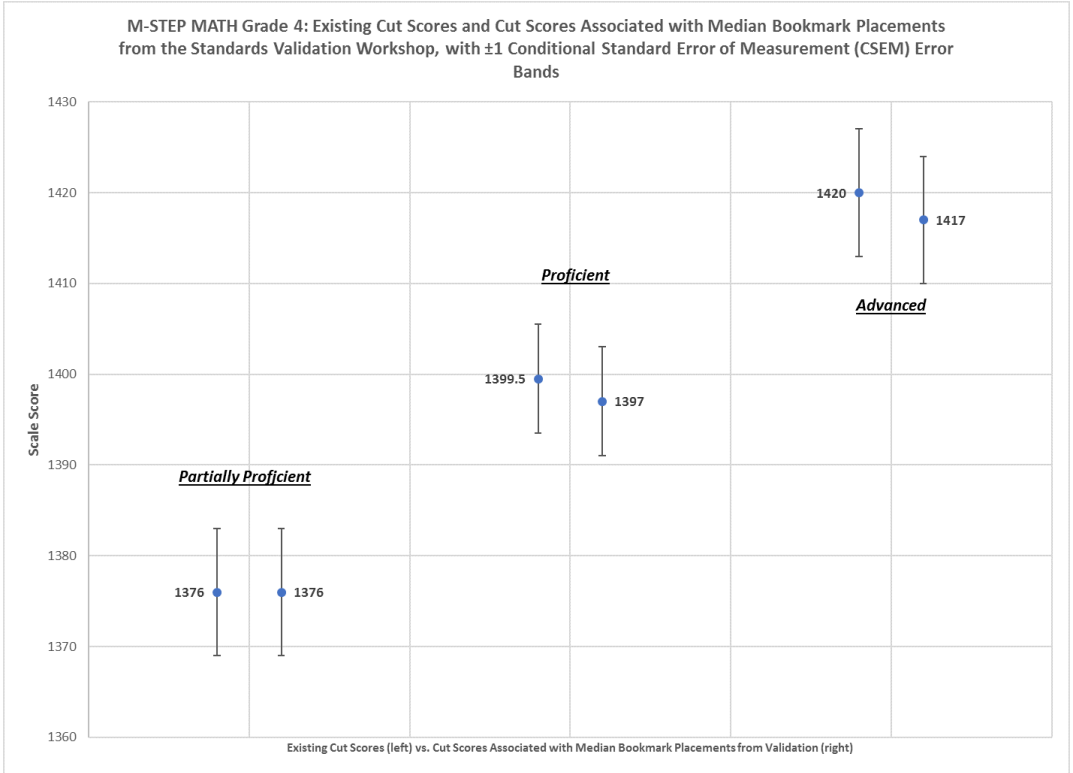
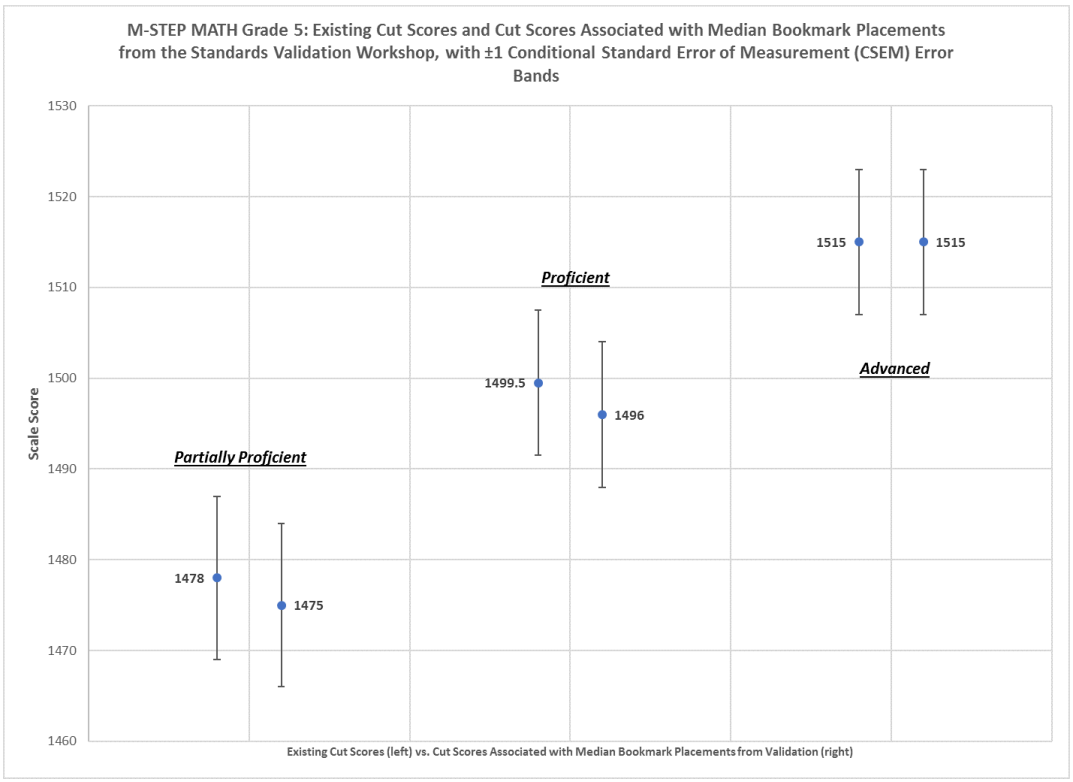
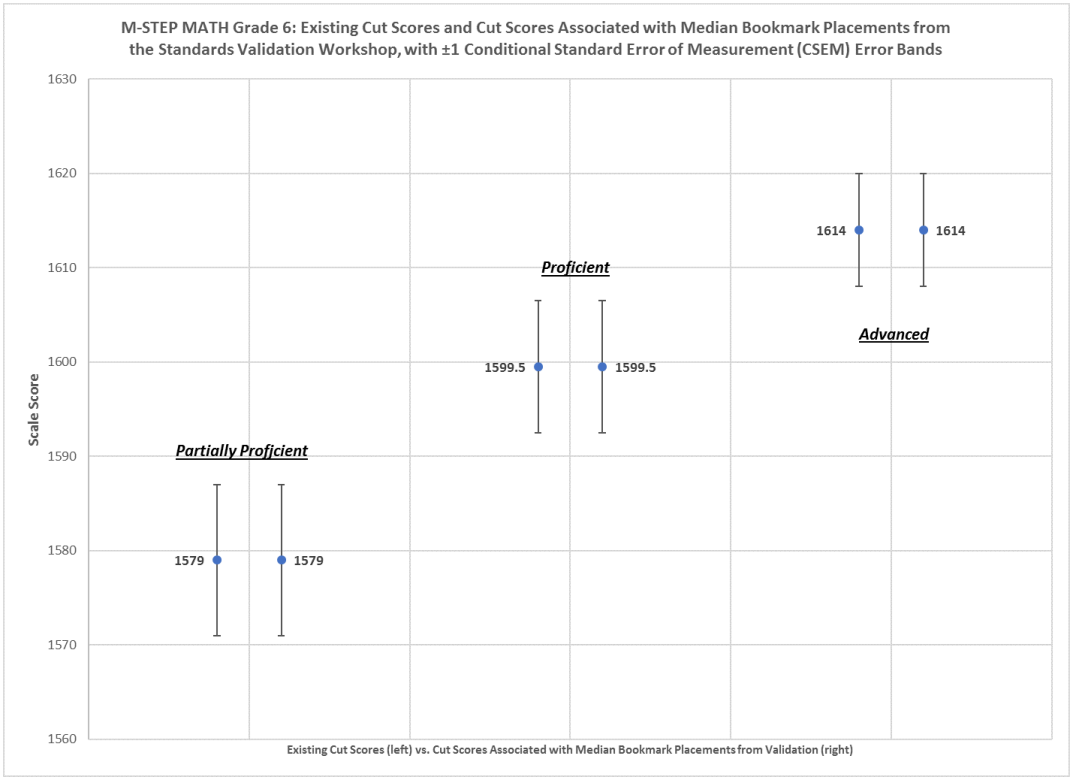

M-STEP ELA Grade 8: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1850
1840
1830
1820
1810
1800
1790
1780
1770
1760
Partially Proficient
Proficient
Advanced
1777
1775
1799.5
1794
1828
1828
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)

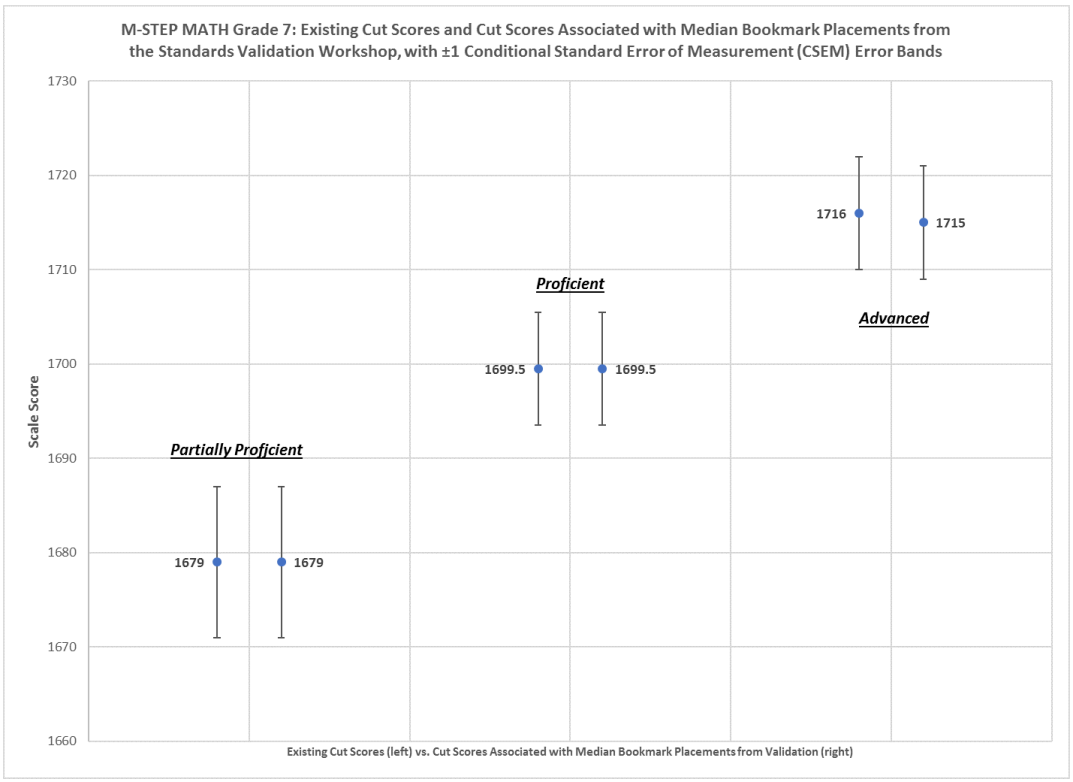
Figure E-2. Mathematics Comparison of Median Cut Score Recommendations and Existing Cut Scores, with Differences Expressed in Terms of Conditional Standard Error of Measurement (CSEM)

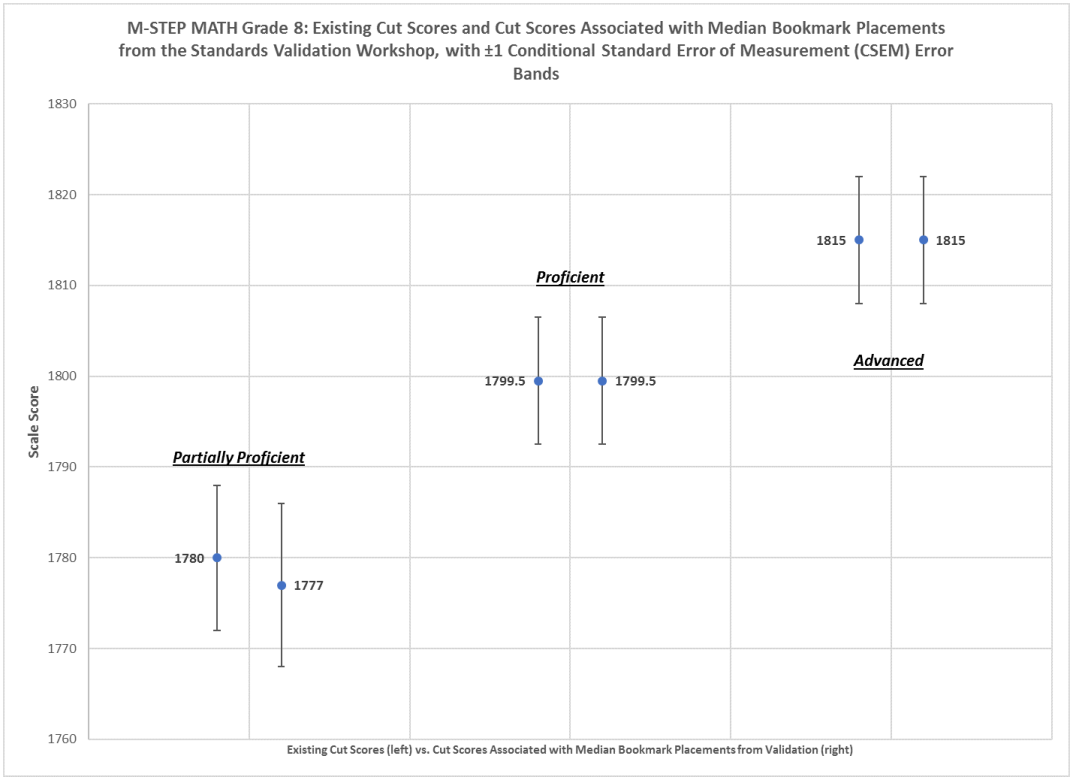





M-STEP MATH Grade 5: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1530
1520
1510
1500
1490
1480
1470
1460
Partially Proficient
Proficient
Advanced
1478
1475
1499.5
1496
1515
1515
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)


M-STEP MATH Grade 6: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1630
1620
1610
1600
1590
1580
1570
1560
Partially Proficient
Proficient
Advanced
1579
1579
1599.5
1599.5
1614
1614
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)


M-STEP MATH Grade 7: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1730
1720
1710
1700
1690
1680
1670
1660
Partially Proficient
1679
1679
Proficient
1699.5
1699.5
Advanced
1716
1715
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)


M-STEP MATH Grade 8: Existing Cut Scores and Cut Scores Associated with Median Bookmark Placements from the Standards Validation Workshop, with ±1 Conditional Standard Error of Measurement (CSEM) Error Bands
Scale Score
1830
1820
1810
1800
1790
1780
1770
1760
Partially Proficient
1780
1777
Proficient
1799.5
1799.5
Advanced
1815
1815
Existing Cut Scores (left) vs. Cut Scores Associated with Median Bookmark Placements from Validation (right)

The MDE considered the recommendations made by the standards validation committee and the existing cut scores. Working with DRC, MDE made three primary findings:

1. The content-based expectations for students in each performance level have not changed significantly since the cut scores were established. Although the tests are now shorter and passage-based writing items have been introduced on the ELA tests, the underlying expectations for students in each performance level have not changed.

2. The impact data observed in spring 2018 is similar to those from the 2017 administration of the tests when the existing cut scores were applied. This similarity supports the contention that the expectations for students in each performance level have not changed, and that the existing cut scores are valid for continued use.

3. The median cut score recommendations were all very close to the existing cut scores, to the point of being statistically indistinguishable. As shown in Table E-2, the average difference from the existing cut scores was -0.11 CSEM, and all were within a range of ±0.7 CSEM. Within this narrow range, it is difficult to argue that scale scores are significantly different.

The available validity evidence suggests that there were no significant differences between the updated ELA and mathematics assessments and the content assessed by the prior assessments; and that the differences between the judgments made at the 2018 standards validation workshop and the existing cut scores were not statistically different. That is, the recommendations made by Michigan educators during the standards validation were consistent with the existing cut scores, and the validity evidence collected during this process supports the continued use of the cut scores.

Table E-3 shows the existing cut scores and associated impact data for ELA and mathematics using spring 2018 administration data. Figures E-3 and E-4 show a graphical representation of the existing cut scores and their associated impact data from spring 2018.

Table E-3a. Existing ELA and Mathematics Cut Scores

Content	Grade	Partially Proficient	Proficient	Advanced
ELA	3	1280	1299.5	1317
ELA	4	1383	1399.5	1417
ELA	5	1481	1499.5	1524
ELA	6	1578	1599.5	1624
ELA	7	1679	1699.5	1726
ELA	8	1777	1799.5	1828
Math	3	1281	1299.5	1321
Math	4	1376	1399.5	1420
Math	5	1478	1499.5	1515
Math	6	1579	1599.5	1614
Math	7	1679	1699.5	1716
Math	8	1780	1799.5	1815

Table E-3b. Associated Impact Data for M-STEP Spring 2018

Content	Grade	Not Proficient	Partially Proficient	Proficient	Advanced
ELA	3	30.90%	24.60%	22.40%	22.10%
ELA	4	33.80%	21.10%	21.50%	23.60%
ELA	5	32.10%	21.20%	28.70%	17.90%
ELA	6	31.30%	27.30%	28.20%	13.20%
ELA	7	29.20%	27.30%	30.70%	12.80%
ELA	8	29.90%	27.30%	31.10%	11.80%
Math	3	27.80%	26.40%	27.20%	18.60%
Math	4	24.70%	33.20%	25.70%	16.40%
Math	5	37.00%	28.50%	17.80%	16.60%
Math	6	34.40%	30.90%	18.60%	16.00%
Math	7	36.20%	28.00%	19.50%	16.40%
Math	8	40.90%	26.30%	14.90%	17.80%

Figure E-3. Existing, Validated Cut Scores and Associated Impact Data for Spring 2018 ELA

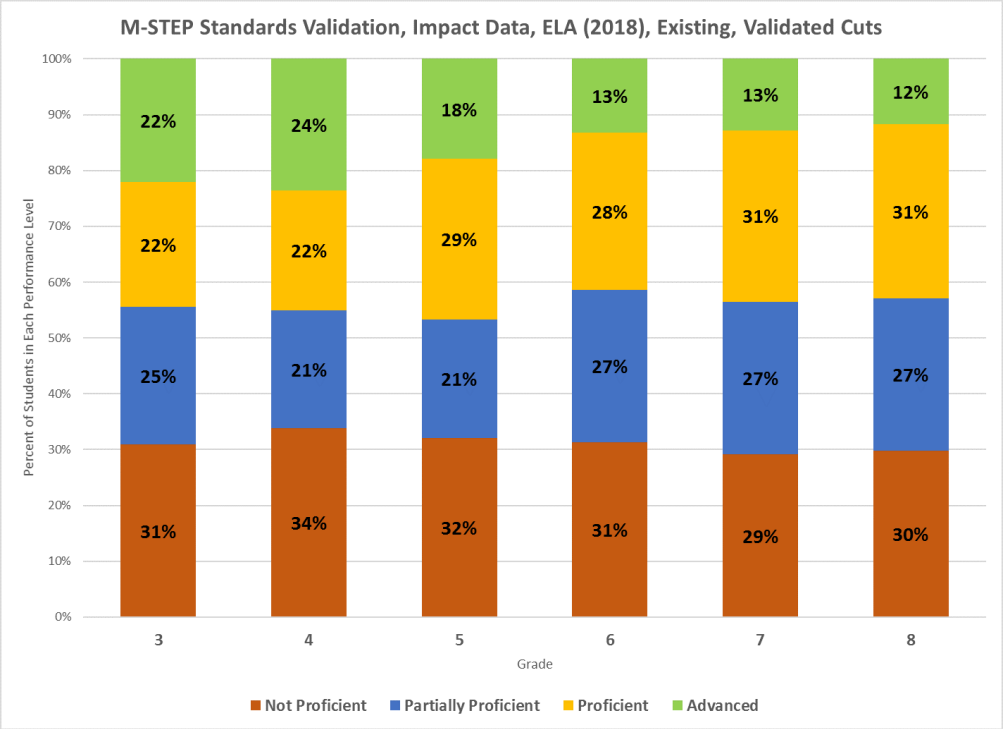
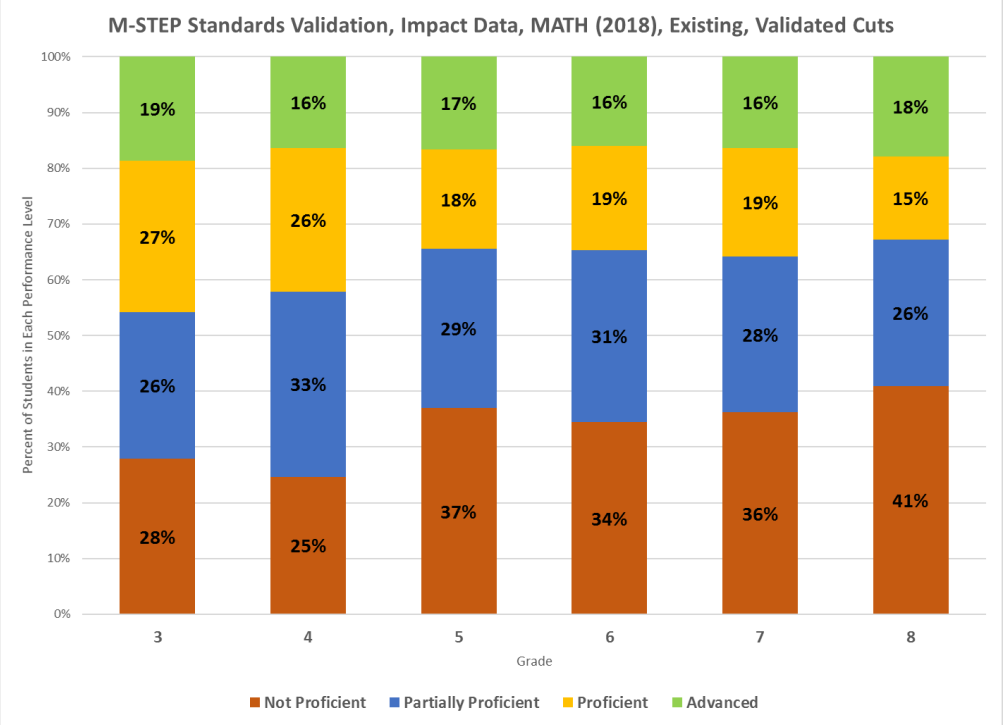


Figure E-4. Existing, Validated Cut Scores and Associated Impact Data for Spring 2018 Mathematics



Appendix E.6 References

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach. Symposium* presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment: Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225-253). New York: Routledge.

Appendix F: Test Mode Comparison

Appendix F:
Conversion Relation Study of Online and Paper-Pencil Administrations for M-STEP Social Studies

Overview

For 2019 M-STEP Social Studies, more than 99% of students statewide took the online forms. When conversion tables were created, no paper-pencil data were available, thus the online forms' conversion tables were applied to the paper-pencil form at each grade. However, whether such approach is appropriate needs to be examined. The current study thus aims at addressing the following question: Is it appropriate to apply the online forms' (will be referred to as CBT hereafter) conversion tables to the paper-pencil form (will be referred to as PPT hereafter) at each grade?

This appendix is organized around three major sections: Propensity Scores and Matched Samples, Comparability Analyses and Results, and Discussion and Conclusion. For the propensity score matching and mode comparison reported here, we follow the procedures listed in Zeng, Yin, and Shedden (2015) with some modifications to better address the question here.

Section 1: Propensity Scores and Matched Samples

This section describes how matched samples are formed. Specifically, the concept of propensity score was introduced, along with the description of propensity score matching procedures.

1.1 Propensity Scores

A propensity score is the conditional probability of assignment to treatment (in current report, take PPT instead of CBT) given the covariates, and it does not depend on the response information (Rosenbaum & Rubin, 1985). When the treatment variable is dichotomous, logistic regressions with the treatment assignment as an outcome are used to estimate the propensity scores (e.g., Harder, Stuart, & Anthony, 2010). In this report, the PPT was coded as 1 and the CBT was coded as 0. After propensity scores are estimated, different approaches such as matching, weighting, and subclassification can be applied to form comparable groups (Harder et al., 2010; Stuart, 2010).

In this report, we only considered pair matching in forming comparable groups with the same sample size. Five issues need to be considered when conducting propensity score matching (Zeng et al., 2015): (1) choice of covariates, (2) dealing with missing data on the covariates, (3) matching methods, (4) assessing the matching quality, and (5) the possible violations of ignorable treatment assignment. All these issues are discussed in section 1.2 below, and the fifth one is also tackled with in the Discussion and Conclusion section.

1.2 Propensity Score Matching Procedures

This section provides detailed information on the five issues mentioned above when conducting propensity score matching.

(1) Choice of Covariates

Three types of covariates can be included in a propensity score estimation model:

- a variable related to both the outcome and the treatment,
- a variable related to only the outcome, and
- a variable related only to the treatment.

Simulation studies found that the optimal propensity score model was the one only included the first two types of variables mentioned above (i.e., a variable related to both the outcome and the treatment, and a variable related to only the outcome) (Brookhart et al., 2006). Steiner et al. (2010) found that the first type mentioned above (i.e., a variable related to both the outcome and the treatment) was crucial for removing bias. Pre-test measures were found to be highly correlated to potential outcomes (Steiner et al., 2010), and were therefore suggested to be included as covariates for estimating propensity scores. Socioeconomic status (SES) is a student-level variable that is possibly relevant to any score differences across mode (Pomplun, Ritchie, & Custer, 2006). Way, Lin, & Kong (2008) used all possible prior achievement scores in their mode comparability study.

To fully utilize the capability of propensity scores in balancing multiple covariates, we included all possible prior year achievement scores and current year achievement scores (excluding the subject area under examination). In addition, we also included all available demographic variables at the student level: female (1 for female, 0 for male), White (1 for White, 0 otherwise), Black (1 for Black, 0 otherwise), Hispanic (1 for Hispanic, 0 otherwise), Asian (1 for Asian, 0 otherwise), Economically Disadvantaged (ED, 1 if yes, 0 if no), Special Education (SE, 1 if yes, 0 if no), and English Language learner (EL, 1 if yes, 0 if no).

Based on the data, we observed that some school buildings (will be referred to as school or schools hereafter) had the corresponding grade level participate via one administration mode only, but some schools had the corresponding grade level participate via both administration modes. Since we used student level data for calibration and equating, we did not consider school level variables in propensity score model building. However, school level variables were used for data imputation. Related details can be found below where the missing data issue is discussed.

(2) Dealing with missing data on the covariates

Since we do not want to exclude any PPT students[1] from this investigation, while quite a few of them were found to have missing data on previous achievement or even some current year achievement, we had to use imputation. Although various missing data handling techniques have been proposed in the context of propensity score estimation, no significant differences in treatment effect estimations were found between various techniques applied to real data sets (Harder et al., 2010). For this report, a multiple imputation procedure was carried out using the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011), which conducts multivariate imputation by chained equations. However, instead of using the multiple imputed values, we only used one set of imputed values to simplify the analysis, thus in essence a single imputation approach.

As mentioned above, the school level data were not used in propensity score model building, but only used in imputation. The school level variables being used are: N-count per school for the grade level under consideration, percent female, percent White, percent Black, percent Hispanic, percent Asian, percent ED, percent SE, percent EL, average scale score in past two years, and the average percent proficient in Social Studies in past two years. Note that the teacher information was not considered at the school level as we did not have student-teacher nesting information for the school year of 2017-2018.

Since we had both the school level and student level variables, we did the imputation in two steps. The first step imputation was conducted at the school level, and the imputed school level data were combined with student level variables to conduct the second step imputations at the student level. The imputed student level data with only student level variables were then used to build propensity score models and to form matched samples. The N-counts for PPT and CBT students used in propensity score models per grade can be found in Table 1.1. Note that the N-counts for CBT in Table 1.1. is not for all CBT students, but for the most similar CBT form (form 2 at each grade, see footnote 2 for more details) for Social Studies. These are the real CBT pool used for propensity score building and pair matching.

[1] Note that duplicated ID records were excluded. Only 1 or 2 (at maximum) such cases were found and excluded per grade.

Table 1.1. N-Counts for CBT and PPT per Grade for Social Studies[2]

Grade	PPT	CBT
5	936	36,069
8	653	36,598
11	728	35,062

(3) Matching Methods

Different matching methods exist in the literature, such as the nearest neighbor matching approach (Stuart, 2010). The optimal matching algorithm is found to be better than the nearest neighbor matching approach (or the greedy algorithm) for pair matching with a large pool of controls (Hansen, 2004). As shown in Table 1.1, there is a huge pool of CBT students in comparison to the PPT students at each grade. Therefore, the optimal matching algorithm was used. Specifically, the R package OPTMATCH (Hansen & Klopfer, 2006) was used to conduct pair matching based on the logit estimated propensity scores per grade.

(4) Assessing the Matching Quality

When judging matching quality, we examined individual covariate (at student level) using either a Chi-square test or a *t*-test, in addition to the overall balance test reported in the R package RItools (Bowers, Fredrickson, & Hansen, 2010). The overall balance check used in RItools tests balance on all linear combinations of the covariates in the propensity score model (Hansen & Bowers, 2008). We observed p-values close to 1 for the overall balance check across all grade levels. We also found that for all grade levels, no individual level covariate had significant difference between the two modes at $\alpha = 0.05$ level.

(5) Considering Possible Violation of Ignorable Treatment Assignment

As stated by Rosenbaum and Rubin (1985), propensity score approaches cannot balance unobserved variables. Therefore, if an unobserved variable is significantly related to both the treatment assignment and the outcome but is unmeasured and thus is not included in the propensity score estimation model, the resulting treatment effect estimates would be biased (Stuart, 2010). Different sensitivity analysis approaches have been proposed in the literature (Caliendo & Kopeinig, 2008; Stuart, 2010). According to Rosenbaum (2010), such analyses are conducted by altering the chances of receiving treatment for those units that appear to have similar chances. The examples included in Rosenbaum (2010) all indicate a significant effect being found, and the sensitivity analyses are conducted trying to specify when such an effect

[2] We kept all PPT students, but only used CBT students from form 2 at each grade. Note that for Social Studies, all CBT forms share the same operational items. However, due to too large a sample size of CBT students during the matching step which exhausts the computer memory, we used the CBT form 2 at each grade since this form shares most operational and field-test items with the paper-pencil form per grade. Moreover, duplicated records were excluded for both CBT and PPT forms.

becomes non-significant statistically. In our case, however, we hoped to conclude that the two modes are comparable (i.e., that no significant differences can be found between the two modes). Therefore, if a sensitivity analysis were conducted, the direction would be the opposite (i.e., trying to specify when the two modes would show statistically significant differences). We considered such analysis unnecessary here, as we reported out the online form conversion tables for the PPT students at each grade. Therefore, the worst scenario (with regard to incomparable samples) in our case would be the reported conversion relations. We further discuss this in the Discussion and Conclusion section.

Section 2: Comparability Analyses and Results

This section describes the methods and results for comparability analyses based on the matched samples obtained per grade.

Three sets of comparison analyses were conducted and all of them focused on the overall test. First, a multigroup confirmatory factor analysis (MGCFA) using MPLUS (Muthén & Muthén, 2019) was conducted. Second, summed score to Expected *A Posteriori* (EAP) conversion tables from separate calibrations (with the fixed item parameter calibration approach) using flexMIRT (Cai, 2017) were compared. Third, proficient classification of PPT students from different conversion tables were compared. Details of the three sets of analyses and corresponding results are presented below.

2.1 Multigroup Confirmatory Factor Analysis (MGCFA) using MPLUS

For this analysis, three nested models were compared to establish measurement invariance (Schroeders & Wihelm, 2011): configural invariance, strong invariance, and strict invariance. For configural invariance, factor loadings and thresholds are freely estimated but the residual variances are fixed at 1 and factor means are fixed at 0 in both groups. For strong invariance, the factor loadings and thresholds are fixed to be equal across the two groups, the residual variances are fixed at 1 for the CBT group but are freed in the PPT group, and the factor means are fixed at 0 in the CBT group but are freed in the PPT group. The only difference between the strict invariance model and the strong invariance model is that the former also fixes the residual variances at 1 in both groups (see Table 2.1 below, which is adapted from Schroeders & Wihelm, 2011).

Table 2.1. Testing for Measurement Invariance with Categorical Data

Invariance Type	Factor Loadings	Thresholds	Residual Variances	Factor Means
Configural invariance	(*	*)	Fixed at 1	Fixed at 0
Strong invariance	(Fixed	Fixed)	Fixed at 1/*	Fixed at 0/*
Strict invariance	(Fixed	Fixed)	Fixed at 1	Fixed at 0/*

Note. From Schroeders & Wihelm (2011). The asterisk (*) indicates that the parameter is freely estimated. Fixed = the parameter dominated in the title of the column is fixed to equity across groups; Fixed at 1 = the residual variances are fixed at 1 in both groups; Fixed at 1/* = the residual variance is fixed at 1 in one group whereas freed in the other group; Fixed at 0 = factor means are fixed at 0 in both groups. Fixed at 0/* = the factor mean is fixed at 0 in one group and freed in the other. Parameters in parentheses need to be varied in tandem.

If strict invariance is established, the observed scores can be considered as interchangeable (Neuman & Baydoun, 1998). If, however, only the strong invariance is established, ability estimates can be considered as comparable when residual item variances can be attributed to random error (Schroeders & Wilhelm, 2011). Same as in Schroeders & Wihelm (2011), here we estimated all models using the default estimator—weighted least squares mean and variance adjusted (WLSMV) estimator with Theta parameterization. Due to problems found with the Chi-square (χ^2) statistics (Chen, 2007; Cheung & Rensvold, 2002), the following fit indices and cutoff criteria were used: the comparative fit index (CFI) ≥ 0.95 and the root mean square error of approximation (RMSEA) < 0.05 (Hu & Bentler, 1998) for indicating a good model fit; and a change of ≥ -0.010 in CFI and a change of ≥ 0.015 in RMSEA (Chen, 2007) for indicating noninvariance for each step of the nested model comparison.

Table 2.2.a. Testing for Measurement Invariance for Social Studies Grade 5 (1-Factor Model)

Invariance Type	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural invariance	0.982	0.010		
Strong invariance	0.966	0.014	-0.016	0.004
Strict invariance	0.963	0.014	-0.003	0.000

Table 2.2.b. Testing for Measurement Invariance for Social Studies Grade 8 (1-Factor Model)

Invariance Type	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural invariance	0.969	0.016		
Strong invariance	0.963	0.018	-0.006	0.002
Strict invariance	0.964	0.017	0.001	-0.001

Table 2.2.c. Testing for Measurement Invariance for Social Studies Grade 11 (1-Factor Model)

Invariance Type	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural invariance	0.989	0.016		
Strong invariance	0.986	0.018	-0.003	0.002
Strict invariance	0.976	0.023	-0.010	0.005

As shown in Tables 2.2.a.—2.2.c., all three invariance models fit for all grade levels when using the fit indices of CFI and RMSEA. When using the change of CFI in combination with the change of RMSEA, results are not so clear. When using the change in CFI as the main criterion (as recommended by Chen [2007]), however, only the configural invariance holds for grade 8 and grade 11, but strong invariance holds for grade 5.

2.2 Conversion Table Comparison

Based on the above findings, we decided to take a conservative approach: assuming that only configural invariance holds. We thus did separate calibrations for the matched samples to compare their conversion relations. Tables 2.3.a.—2.3.c. present the results for separate calibrations. In addition, the reported conversion tables for PPT forms (i.e., the conversion tables created for the online forms) for each grade are also included in these tables, as our focus here is to address if it is appropriate to apply the conversion tables established for the online forms to the PPT form at each grade.

Based on the separate calibration results, we did two comparisons: (1) between PPT and matched CBT, and (2) between PPT and the reported results. All raw to scale score conversion relations can be found in Tables 2.3.a. to 2.3.c., and the maximum absolute differences between the PPT and the matched CBT, as well as those between the PPT and the reported conversion relations can be found in Table 2.4. As shown in Table 2.4, the maximum absolute difference between the PPT and the matched CBT is smaller than the smallest SE found in both calibrations. Same conclusions can be made when the PPT calibration results are compared to the reported results. Test Characteristic Curves comparisons are skipped here as the same information is contained in Tables 2.3.a. to 2.3.c..

Table 2.3.a. Conversion Tables for Social Studies Grade 5 Matched Samples

RawScore	PPT_EAP	PPT_SE	CBT_EAP	CBT_SE	Reported_EAP	Reported_SE
0	-2.955	0.511	-3.100	0.488	-3.077	0.484
1	-2.760	0.517	-2.931	0.503	-2.899	0.496
2	-2.563	0.511	-2.754	0.506	-2.714	0.495
3	-2.371	0.498	-2.575	0.500	-2.530	0.485
4	-2.188	0.482	-2.400	0.488	-2.352	0.470
5	-2.015	0.466	-2.231	0.475	-2.183	0.454
6	-1.851	0.452	-2.070	0.461	-2.023	0.439
7	-1.696	0.439	-1.916	0.447	-1.872	0.426
8	-1.549	0.427	-1.768	0.436	-1.727	0.414
9	-1.408	0.417	-1.626	0.425	-1.589	0.404
10	-1.274	0.409	-1.490	0.416	-1.457	0.395
11	-1.144	0.401	-1.358	0.409	-1.330	0.388
12	-1.019	0.395	-1.231	0.402	-1.207	0.382
13	-0.897	0.390	-1.107	0.396	-1.088	0.376
14	-0.778	0.385	-0.986	0.392	-0.973	0.372
15	-0.662	0.382	-0.868	0.388	-0.859	0.368
16	-0.548	0.379	-0.752	0.385	-0.749	0.365
17	-0.436	0.377	-0.638	0.383	-0.640	0.363
18	-0.325	0.376	-0.525	0.381	-0.532	0.361
19	-0.216	0.375	-0.414	0.380	-0.426	0.360
20	-0.107	0.374	-0.303	0.380	-0.320	0.359
21	0.002	0.375	-0.192	0.380	-0.215	0.359
22	0.111	0.376	-0.082	0.381	-0.110	0.359
23	0.219	0.377	0.028	0.382	-0.005	0.360
24	0.329	0.379	0.139	0.384	0.101	0.362
25	0.439	0.382	0.250	0.386	0.207	0.364
26	0.550	0.385	0.363	0.388	0.314	0.366
27	0.663	0.389	0.476	0.392	0.422	0.369
28	0.777	0.393	0.591	0.396	0.532	0.372
29	0.893	0.398	0.708	0.400	0.643	0.377
30	1.012	0.403	0.827	0.405	0.757	0.381
31	1.133	0.410	0.949	0.411	0.873	0.387
32	1.257	0.417	1.074	0.417	0.993	0.393
33	1.385	0.425	1.202	0.424	1.115	0.400
34	1.517	0.434	1.333	0.432	1.242	0.408
35	1.653	0.443	1.469	0.441	1.373	0.417
36	1.794	0.454	1.610	0.450	1.510	0.427
37	1.940	0.465	1.756	0.461	1.652	0.438
38	2.093	0.478	1.909	0.472	1.801	0.450
39	2.251	0.490	2.068	0.485	1.958	0.463
40	2.414	0.502	2.234	0.498	2.123	0.478

RawScore	PPT_EAP	PPT_SE	CBT_EAP	CBT_SE	Reported_EAP	Reported_SE
41	2.582	0.511	2.408	0.510	2.298	0.493
42	2.750	0.514	2.587	0.519	2.481	0.507
43	2.914	0.510	2.768	0.521	2.671	0.516
44	3.069	0.496	2.947	0.514	2.862	0.515
45	3.208	0.473	3.115	0.495	3.047	0.502

Note. PPT_EAP is the EAP theta from the separate PPT calibration, CBT_EAP is the EAP theta from the separate matched CBT calibration, and Reported_EAP is the EAP theta from the most similar CBT form applied to the PPT students for reporting.

Table 2.3.b. Conversion Tables for Social Studies Grade 8 Matched Samples

RawScore	PPT_EAP	PPT_SE	CBT_EAP	CBT_SE	Reported_EAP	Reported_SE
0	-2.815	0.535	-2.865	0.531	-2.899	0.511
1	-2.629	0.537	-2.675	0.533	-2.699	0.511
2	-2.438	0.527	-2.484	0.524	-2.501	0.499
3	-2.249	0.511	-2.297	0.508	-2.310	0.480
4	-2.068	0.492	-2.120	0.488	-2.132	0.459
5	-1.897	0.472	-1.952	0.468	-1.965	0.439
6	-1.734	0.454	-1.793	0.448	-1.809	0.421
7	-1.581	0.437	-1.643	0.430	-1.663	0.404
8	-1.436	0.422	-1.502	0.414	-1.525	0.390
9	-1.298	0.408	-1.367	0.400	-1.395	0.378
10	-1.167	0.397	-1.238	0.387	-1.270	0.367
11	-1.041	0.387	-1.114	0.376	-1.151	0.358
12	-0.920	0.378	-0.995	0.366	-1.037	0.350
13	-0.803	0.371	-0.879	0.358	-0.926	0.343
14	-0.689	0.365	-0.767	0.352	-0.819	0.338
15	-0.578	0.360	-0.658	0.347	-0.714	0.334
16	-0.470	0.357	-0.552	0.343	-0.612	0.331
17	-0.363	0.354	-0.447	0.341	-0.511	0.329
18	-0.258	0.352	-0.344	0.339	-0.411	0.328
19	-0.153	0.352	-0.242	0.339	-0.312	0.328
20	-0.049	0.352	-0.140	0.340	-0.214	0.328
21	0.054	0.353	-0.039	0.342	-0.116	0.329
22	0.158	0.355	0.063	0.344	-0.017	0.331
23	0.263	0.357	0.166	0.347	0.082	0.333
24	0.368	0.361	0.270	0.351	0.182	0.337
25	0.475	0.365	0.375	0.356	0.283	0.340
26	0.583	0.370	0.482	0.362	0.386	0.345
27	0.694	0.376	0.591	0.368	0.491	0.350
28	0.806	0.382	0.703	0.375	0.598	0.356
29	0.922	0.390	0.817	0.383	0.708	0.363
30	1.040	0.398	0.935	0.391	0.821	0.370
31	1.163	0.407	1.057	0.401	0.938	0.378
32	1.289	0.417	1.182	0.411	1.059	0.387
33	1.419	0.427	1.312	0.422	1.184	0.397
34	1.554	0.439	1.446	0.434	1.314	0.408
35	1.694	0.451	1.586	0.446	1.450	0.420
36	1.840	0.465	1.732	0.460	1.592	0.434
37	1.991	0.479	1.883	0.475	1.741	0.448
38	2.148	0.493	2.042	0.490	1.898	0.464
39	2.311	0.507	2.207	0.505	2.064	0.480
40	2.479	0.519	2.379	0.518	2.239	0.497

RawScore	PPT_EAP	PPT_SE	CBT_EAP	CBT_SE	Reported_EAP	Reported_SE
41	2.648	0.525	2.555	0.528	2.423	0.513
42	2.816	0.524	2.733	0.532	2.614	0.523
43	2.976	0.515	2.906	0.526	2.806	0.525
44	3.123	0.496	3.070	0.509	2.992	0.515

Note. PPT_EAP is the EAP theta from the separate PPT calibration, CBT_EAP is the EAP theta from the separate matched CBT calibration, and Reported_EAP is the EAP theta from the most similar CBT form applied to the PPT students for reporting.

Table 2.3.c. Conversion Tables for Social Studies Grade 11 Matched Samples

RawScore	PPT_EAP	PPT_SE	CBT_EAP	CBT_SE	Reported_EAP	Reported_SE
0	-2.730	0.510	-2.669	0.519	-2.746	0.509
1	-2.496	0.493	-2.439	0.502	-2.518	0.494
2	-2.277	0.466	-2.219	0.474	-2.299	0.469
3	-2.078	0.437	-2.016	0.445	-2.098	0.441
4	-1.897	0.411	-1.831	0.417	-1.914	0.415
5	-1.733	0.388	-1.663	0.393	-1.747	0.393
6	-1.582	0.369	-1.509	0.372	-1.594	0.374
7	-1.443	0.353	-1.367	0.354	-1.452	0.358
8	-1.313	0.339	-1.236	0.339	-1.320	0.344
9	-1.191	0.328	-1.113	0.327	-1.196	0.333
10	-1.075	0.319	-0.997	0.317	-1.078	0.324
11	-0.964	0.312	-0.886	0.308	-0.965	0.316
12	-0.858	0.306	-0.780	0.301	-0.857	0.310
13	-0.754	0.302	-0.678	0.296	-0.753	0.305
14	-0.653	0.299	-0.578	0.292	-0.651	0.302
15	-0.554	0.297	-0.481	0.289	-0.551	0.299
16	-0.456	0.296	-0.385	0.287	-0.452	0.297
17	-0.359	0.296	-0.290	0.286	-0.354	0.296
18	-0.262	0.297	-0.195	0.286	-0.257	0.296
19	-0.164	0.299	-0.101	0.287	-0.160	0.297
20	-0.065	0.302	-0.005	0.290	-0.062	0.299
21	0.034	0.306	0.091	0.293	0.037	0.302
22	0.136	0.311	0.189	0.297	0.138	0.305
23	0.240	0.316	0.290	0.302	0.240	0.310
24	0.347	0.323	0.393	0.308	0.346	0.315
25	0.458	0.330	0.499	0.315	0.454	0.322
26	0.572	0.339	0.610	0.324	0.567	0.329
27	0.692	0.349	0.726	0.333	0.684	0.339
28	0.817	0.360	0.847	0.345	0.806	0.349
29	0.949	0.373	0.975	0.357	0.936	0.361
30	1.088	0.388	1.111	0.372	1.073	0.375
31	1.237	0.404	1.255	0.388	1.218	0.391
32	1.395	0.423	1.411	0.407	1.375	0.409
33	1.566	0.444	1.578	0.428	1.543	0.430
34	1.750	0.467	1.759	0.451	1.725	0.453
35	1.948	0.492	1.957	0.477	1.924	0.479
36	2.163	0.519	2.172	0.504	2.140	0.506
37	2.392	0.542	2.405	0.529	2.375	0.531
38	2.628	0.554	2.647	0.543	2.622	0.547

Note. PPT_EAP is the EAP theta from the separate PPT calibration, CBT_EAP is the EAP theta from the separate matched CBT calibration, and Reported_EAP is the EAP theta from the most similar CBT form applied to the PPT students for reporting.

Table 2.4. Comparison of PPT Calibrations from Matched CBT Calibrations and the Reported

Subject	Grade	Maximum Absolute Difference		Minimum SE		
		CBT vs. PPT	Reported vs. PPT	PPT	CBT	Reported
Social Studies	5	0.220	0.293	0.374	0.380	0.359
	8	0.108	0.250	0.352	0.339	0.328
	11	0.078	0.025	0.296	0.286	0.296

Note. CBT here indicate matched CBT data.

2.2 Comparison of Cut Scores and Proficiency Classification

We also compared classification results. The yellow cells in Tables 2.3.a. to 2.3.c. are the minimum theta values at or above the threshold for each performance levels. Among the three cuts, i.e., Not Proficient vs. Partially Proficient, Partially Proficient vs. Proficient, Proficient vs. Advanced, we care most about the Partially Proficient vs. Proficient cut. We found that for Social Studies grade 11, the raw score point associated with Partially Proficient vs. Proficient cut is the same for all three conversion relations, and there is a two raw score points' difference for grade 5 and grade 8. Since only PPT students would be affected if they were reported based on the separate calibration results, only PPT students were examined for possible impact for different classification with regard to Partially Proficient vs. Proficient. Table 2.5. reports the number of students who would be classified differently.

Table 2.5. Partially Proficient vs. Proficient Classification Impact for PPT Students

Subject	Grade	PPT Students Impacted	
		Number	% of All PPT
Social Studies	5	24	2.56
	8	28	4.29
	11	0	0

According to Table 2.5., a small portion of students (2.56% for grade 5 and 4.29% for grade 8) would be classified differently. Moreover, based on Tables 2.3.a. to 2.3.c., for grade 5 and grade 8, the reported conversion tables classified correspondingly impacted students (reported in Table 2.5.) as Partially Proficient, while the PPT only separate calibration would classify them as Proficient if the separate calibration for PPT students were used.

Section 3: Discussion and Conclusion

In this mode comparison study, we used propensity score matching to form a matched set from the most similar CBT form to each PPT form per grade. We have mentioned before that for a proper use of propensity score matching, we need to consider the possible violation of ignorable treatment assignment. We also mentioned that a sensitivity analysis would best address this

consideration. However, we decided to skip this analysis, as we are only considering if it is appropriate to apply the online forms' conversion tables to the PPT students at each grade. The reported relation is thus the possible worst scenario. Based on the conversion table comparison, we found similar conclusions between the PPT and the matched CBT, as well as between the PPT and the reported.

We checked the marginal reliabilities reported from the flexMIRT separate calibration results and found them to be similar across the two modes per grade. Specifically, for Social Studies at grade 11, with two decimal points, they are the same across the two modes, and there is only 0.01 difference between the modes on the reliabilities for Social Studies at grades 5 and 8. Such similar internal consistency level between the two modes is to be expected, based on the reported results from the MGCFA analysis mentioned in Section 2.1.

Some states reported t-test results, and we considered this inappropriate here. First, when sample size is large, t-test usually ends up with significant results. Second, t-test only compares the means, and at most also tests the equality of variance assumption. However, for mode comparison at test score level, a better way would be to conduct the Kolmogorov-Smirnov test as described in Zeng et al. (2015) to compare the equality of distributions. We skipped this analysis here, because our focus here is not on mode comparison per se, but rather to address the question if the application of online forms' conversion tables to the PPT form is appropriate, i.e., if non-significant conversion relation in statistical sense could be established between separate calibrations of the PPT and the matched CBT. In addition, we compared the separate PPT calibration results to the reported conversion tables used for PPT.

Based on the comparisons reported in Section 2 above, we concluded that to apply the Social Studies online forms' conversion tables to the PPT form at each grade is acceptable.

References

Bowers, J., Fredrickson, M., & Hansen, B. (2010). RItools: Randomization Inference Tools. R package version 0.1-11.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156.

Cai, L. (2017). flexMIRT® version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Hansen, B. B. & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2):219--236.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99(467), 609-618.*

Hansen, B.B. and Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*, 609-627.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-249.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Muthén, L. K., & Muthén, B. O. (2019). *Mplus* (Version 8.3) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71-83.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127-143.

Rosenbaum, P. R. (2010). *Observational studies* (2nd Ed.). New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Schroeders, U. & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849-869.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267.

Stuart, E. A. (2010). Matching methods for causal inference: *A review and a look forward. Statistical Science, 25*(1), 1-21.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67.

Way, W. D., Lin, C.-H., & Kong, J. (March, 2008). *Maintaining score equivalence as test transition online: Issues, approaches and trends*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Zeng, J., Yin, P., & Shedden, K. A., (2015). Does matching quality matter in mode comparison studies? *Educational and Psychological Measurement, 75*(6), 1045-1062.

Appendix G: Michigan Assessment System Participant Groups

This appendix provides more details on the stake holders and participants involved in the Michigan Assessment System.

Appendix G.1 Michigan Educators

Michigan educators (including classroom teachers from K–12 and higher education, curriculum specialists, and administrators) play a vital role in all phases of the test development process. Committees of Michigan educators review the test specifications and provide advice on the model or structure for assessing each content area. They also work to ensure that test content and question types align closely with best practices in classroom instruction.

Appendix G.2 Technical Advisory Committee

Michigan's Technical Advisory Committee (TAC) serves as an advisory body to MDE. The TAC provides recommendations on technical aspects of large-scale assessments, including item development, test construction, administration procedures, scoring and equating methodologies, and standard-setting workshops. The TAC also provides guidance on other technical matters, such as practices not already described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), and continues to provide advice and consultation on the implementation of new assessments and adherence to the federal requirements set forth by the Every Student Succeeds Act. Table G-1 can be referenced for TAC member information.

Table G-1. Technical Advisory Committee

Name	Position	Organization
Dr. Mark Reckase, Chair	Distinguished Professor of Measurement and Quantitative Methods (retired)	Michigan State University
Dr. Damian Betebenner	Senior Associate	National Center for the Improvement of Educational Assessment
Dr. Gregory J. Cizek	Distinguished Professor of Educational Measurement and Evaluation	University of North Carolina, Chapel Hill
Dr. George E. Engelhard, Jr.	Professor Emeritus of Educational Measurement and Policy	University of Georgia
Dr. Christine Carrino Gorowara	Interim Director	Delaware Center for Teacher Education, University of Delaware
Dr. Joseph Martineau	Senior Associate	National Center for the Improvement of Educational Assessment
Dr. Dave Treder	Coordinator of Research, Evaluation, and Assessment	Genesee Intermediate School District, Flint, Michigan

Appendix G.3 Michigan's Division of Educator, Student, and School Supports (DESSS) Advisory Committee

The DESSS Advisory Committee meets quarterly to provide input, ideas, expert advice, and/or recommendations to MDE and DESSS on matters related to assessment and accountability, professional preparation, educator evaluations, assessment policy, and related communications to the field. The committee also meets to keep its respective organizations abreast of changes to the above areas that will affect Michigan's schools and students. The committee comprises representatives from educational agencies, organizations, and representatives from both two-year and four-year colleges and universities across the state. Table G-2 shows the members of the DESSS Advisory Committee.

Table G-2. Division of Educator, Student, and School Supports Advisory Committee

Last Name	First Name	Organization
Anand	Johanna	Michigan Department of Education/Low Incidence Outreach
Arnswald	Jennifer	Michigan Science Teachers Association
Berry	Kathy	Michigan Council of Teachers of Mathematics
Clingman	Cindy	Michigan Reading Association
Cox	Mary	Michigan Council of Teachers of English
Czerwinski	Harvey	Michigan Education Research Association
Dewsbury-White	Kathryn	Michigan Assessment Consortium
DeYoung	Ann	Michigan Elementary and Middle School Principals Association
Flukes	Jonathan	Michigan Education Research Association
Gordon	Casey	MI Council of Teachers of English to Speakers of Other Languages
Greer	Doug	Oakland Area Intermediate School District
Kher	Neelam	Michigan State University
Koekkoek	Matthew	Michigan Association of Administrators of Special Education
Langdon	Thomas	Michigan Association of School Administrators
Mastie	Marge	Washtenaw Intermediate School District - Retired
McIntyre	Rebecca	Michigan Association of Administrators of Special Education
Miller	Kathy	Michigan School Facilitators Network
Trout	Kelly	Ingham Intermediate School District
Vespremi	Stacy	Michigan Association of State and Federal Programs Specialists
Vorenkamp	Ellen	Wayne Regional Educational Services Agency
Zdeb	Wendy	Michigan Association of Secondary School Principals
Substitutes		
McGoran	Holly	Michigan Science Teachers Association
Musial	Joe	Wayne Regional Educational Services Agency
Ripmaster	Colin	Michigan Association of Secondary School Principals
Taraskiewicz	Cindy	Wayne Regional Educational Services Agency