Research Brief

April 2011 Office of Shared Accountability

Establishing Reading Proficiency Benchmarks for Grades 3–8

Huafang Zhao, Ph.D.

Executive Summary

This research brief describes how Grades 3–8 end-of-year reading benchmarks were established in Montgomery County Public Schools (MCPS). The following topics are discussed in this brief.

- MCPS reading benchmarks for proficient and advanced levels in Grades 3–8
- Methods used to set the benchmarks
- Validity evidence of the benchmarks
- Validation results for the benchmarks

Background

One milestone of the MCPS strategic plan, *Our Call to Action: Pursuit of Excellence* is that all students will achieve or exceed proficiency standards in reading on local or state assessments (MCPS, 2010). In MCPS, end-of-year reading benchmarks have been set and revised since 2002 for students in kindergarten to Grade 2. The focus on early literacy and continuous monitoring intends to close achievement gaps among different groups of students. To date, no reading benchmarks have been set for Grades 3–8.

MCPS has published a pathway to college readiness that includes seven important indicators (keys) for students from kindergarten to Grade 12 (Von Secker, 2009). The pathway aims at helping students reach important benchmarks during their elementary and secondary education in preparation for college. Establishing Grades 3–8 proficient and advanced reading benchmarks can help teachers and parents monitor student progress towards meeting expected MCPS end-of-year performance and attainment of the third key (advanced reading) in reading from Grade 3 to Grade 8.

In 2008, an MCPS committee was convened to discuss how to set the reading benchmarks with appropriate methodology. After considerable research on appropriate assessment options, the committee decided to set the benchmarks based on the Measures of Academic Progress—Reading (MAP-R) currently used successfully in MCPS.

Since 2004, MCPS has been administering the MAP-R, a computer adaptive achievement test developed by Northwest Evaluation Association (NWEA) in Grades 3–8. Students in all MCPS schools take MAP-R at least twice a year, fall and spring. MAP-R winter administration is optional for schools.

MAP-R measures six reading areas: word recognition, reading comprehension, inferential or interpretive comprehension, evaluative comprehension, literary responses or analyses, and general reading. The MAP-R is aligned with Maryland state curriculum standards in reading (Bowe and Cronin, 2005; Adkins, 2007). MAP-R scores are reported in RIT (Rasch unIT) scale. The RIT scale reports student performance on an equal-interval and measures student's academic growth (NWEA, 2008). The test results are available within a few days after test administration. The fast delivery and comparability of RIT scores across grade levels provide frequent formative data to teachers for monitoring a student's reading progress in a timely manner.

The Maryland School Assessment (MSA) is a criterion-referenced achievement test that meets the *No Child Left Behind Act of 2001* (NCLB) testing requirements. The test is given in spring each year to Maryland students in Grades 3 through 8. Student performance on the MSA is classified as basic, proficient, and advanced.

While the MSA provides valuable information regarding the annual progress required by the NCLB, the test results are only available at the end of a school year. The benchmarks can provide teachers information about student progress toward the end-of-year goals. This brief addresses the following research questions:

1. How were end-of-year benchmark scores for Grades 3–8 reading proficient and advanced levels established?
2. What was the validity evidence for reading benchmarks?

3. How were Grade 5 benchmarks verified using different methods in a pilot study?

Methodology

Standard Setting Procedures

Setting a benchmark is a process of establishing expected performance standards. Two major methods are commonly used to set a benchmark—test-centered or examinee-centered. The test-centered method involves examination of test items and judgment of expected performance on each test item, while the examinee-centered method relies on judgment of student ability (Cizek, 1996; Jaeger, 1993; Kane, 1998).

Because MAP-R test items were not available to MCPS educators, it was not feasible to review test items using any test-centered methods. As a result, the two examinee-centered methods (Livingstone & Zieky, 1982, Zieky & Livingston, 1977), namely the borderline group and contrasting groups methods, were selected to set reading benchmarks.

In general, the borderline group method establishes benchmarks based on competency of borderline students, defined as just proficient or just advanced. Once the borderline group of students is selected, students' median scores are calculated as the benchmarks (Livingstone & Zieky, 1982, Zieky & Livingston, 1977).

All tests are subject to Standard Error of Measurement (SEM). If a student were to take the same test repeatedly, with no change of knowledge and preparation, it is possible that the resulting scores would be slightly higher or lower than the score that precisely reflects the student's actual level of knowledge and ability. MSA reading proficient and advanced cut scores, associated SEM at cuts (Maryland State Department of Education, 2008), and score ranges for borderline students are displayed by grade in Table A1 (Appendix A).

At each grade level, the proficient borderline group includes students who scored one SEM at or above the MSA proficient cut score, while the advanced borderline students were those who scored one SEM at or above the MSA advanced cut score. For example, in Grade 3 the MSA reading cut score for proficient level is 388 with a SEM of 11 (Table A1, Appendix A). The proficient borderline students are those whose scores fall between 388 and 399. After borderline students were selected, their median scores on the spring MAP-R were calculated as the end-of-year proficient benchmarks. The same method applies to the advanced benchmark.

In the contrasting groups method, examinees are sorted into different groups based on predetermined criteria or descriptors. Their test score distributions are plotted and the intercepts of score distributions are selected as the performance standards or benchmark cut scores (Hambleton & Pitoniak, 2006). The contrasting groups method was used in a Grade 5 pilot study to verify Grade 5 benchmarks established with the borderline group method. The pilot study is described in Appendix B.

An assessment has criterion-related validity if it has concurrent validity with other assessments (Kane, 2006). Evidence of the concurrent validity of MAP-R can be found by examining how well it is related to MSA reading administered at about the same time.

Students who had MSA reading scores and MAP-R scores in 2009 were included in MCPS reading benchmark setting process.

Results

Grades 3–8 Reading Benchmark RIT Scores

Table 1 presents Grades 3–8 end-of-year reading benchmark RIT scores established with the borderline group method. For instance, the Grade 3 MAP-R reading benchmark is 194 for the proficient level and 216 for the advanced level by the end of the school year.

Table 1
Grades 3–8 MAP-R End-of-Year Benchmark Scores Established with Borderline Group Method by Grade

	MAP-R end-of-year benchmark			
Grade	Proficient borderline group *N*	Proficient benchmark RIT scores	Advanced borderline group *N*	Advanced benchmark RIT scores
3	790	194	1347	216
4	623	198	1793	221
5	605	203	1165	223
6	851	209	1575	225
7	733	213	1547	227
8	1100	217	1527	231

Impact Data for Reading Benchmarks

Once the reading benchmarks are established, it is important to examine the impact data when the

benchmarks are applied. As shown in Table A2 (Appendix A), 19.6% of students were identified as below proficient, 55.3% as proficient and 25.1% as advanced in 2009 if Grade 3 reading benchmarks (194 and 216) established with the borderline group method are used.

Percentages of Asian American and White students meeting the advanced benchmark are higher than those for African American and Hispanic students.

Validity Evidence

The correlation between 2009 MSA reading scale scores and MAP-R RIT scores in fall, winter, and spring 2008–2009 is reasonably high as shown in Table 2, with Pearson correlation coefficients ranging from 0.58 to 0.77. The positive correlation means that students who score high on the MAP-R tend to score high on the MSA.

Table 2
Pearson Correlation Coefficients Between 2009 MSA Reading Scale Scores and MAP-R RIT Scores in 2008–2009 School Year

	Fall RIT	Winter RIT	Spring RIT
Grade 3 MSA	0.73	0.75	0.75
Grade 4 MSA	0.75	0.77	0.77
Grade 5 MSA	0.75	0.75	0.76
Grade 6 MSA	0.67	0.65	0.68
Grade 7 MSA	0.64	0.63	0.65
Grade 8 MSA	0.60	0.58	0.60

Classification consistency between the MAP-R and the MSA can provide concurrent validity evidence for benchmark scores. Classification consistency means students who scored below the proficient benchmark scores on MAP-R also scored basic on MSA, those who scored proficient on MAP-R also scored proficient on MSA, and those who met the advanced benchmark on MAP-R also scored advanced on MSA.

Underestimation may occur when students who were identified as below proficient on MAP-R performed at the higher level on MSA. Overestimation may occur when students who were identified as proficient actually performed at lower level on MSA.

Table 3 presents classification consistency between MAP-R and MSA. For instance, Grade 3 MAP-R reading benchmarks accurately identified 69.3% of students. The underestimation rate was 21.1% in Grade 3. This means that Grade 3 students who did not meet the end-of-year proficient or advanced reading benchmarks on MAP-R actually scored proficient or advanced on the MSA. Only 2.6% of students who were below Grade 3 MAP-R benchmark for the proficient level scored basic on MSA reading, and 7% of students who met Grade 3 MAP-R advanced benchmark scored below the MSA advanced level.

Table 3
Classification Consistency Between 2009 MSA Reading and MAP-R End-of-Year Benchmarks Established with Borderline Group Method

	MAP-R end-of-year benchmark			
Grade	Accuracy %	Underestimate (perform better on MSA) %	Overestimate for proficient (perform worse on MSA) %	Overestimate for advanced (perform worse on MSA) %
3	69.3	21.1	2.6	7.0
4	70.3	21.1	2.6	6.0
5	65.1	30.8	1.6	2.5
6	68.1	23.7	2.4	5.8
7	68.5	23.5	2.5	5.6
8	65.2	24.0	2.4	8.4

In general, the higher a cut score is set, the higher the rate of underestimation. To ensure success on MSA, it is better to have a lower rate of overestimation. Since the SEM for MSA advanced cut scores are higher than those for the proficient cut scores at all grade levels, the overestimation for MSA advanced is higher than that for MSA proficient level.

Validation of Grade 5 Reading Benchmarks

The Grade 5 pilot study is described in Table B1 (Appendix B). The validation results are summarized in Table B2 (Appendix B). As shown in Table B2, panelists' judgment in the Grade 5 pilot study produced a lower proficient benchmark (200) and the same advanced benchmark (223), compared with the borderline group results (203 and 223). When report card ranking was used, the results (204 and 224) were close to the benchmark scores yielded by the borderline group method (203 and 224). The validation results for Grade 5 increased our confidence in MCPS benchmark setting methodology.

In Grades 3 and 4, the report card ranking also was used with the contrasting groups method to validate results obtained with the borderline group method. Similar to Grade 5, the results produced with the two methods were very close.

Recommendations

Based on the benchmark setting study, we recommend the following:

- Use the MAP-R RIT cut scores (Table 1) established with the borderline group method as end-of-year reading benchmarks for Grades 3–8.
- Use RIT scores to monitor student progress across test administrations in a school year or across grade levels.

The recommendations are based on the following rationales:

- The borderline group method is an acceptable standard setting process for setting reading benchmarks.
- The Grade 5 pilot study results validated the benchmarks with different methods.
- MAP-R RIT scores were vertically equated so results are comparable across test administrations and across grades. Therefore, student progress can be monitored in a school year and across grade levels.
- There is a reasonably high correlation between MAP-R and MSA reading. Students who score at or above the benchmarks, have a higher probability of being proficient or advanced on MSA reading.

Caution

MAP-R tests include only multiple-choice items so performance in writing is not measured by MAP-R end-of-year reading benchmarks. No accommodations were provided to English language learner students and students receiving special education services on MAP-R. Therefore, teachers may need to use other data points when evaluating performance of these students.

References

Adkins, D. (2007). *A study of the alignment of the NWEA RIT scale with the Maryland Assessment System.* Lake Oswego, OR: Northwest Evaluation Association.

Bowe, B. and Cronin, J. (2005). *Aligning the NWEA RIT scale with the Maryland School Assessment (MSA).* Lake Oswego, OR: Northwest Evaluation Association.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20–31.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards (pp. 433–470) in R.L. Brennan (edition). *Educational Measurement*. Westport, CT: Praeger.

Jaeger, R. (1993). Certification of student competency. *Educational Measurement* (4th ed.). Westport, CT: Praeger Publisher.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard setting methods. *Educational Assessment, 5*, 129–145.

Kane, M.T. (2006). *Validation,* in Educational Measurement (4th ed.). Westport, CT: Praeger Publisher.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Maryland State Department of Education (2007). *Maryland School Assessment–Reading: Grades 3 through 8 technical report.* Baltimore, Maryland.

Maryland State Department of Education (2008). *Maryland School Assessment–Reading: Grades 3 through 8 technical report.* Baltimore, Maryland.

Montgomery County Public Schools (2010). *Our Call to Action: Pursuit of Excellence.* Rockville, Maryland: Montgomery County Public Schools.

Northwest Evaluation Association (2008). *RIT scale norms for use with Measures of Academic Progress*. Lake Oswego, OR.

Von Secker, C. (2009). *Closing the gap: Seven keys to college readiness for students of all races/ethnicities*. Rockville, MD: Montgomery County Public Schools.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment test.* Princeton, NJ: Educational Testing Service.

Author Note. The author appreciates valuable comments received from Mrs. Jody Leleck, Mrs. Ebony Y. Langford-Brown, Mr. Steven Fink, Mrs. Cindy L. Loeb and Ms. Suzanne R. Merchlinsky.

Appendix A

Table A1
Grades 3–8 MSA Cut Scores, Standard Error of Measurement (SEM) at MSA Cut Scores, and Score Range For Borderline Group Students by Grade

	Proficient on MSA Reading			Advanced on MSA Reading		
Grade	Proficient Cut Score	SEM at Proficient Cut*	Proficient Borderline Group Score Range	Advanced Cut Score	SEM at Advanced Cut*	Advanced Borderline Group Score Range
3	388	11	388–399	456	16	456–472
4	371	12	371–383	437	17	437–454
5	384	12	384–396	425	13	425–438
6	381	11	381–392	421	13	421–434
7	385	11	385–396	425	14	425–439
8	391	11	391–402	425	12	425–437

*Source: Page 119 of MSA Reading 2007 Technical Report (MSDE, 2007).

Table A2
2009 Impact Data for MAP-R RIT Cut Scores Established With Borderline Groups Method by Race/Ethnicity

Grade	Group	Total *N*	Below Proficient *N*	Proficient *N*	Advanced *N*	Below Proficient %	Proficient %	Advanced %
3	All	9,688	1,902	5,359	2,427	19.6	55.3	25.1
	African Am.	2,165	638	1,264	263	29.5	58.4	12.1
	Asian Am.	1,577	148	882	547	9.4	55.9	34.7
	Hispanic	2,059	740	1,173	146	35.9	57.0	7.1
	White	3,866	372	2,027	1,467	9.6	52.4	37.9
4	All	9,536	1,482	5,294	2,760	15.5	55.5	28.9
	African Am.	2,160	511	1,358	291	23.7	62.9	13.5
	Asian Am.	1,464	94	740	630	6.4	50.5	43.0
	Hispanic	2,048	631	1,225	192	30.8	59.8	9.4
	White	3,829	239	1,948	1,642	6.2	50.9	42.9
5	All	9,728	1,440	4,375	3,913	14.8	45.0	40.2
	African Am.	2,275	528	1,238	509	23.2	54.4	22.4
	Asian Am.	1,514	109	565	840	7.2	37.3	55.5
	Hispanic	2,021	573	1,150	298	28.4	56.9	14.7
	White	3,889	227	1,405	2,257	5.8	36.1	58.0
6	All	9,505	1,704	3,691	4,110	17.9	38.8	43.2
	African Am.	2,236	641	1,034	561	28.7	46.2	25.1
	Asian Am.	1,511	130	504	877	8.6	33.4	58.0
	Hispanic	2,051	701	948	402	34.2	46.2	19.6
	White	3,680	230	1,191	2,259	6.3	32.4	61.4
7	All	9,802	1,661	3367	4,774	16.9	34.4	48.7
	African Am.	2,163	568	953	642	26.3	44.1	29.7
	Asian Am.	1,583	127	452	1,004	8.0	28.6	63.4
	Hispanic	2,091	699	899	493	33.4	43.0	23.6
	White	3,940	258	1,054	2,628	6.5	26.8	66.7
8	All	9,824	1,895	3,554	4,375	19.3	36.2	44.5
	African Am.	2,257	676	970	611	30.0	43.0	27.1
	Asian Am.	1,483	124	482	877	8.4	32.5	59.1
	Hispanic	2,006	761	841	404	37.9	41.9	20.1
	White	4,049	329	1,251	2,469	8.1	30.9	61.0

Note. Percent may not add up to 100 due to rounding. American Indian students were included but not reported separately.

Appendix B

Grade 5 Pilot Study

Background

In November 2008, a group of teachers and reading specialists developed descriptors for Grade 5 reading proficiency based on the Montgomery County Public Schools reading curriculum (Table B1). As part of the pilot study, Grade 5 reading teachers and reading specialists were invited to participate in a standard setting meeting in spring 2009. Over 50 participating teachers were first trained to understand proficiency descriptors, the standard setting process, and the method to calculate cut scores. Panelists were asked to fill out an evaluation form (located at end of Appendix B) before proceeding to rank their own students. The panelists indicated that they understood the MCPS reading curriculum, and felt comfortable ranking their own students according to the descriptors. Then panelists classified their students into different proficiency groups according to descriptors.

Table B1
Performance Descriptors for Grade 5 Reading Proficiency

BELOW	Somewhat Below	ON	Somewhat Above	ABOVE
• Comprehend orally or in writing literary and informational texts **that are below grade level** • During teacher directed small group instruction • **Significant teacher support** • **Struggles when confronted with grade-level texts**	• Comprehend orally or in writing literary and informational grade-level texts • During teacher directed small group instruction • **More teacher support when approaching more challenging reading situations**	• Comprehend orally or in writing literary and informational grade-level texts • During teacher directed small group instruction	• Comprehend orally or in writing literary and informational grade-level texts • During teacher directed small group instruction • **Less teacher support when approaching more challenging reading situations**	• Comprehend orally or in writing literary and informational grade-level **or above** texts • During teacher directed small group instruction **or in an independent setting** • **Can synthesize information during group discussions**

Findings

Results of Contrasting Groups Method with Panelists' Judgment

About 1,475 students were included in the analyses. The MAP-R spring score distributions of three groups were plotted (below-grade, on-grade, and above-grade). After statistical adjustment (smoothing), two score intercepts of the score distributions were selected as on-grade-level and above-grade-level benchmarks. Figure B1 shows the on-grade and above-grade cut scores are 200 and 223 respectively, based on panelists' judgments.

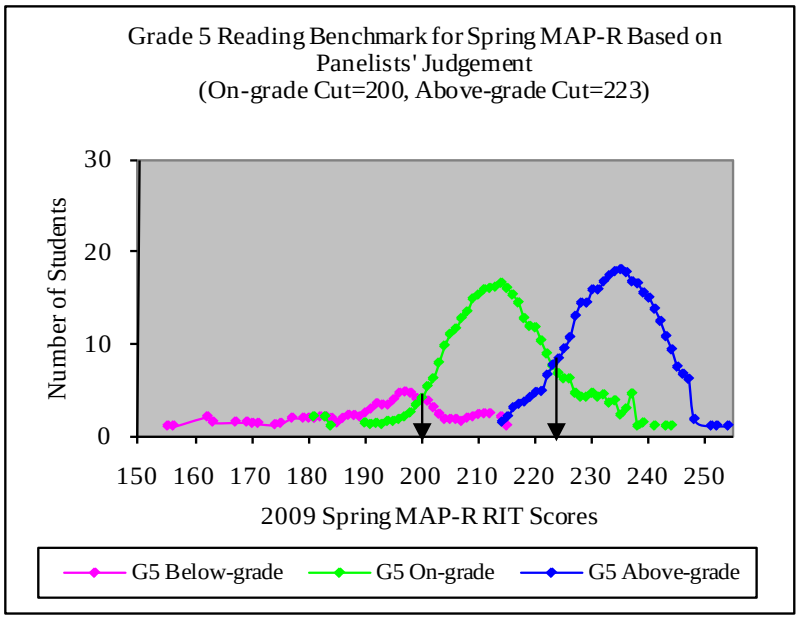

Figure B1. MAP-R cut scores for Grade 5 on-grade and above-grade levels with the contrasting groups method based on panelists' judgments of their own students.

Results of Contrasting Groups Method with Teachers' Rating on Report Card

The second validation used teachers' rating on report cards in Grade 5. On MCPS elementary school report cards, teachers categorize students into three reading levels: below-grade, on-grade, and above-grade levels depending on how well students understand the narrative, expository, and procedural text materials.

All Grade 5 students (n = 9,700) with valid data on the report cards and spring MAP-R were included. Based on report cards, students' spring MAP-R score distributions were plotted. After smoothing, two intercepts of the MAP-R score distributions were defined as on-grade and above-grade benchmarks. Figure B2 shows the MAP-R scores distribution for students based on their group membership according to teachers' ranking on the report card. The first intercept between below- and on-grade groups was 204 for on-grade reading level, and the second intercept between on-grade and above-grade groups was 224 for above-grade reading level.

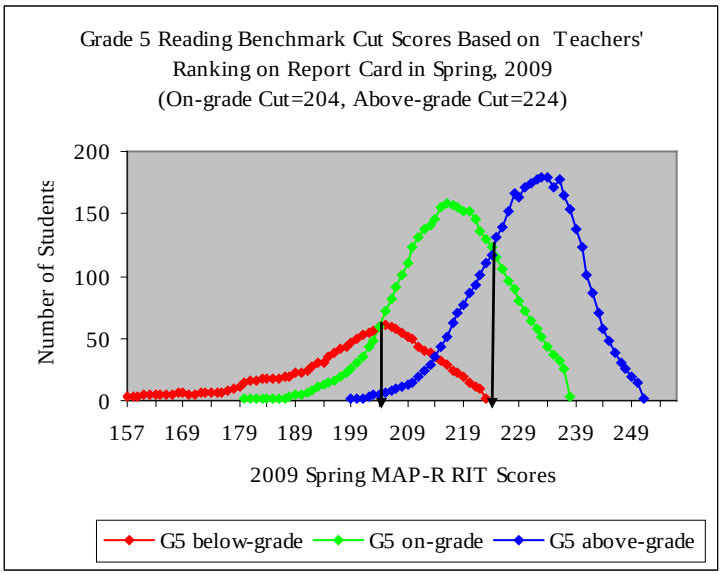

Figure B2. MAP-R cut scores for Grade 5 on-grade and above-grade levels with the contrasting groups method based on teachers' ranking of students on the report card.

Table B2
Comparison of Grade 5 MAP-R Reading Benchmarks and Cut Scores Obtained with Two Contrasting Groups Methods

Method	Proficient or on-grade RIT cut	Advanced or above-grade RIT cut
Two Contrasting Groups		
Panelist judgment	200	223
Report card ranking	204	224
Borderline group	203	223

Summary of Evaluation for the Grade 5 Reading Standard Setting Training

May 5, 2009

The purpose of the evaluation form is to obtain participants' feedback about the training on standard setting process and descriptors. Feedbacks are anonymous so no individuals can be identified. The summary is based on 46 respondents.

Group representing: Teachers 58.7%
Reading specialists 41.3%
Type of school you teach/work in: 15.2% from Title 1 School
How many years have you been teaching reading? Average years = 14
Gender: 95.7% Female 4.3% Male

Race/ethnicity: 6.5% Asian American 6.5% African American 2.2% Hispanic 76.1% White 8.7% Other

1. Place a check mark (√) under only one category (Strongly Agree, Agree, Disagree, or Strongly Disagree) to indicate the degree to which you agree with each statement.

	Strongly Agree	Agree	Disagree	Strongly Disagree	No Response
a. I understand the purpose of training.	41.3%	52.2%	4.3%		2.2%
b. I understand the tasks I need to do.	39.1%	56.5%	4.3%		
c. I have a clear understanding of the MCPS Grade 5 reading curriculum content standards.	54.3%	41.3%	2.2%		2.2%
d. The training on standard setting methods was sufficient.	23.9%	56.5%	17.4%		2.2%
e. I understand how cut scores are established.	15.2%	60.9%	19.6%	4.3%	
f. The training on the performance level descriptions was appropriate in giving me the information I need to complete the tasks.	26.1%	65.2%	6.5%	2.2%	
g. I feel comfortable ranking my own students.	54.3%	34.8%	2.2%		8.7%
h. I understand how to record data for my ranking.	50.0%	43.5%	4.3%		2.2%
i. The standard setting experience is valuable for professional development.	23.9%	63.0%	6.5%		6.5%

2. Have you participated in a standard setting workshop before today?

   Only one out of 46 respondents participated before.

3. Do you have any comments about the procedure?