

Reporting aggregated data at the System, School and Classroom Level¹

Robert W. Lissitz and Ying Li

MARCES

University of Maryland

Introduction

To enhance the information available for improving instruction, Maryland State Department of Education (MSDE) desires an examination of the feasibility of reporting strengths and weaknesses for aggregated student data at the state, school system and school level. The analysis included possible reporting strategies at the indicator and or expectation level with appropriate measures of accuracy that provide educators with potentially diagnostic and prescriptive information that can be applied at the school. The research is focusing on the High School Assessment (HSA) and Maryland State Assessment (MSA), although it is important to recognize that they report different outcomes. The HSA test reports a scale score and a pass-fail determination. The MSA reports scale scores and three proficiency levels. The MSA assesses the state content standards in reading and mathematics.

The Maryland HSAs are end-of-course tests that a student takes as they complete the appropriate high school level course. All high school students in Maryland take four courses, and the corresponding High School Assessments. These courses currently include Algebra/Data Analysis, English 2, Biology and Government. All students receive a score for each test they take. Scores are also reported at an aggregated level for the state, school systems, and schools. Performance standards were set for HSA in the summer of 2003, so a pass or failure is also reported, in addition to a scale score for each test. Aggregated Scores are reported as a percentage of students who passed the assessment.

In the MDK12.ORG site, there are sections for analyzing results for both MSAs and HSAs at different aggregated levels (e.g. school, district and the state). In the MSA section, data are organized to answer the following questions: At which proficiency levels are our students performing within each group? How did students in that group

perform on specific content standards? How does performance compare across subjects in that group? Which group (school or district) with similar poverty bands or specified populations have been most successful? In the HSA section, data are organized to answer the following questions: How did the group (school, district or the state) perform on HSA in terms of percent of passing? How did students perform on specific content standards within that group? Which groups (schools or districts) with similar poverty bands or specified population are the most successful groups? The main difference on score reporting between HSA and MSA is at the proficiency levels. HSA has only pass or fail while MSA has three achievement levels.

This project involves developing score reports to satisfy concerns for different stakeholders such as classroom teachers, and district and state administrators. We are hoping to use a focus group approach by providing initially developed score reports to several groups of district administrators and teachers for their feedback. Then we will revise the forms based on this feedback. After some initial feedback, we have organized the score report forms by their potential audiences and these are presented in this report. Then several limitations of the reports are presented and briefly discussed. Finally, a reflection on the state assessment system is offered to facilitate targeted score reporting to serve the purpose outlined here.

Literature Review

We reviewed the literature and found that most of the work in this area concerns score reporting at the individual level. A very insightful paper is that by Goodman & Hambleton (2003) titled *Student Test Score Reports and interpretive guides*, which focuses on individual score reporting with a number of suggestions for presenting that information. A particularly helpful paper was by Joe M. Ryan (2003), based on the item mapping and test reporting strategies that he developed with South Carolina. We also reviewed the websites of several states' Departments of Education, including Ohio, California, South Carolina, Maine, and Virginia. Most of them release their state-wide assessment score reports on their WEB sites, organized by the need of the users. For example, you could choose to see the score report within a school or within a district, or

within the state easily on their sites by drilling down. However, they did not provide detailed interpretive information on the students' weakness and strengths that would help a teacher or a school administrator trying to raise their level of test performance. The work done by Massachusetts on their Massachusetts Comprehensive Assessment System (MCAS, 2007) reports was also especially useful. The sample HSA score reporting forms we developed for this project are primarily based on the work by Massachusetts and that by Joseph M. Ryan. Some additional guidance was provided by Thanos Patelis of the College Board and the work they have done.

Initial Review

The initial sample of score reporting forms were created and sent to a small group of assessment professionals in Maryland and to Thanos Patelis at the College Board. We particularly want to thank Dr. Von Secker from Montgomery, Dr. Alban from Howard, and Mr. Bruce Hislop from Prince Georges. The following summarizes the feedback that we obtained from this first group of reviewers. Specific sample presentations referenced in these remarks can be found later in the report.

1. Summary tables A and B are fine for sharing school level, system level and state level comparisons; they offer limited information about how we compare to others and little value in guiding instructional decisions.
2. County people appreciate graphs more than tables.
3. The most valuable reports would be C1 and C2, which present information on student performance, if the purpose of a report is to help with teaching and planning instruction.
4. Teachers understand the "percent correct" metric much better than scale scores.
5. Teachers want a complete item analysis of the test so that they can address very specific needs.
6. Teachers want the most targeted information they can get so they will know how their students are performing and can adjust their teaching accordingly.
7. Teachers want a report that brings together the content standards and the performance standards to provide data that is useful and usable to help teachers make instructional decisions.

8. It would be helpful for teachers to understand within each strand how students are performing on the various indicators and which are presenting particular challenges.
9. No other information can take the place of item-by-item information.
10. It would be helpful if we know for a certain item, the percent correct by the students who scored at basic, proficient and advanced, separately.
11. It would be helpful if a report would include item difficulties or scale values (some sort of narrative format would be best) with the public release items so teachers could understand how students were responding to certain kinds of items.

We noticed that some of the feedback shows an interest in strand level reporting. Reporting strand performance information can help teachers and instructional leaders pinpoint areas of students' strengths and weaknesses, without encouraging a teaching the test mentality. In other words, we want the teachers to be aware of the Maryland Curriculum and to focus on the ideas, but we do not want them to focus on specific test items. Based on the feedback from the initial focus group and a discussion with the state assessment professionals we revised some of the forms and added some new ones.

Reporting Scores for Target Audiences

Depending on the audience, different aspects of the information should be emphasized. The following are suggestions for providing the most relevant information divided into that which is aggregated above the student level (for school, system or state level administrators) and that which is at the student level (for teachers and school level administrators). The scores and numbers we are using in our score reports below are hypothetical for illustration purpose.

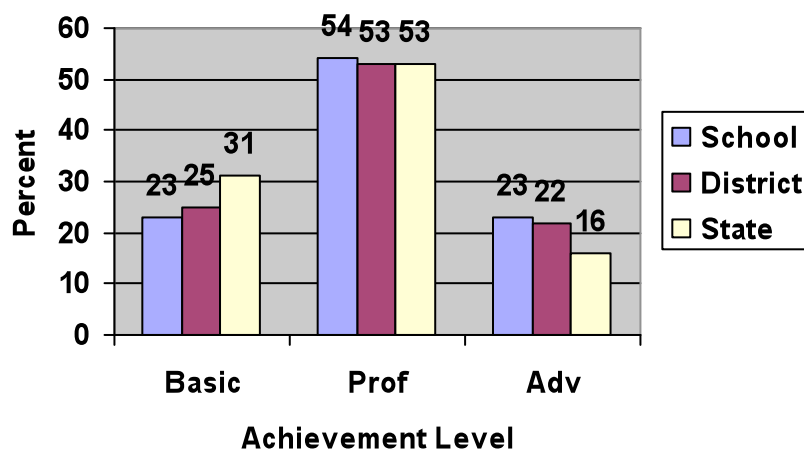
Section One: Score Reporting at Above Student Level Audience for administrators

Three kinds of information are presented and discussed in this section. They are 1) average performance of schools, districts and the state, 2) comparison with demographically, SES, or linguistically similar schools, and 3) three-year trend data for school, district and the state.

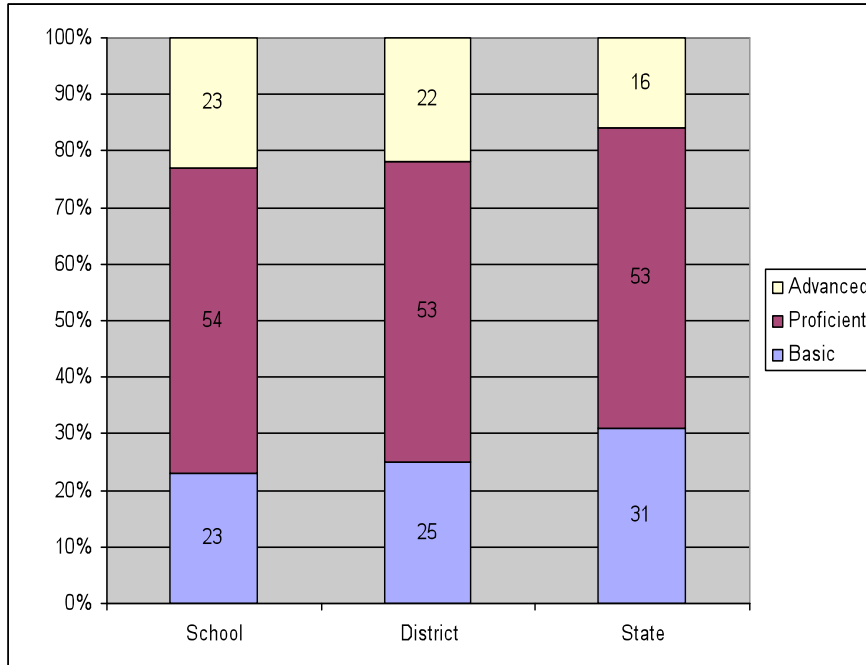
1. Overall Performance and Achievement Level Result

Average scale scores as a measure of overall performance should be reported and compared among school, district and the state. Achievement level results for various groups could also be reported and compared. Tables A1, A2 and A3 below provide such information so that a single school can compare its average scale score and achievement level results with the district and the state; similarly, a single district can compare its' average scale scores and achievement level results with the state.

A1a. Comparison of Achievement Levels at a School, District and the State



A1b. Comparison of Achievement Levels at a School, District and the State



A2. Comparison of Achievement Levels and Scaled Scores on a School, District and the State: Mean Performances

	Achievement Level		
	Basic	Proficient	Advanced
	350-411	412-449	450-650
School Average (427)	↑		
District Average (419)	↑		
State Average (403)	↑		

A3. Comparison of Achievement Levels of a School, District and the State: Percent Performing at each level

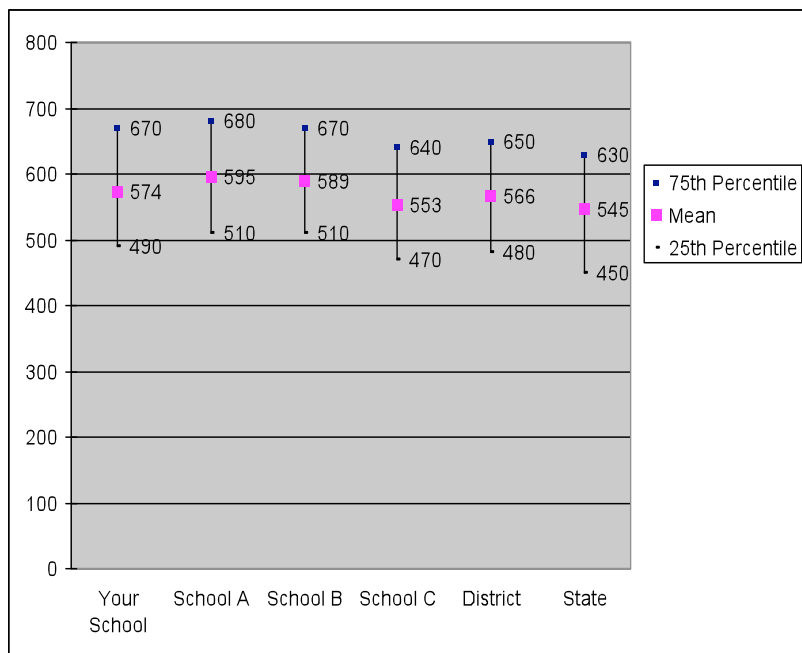
Achievement Level	Achievement Level Description A student is at this level :	% of Students		
		School	District	State
Advanced	Has demonstrated superior performance beyond the proficient level of mastery	23	22	16
Proficient	Has demonstrated competency over challenging subject matter; well-prepared for the next level of schooling	54	53	53
Basic	Has demonstrated only the fundamental knowledge and skills needed for the next level of schooling	23	23	31

- **School Comparison with Similar Groups**

In order to provide school administrators information on relative standing of their school's performance, schools with similar SES, demographics, ELS, or minority can be reported.

Ohio State has a *Similar District Methodology*, which can generate comparison groups of up to 20 districts for any individual district. The method concerns six dimensions of district characteristics simultaneously: District Size, Poverty, SES, Rural/Urban Continuum, Race/Ethnicity and Non-Agricultural and Non-Residential Tax Capacity. The districts with the smallest six-dimensional distance from the target district will be regarded as the comparable group for the targeted individual district. The more detailed information about this method can be found at https://webapp2.ode.state.oh.us/similar_districts/ For instance, schools with similar SES, demographics or minority combinations may be defined as comparable groups. The Table A4 below shows the comparison of school score distribution with comparable groups.

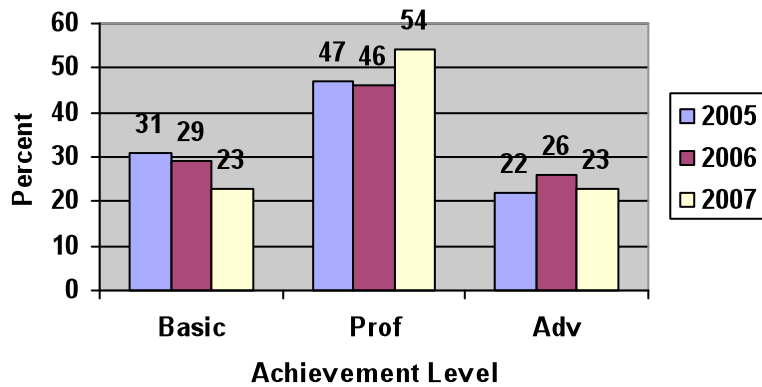
A4. Comparison of Comparable Schools within the district and state: Score Distribution



- **Three Years' Performance Trends**

Showing the trends of the school performance for consecutive years will be helpful for school and district administrators to see their school progress across time. Tables B1, B2, B3 and B4 provide such information.

B1. Three Year Comparison on Achievement Levels for a single Group (School or District)



B2. Three Years Comparison on Achievement Levels and Scaled Scores for a single Group (School or District)

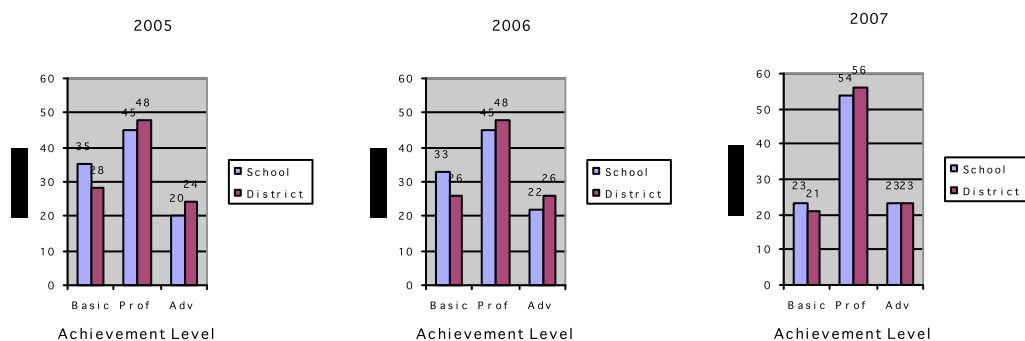
Achievement Level	Scaled Score	2005		2006		2007	
		N	%	N	%	N	%
Advanced	450-650	89	20	105	24	113	25
Proficient	412-449	182	41	188	43	241	53
Basic	350-411	174	39	144	33	101	22
Total N of Students		445		437		455	

B3. Three-Year Comparison on Achievement Levels and Scaled Scores between a School and District: Percent Performing at Each Level

Achievement Level	Scaled Score	2005		2006		2007	
		School	District	School	District	School	District
Advanced	450-650	20*	24	24	29	25	24
Proficient	412-449	41	50	43	46	53	55
Basic	350-411	39	26	33	25	22	21
Total N of Students		445	1023	437	991	455	1109

* 20% of students in that school are at the advanced achievement level.

B4. Three Year Comparison of Achievement Levels between a School and District



Section Two: Score Reporting at Student Level Audience for administrators

School administrators not only would like to know their school's relative performance with other schools and with its district and the state, but also want to examine its own performance in a much more detailed way. Specifically, school administrators and classroom teachers want to know how their students performed on each item and each

strand, and what are the strengths and weaknesses in the student's performance. This information can be used to monitor their students' learning as well as to adjust the instruction to increase the emphasis on areas with weak performance. Two types of score information are provided for at or below school audience: strand information reporting and analysis of released items.

- **Strand Information Reporting**

C1. Comparison of Achievement on Strands by Students at Each Performance

Level: State versus District

Strand	N of Item	District % Correct by Students at Each Level			State % Correct by Students at Each Level		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
Algebra/ Patterns	9	53 ¹	79	83	58	78	86
Geometry/ Measurement	7	67	70	91	69	84	89
Statistics/ Probability	10	69	84	89	66	86	93
Number Concepts /Computation	6	46	67	73	44	67	78
Processes of Mathematics	4	33	49	63	39	61	72

Note: ¹ In the district, 53% of the items in Algebra/Patterns were answered correctly by students who are at the Basic achievement level. Comparisons among basic, proficient and advanced can only be made within a strand; in other words, no comparison can be made across strands since strands are not equated for difficulty.

C2. Average Percent of Items Correct

Strand	N of Items	Average Percent of Items Correct					
		School		District		State	
		N	%	N	%	N	%
Algebra/Patterns	9	6	67	5	55	4	44
Geometry/Measurement	7	6	86	5	72	4	57
Statistics/Probability	10	7	70	6	60	5	50
Number Concepts/ Computation	6	2	33	3	50	4	67
Processes of Mathematics	4	3	75	2	50	3	75

Note: Comparisons among school, district and state can only be made within a strand; in other words, no comparison can be made across strands since strands are not equated for difficulty

- **Released Item Analysis**

HSA has released items for all the subjects it tests since 2003. Take algebra as an example, there are 35-38 items released publicly each year, and those items can be used as mini-assessments for classroom teachers. For each of the released items, the goal, expectation, indicator and the correct answer or scoring rubric have been provided on the MDK12.ORG. An example of a released item appearing on the site is shown below.

Released Item No.2 out of 37 (2007)

Look at the pattern: $a + 5, 2a + 8, 3a + 11, 4a + 14, \dots$

If this pattern continues, what will be the next term?

- A. $4a + 16$
- B. $4a + 17$
- C. $5a + 14$
- D. $5a + 17$

Goal 1: The student will demonstrate the ability to investigate, interpret, and communicate solutions to mathematical and real-world problems using patterns, functions, and algebra.

Expectation 1.1: The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology.

Indicator 1.1.1: The student will recognize, describe, and/or extend patterns and functional relationships that are expressed numerically, algebraically, and/or geometrically.

The correct answer for this item is: D

More organized information can be provided to educators through the website with the following as one possibility. The released items can be organized by strands and, in addition to the information above on this item, the examinees' percent choosing each alternative response can be reported at the school, district and the state level as shown below. When classroom teachers use these released items as a mini-assessment, they can compare their classroom level results with the school, district and state level results and estimate the weaknesses and strengths of their students as well as the implications for teachers' instruction, curriculum and other factors that might impact student's success. The inference should be very tentative since the sample size in a classroom is not large and many factors outside a teacher's control can influence the success of a class.

More detailed information regarding a released test item

Strand	Item No.	Percent Distribution			Response or Score
		School	District	State	
Algebra/ Patterns	1	3.6	14.5	15.1	A
		3.6	2.8	2.3	B
		0	0	2.8	C
		92.9	82.3	79.3	D*
		0	0.4	0.5	NR
	4	28.6	39.1	32.9	0
		41.1	37.5	34.3	1
		28.6	19.0	28.6	2
		1.8	4.4	4.2	NR
		1.0	0.8	1.0	Mean
	7				
...					
Geometry/ Measurement	5				
	19				
	...				
Statistics/ Probability	12				
	...				
...					

Note: * is correct answer;

NR means Not Respond.

Limitations

The following are three particular limitations.

First, it may be possible to provide meaningful characterizations for intermediate points on the score scale so that a more detailed presentation would become possible. It would be nice if each content area (Algebra, English, Biology and Government) in the HSA, for example, had a more detailed classification than just pass-fail. We can also focus on the strand level descriptions and their items and that is helpful, but other possibilities may exist for anchoring the scale at intermediate points, as NAEP has done.

Second, reporting the measurement error with scores has been neglected in our presentations above. This is a requirement in the AERA, APA and NCME test standards (1999) on score reporting as stipulated below, although one might argue that these are not actually score reports. Our recommendation is to present some indication of the accuracy of each presentation.

“5.10 (Standards) When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.”

“13.14 (Standards) Score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.”

Third, reporting strands must be interpreted cautiously because of two very important limitations:

1. Reporting strands are based on different numbers of questions and, in some cases, the number of questions that makes up a reporting strand may be quite small. The smaller

number of questions results in scores that are less accurate than are needed for certain purposes, and attention needs to be given to this issue.

2. Strand scores are reported in terms of number correct and percent correct. The difficulty of the questions tested for each strand may vary from one administration to the next and vary across strands. Comparison by strands across administrations by number and percent correct may be misleading unless the proper base rates are provided. We recommend presenting this information as normative, so that a school is always compared to other schools or to the district or the state.

A Final Reflection: From Score Reports to State Assessment System

There is an increasing demand for using large-scale achievement test results to generate more specific inferences about examinees and the interventions that might improve their performance. The specific inferences mean that we cannot only rank order the students but we can also identify their strengths and weaknesses. National Research Council (NRC) 2001 claimed that:

“On the whole, most current large scale tests provide very limited information that teachers and educational administrators can use to identify why students do not perform well or to modify the conditions of instructions in ways likely to improve student achievement. (NRC, 2001, p, 27)”

As NCLB develops, many states are realizing that the most critical development is the reporting and using of test results to inform instruction to improve learning. States are searching for ways of reporting aggregated scores at the school level or the district level for the diagnostic information available from the state assessments. The hope is that this information will provide stakeholders with a profile of their students’ performance, and that this information can be used to improve the curriculum and guide instruction. However, not every assessment has a diagnostic function because the purpose of some assessment efforts is to provide summative information. The HSA is an end of course exam and unless a student fails and must repeat the course, there is really no natural opportunity or necessity to improve that student’s performance. The MSA exams are

quite different in their nature. For the most part they are intended to be a measure of the ongoing learning that occurs in school. In order to provide the kinds of information to satisfy different needs in education, states need to think about their current assessment system and what they wish to get from that effort. It can be something different for each stakeholder, from the classroom teacher, to the principle, to the district and state administrator. Ohio State, is one place that has developed a comprehensive assessment system. Maryland is another, although improvement in the reporting process can also be attained.

The WEB URL for Ohio State is:

<http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=16758>) Theirs is not the only way to do this and Maryland already has a comprehensive system of its own. The system could be more integrated with explicit coordination between the different parts. For example, teacher's exams could be related or coordinated to the county and/or state examinations.

This importance of designing a comprehensive state assessment system shows us that certain score reporting forms naturally come from certain types of assessments. Maryland can capitalize on the richness of their testing system to provide evidence required for the different purposes of testing and for the different constituencies demanding such information. It is important that the State continue to refine their system and one of the directions for improvement is the reporting of information for staff charged with improving the performance of the students in the state of Maryland.

Reference

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education* 17(2), 145–220.

Massachusetts Department of Education. (2007). *Guide to Interpreting the Spring 2007 MCAS Reports for Schools and Districts*.

http://www.doe.mass.edu/mcas/2007/interpretive_guides/full.pdf

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Ohio State Comprehensive Assessment System

<http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=16758>

Ohio State Similar District Methodology

https://webapp2.ode.state.oh.us/similar_districts/.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies*. Greensboro, NC: SERVE.

School Improvement in Maryland (MDK12). www.mdk12.org

1 We thank the Maryland State Department of Education for funding this project through the Maryland Assessment Research Center for Education Success.