The Missouri Assessment Program

An Independent Evaluation

William D. Schafer
University of Maryland

September 2002

Commissioned by
Missouri National Education Association

TABLE OF CONTENTS

The Missouri Assessment Program

An Independent Evaluation

William D. Schafer
University of Maryland

EXECUTIVE SUMMARY

The Missouri National Education Association (MNEA) commissioned this study of the Missouri Assessment Program to evaluate the testing program's ability to contribute to Missouri's standards-based school-reform efforts. The goal is to provide recommendations that can enhance the value of the MAP within the state's school accountability system.

The Missouri Department of Elementary and Secondary Education (DESE) cooperated with MNEA and the author by supplying materials and responding to questions as they arose during the course of this study. Further, DESE has had input into the body of information that forms the basis for this evaluation. DESE's cooperation was essential to the completion of this study, and is greatly appreciated.

While this evaluation concludes that the MAP is a successful element of Missouri's student assessment program, the implementation of several recommendations could enhance its value to the state and to educators. To facilitate understanding, the recommendations from the report are grouped into four general themes below: Clarifications of Standards, Technical Considerations, Fairness and Consequences, and Reporting Results.

Clarification of Standards

Two types of standards are commonly used in an assessment program. These are content standards (what is tested) and performance standards (how well students are expected to do). These first recommendations are intended to clarify, especially for Missouri educators, precisely what Missouri's standards are.

1. For clarity of language, communicate to educators who are used to other states' terminology that Missouri's "Achievement Levels" are what others commonly call "performance standards" and that Missouri's "Performance Standards" are what others commonly call "process standards."

2. Develop assessment limits (specifications for the content that can be tested) for each sub-domain description in the state's curriculum frameworks, and describe where and how each sub-domain is assessed in the MAP through test maps (i.e., blueprints) that are useful both to educators in planning instruction as well as to test developers in creating items and other test prompts. (The supplement to the curriculum frameworks provides a good start in mathematics.)

3. Add clarity to the released scoring guides by identifying the achievement level (or scale score) that each anchor paper (example of student work) represents.

Technical Considerations

Both assessment systems (how data are generated) and accountability systems (how data are used) should conform to rigorous professional standards. Especially important in a statewide program are the quality of the data for purposes of making decisions about students, districts and schools and the effects of the program on improvement of education practices throughout the state. The purpose of this set of recommendations is to further these quality goals as well as to enhance the evidence about the degree to which they are met.

4. Replace the Terra Nova with sampling of statewide item pools as is done on the other two portions of the MAP to avoid the possibility that teachers are teaching to the specific items used on accountability measures.

5. Ensure that data from pilot and field testing are used in the revision of MAP assessments. Performance data should include frequencies and samples of student responses at all scoring-tool anchor points, especially the highest levels. Anecdotal data related to ease of administration and item validity should be collected from teachers as they administer the items.

6. Develop and disseminate the rules that are used to score MAP items so that parents, students and school personnel clearly understand how all student responses, including blanks and non-readable responses, are used in determining student test scores.

7. Conduct correlational studies relating MAP results to existing data such as teacher grades and local tests to provide evidence that scores resulting from MAP are consistent with those from other ways of measuring the same achievement.

8. Reconsider the advisability of using multiple ways to generate district performance scores for accreditation. This is a highly unusual feature of MAP that is not commonly used in other states. This feature leads to different criteria for different districts and schools, is not likely to be understood by educators or the public, and is not conducive to setting effective district and school educational goals.

9. If the state decides to continue using multiple criteria, then it should develop and include in a finalized version of the Missouri School Improvement Program (MSIP) scoring guide a justification for the different ways in which districts are awarded points for student performance. Studies to evaluate the consequences of these alternatives should be initiated as data accumulate.

10. Develop scoring and reporting systems for MAP-A (an alternate portfolio system used for the lowest-functioning students) that support accountability. These systems should assess the educational goals that students in these programs are pursuing and provide sound data that can form the basis for making outcomes-based judgments about program success.

Fairness and Consequences

Students and schools should be motivated to score well on assessments that are fair to everyone. To be fair, assessment results should have the same meaning for all. For example, they should be administered under the same conditions for everyone (i.e., be standardized). Schools should be motivated to educate toward and students should be motivated to demonstrate the highest possible levels of achievement. The recommendations in this section are oriented toward increasing motivation for students and schools, and toward ensuring that the assessments are as fair as possible.

11. Consider adding student incentives for test performance, especially for older students, to increase students' motivation to perform well on MAP assessments.

12. Monitor districts' and schools' administration of the MAP assessments to ensure that they comply with the procedures detailed in administration manuals, including procedures related to accommodations and inclusion.

13. To measure districts and schools against the same standards, establish and communicate school and district achievement targets, both for overall performance and for adequate yearly progress, based on percents of students at and above selected achievement levels for schools and districts.

14. Incorporate the Level Not Determined (LND) category into district and school accountability decisions since the mission of districts and schools is to educate all students, not just those who are tested.

Reporting Results

Assessment results are most valuable when users understand them. Fundamental to effective use is knowing what content is being assessed and how to use the information for improvement. Scores should be available on education outcomes that cover the spectrum of important goals, but at the same time are focused enough to lead to informed educational change. Finally, the scores should not be over-

interpreted beyond the quality of the data. These final recommendations are intended to enhance the impact of Missouri assessments upon schooling through the ways they are reported.

15. Monitor the success of districts with as broad an array of standardized student achievement indicators as possible—including such criteria as attendance and dropout rates—and, at the same time, minimize the impact of non-achievement factors in making decisions about district and school success.

16. Include confidence bands in reporting individual-student scale scores to convey the degree to which random factors contribute to score imprecision and to help educators avoid over-interpreting small score differences.

17. Group items in each content area tested by MAP into instructionally relevant categories and report scale-score summaries on these item categories for groups of students, such as classrooms, schools, districts and the state, so that decision making about curriculum and instructional programs can be based upon meaningful student-outcome results.

18. Discontinue reporting percent of items answered correctly for each content and process standard because differences in difficulty between the item pools have unknown impacts upon those percents.

19. Discontinue reporting process sub-scores (i.e., scores for thinking skills and processes) since they may not generalize well from one content area to another in terms of either what students are asked to do or their success in doing it.

William D. Schafer, Ed.D., is professor emeritus and affiliated professor with the Maryland Assessment Research Center for Education Success, University of Maryland-College Park. Prior to his work there, he served as state director of student assessment for the Maryland State Department of Education (1997-1999) and as a faculty member in the Department of Measurement, Statistics and Evaluation at the University of Maryland (1969-2000).

Schafer earned a doctorate in measurement, statistics and evaluation from the University of Rochester in 1969. He has authored numerous journal articles and book chapters on assessment throughout his career. He can be contacted by e-mail at ws7@umail.umd.edu.

The Missouri Assessment Program

An Independent Evaluation

William D. Schafer
University of Maryland

INTRODUCTION

The Missouri National Education Association (MNEA) commissioned this study of the Missouri Assessment Program to evaluate the testing program's ability to contribute to Missouri's standards-based school-reform efforts. The goal is to provide recommendations that can enhance the value of the MAP within the state's school accountability system.

The Missouri Department of Elementary and Secondary Education (DESE) cooperated with MNEA and the author by supplying materials and responding to questions as they arose during the course of this study. Further, DESE has had input into the body of information that forms the basis for this evaluation. DESE's cooperation was essential to the completion of this study and is greatly appreciated.

This report begins with the QUESTIONS, posed by MNEA, that guided the study from the outset. An OVERVIEW of the Missouri school assessment and accountability program follows. The overview has been reviewed by MNEA and DESE. Next, the EVALUATION section is oriented around the initial questions, followed by additional issues. Recommendations are embedded in this material. Finally, all the RECOMMENDATIONS are reorganized by themes in a special section at the end for convenience.

In the spirit of program enhancement, the general orientation of this evaluation is formative. While this study concludes that the MAP is a successful element of Missouri's student assessment program, the implementation of several recommendations could enhance its value to the state and to educators. It should be understood that this report was developed outside Missouri and that some of its recommendations may not be feasible given local considerations not known by the evaluator. Nevertheless, the issues addressed by the recommendations should receive careful consideration whether or not the specific recommendations are implemented.

GUIDING QUESTIONS

This study attempts to make suggestions that are both educationally justifiable and reasonable to implement. Several specific questions were identified at the outset:

Content and Performance Standards

How is the domain of each assessment in the program described? What methods are used to communicate the domain to local education professionals? Do these products clarify effectively what is to be assessed? Are performance standards communicated effectively?

Alignment

How is the domain of each assessment sampled? Do the assessments remain consistent from year to year in their domain coverage? How are the demands of the assessments chosen or developed to be appropriate for the ages and grades at which they are used?

Technical Aspects

What is the technical quality of the assessments? Does their reliability support their intended uses? Is the validity evidence compelling? Are there important areas where validity evidence is missing (e.g., content evidence, convergent evidence, divergent evidence, consequential evidence, fairness)? How have performance standards been developed?

Administration

To whom are the assessments given? Are they given under standardized conditions? Are inclusion and accommodation policies administered fairly? Are accommodations effective?

Scoring

How adequately are the assessments scored? What training do scorers receive? How is scoring quality maintained throughout the project?

Consequences

What are the information sources and procedures used to evaluate programs (e.g., schools) or individuals (e.g., students)? What are the consequences of these evaluations? To what extent are the procedures used statistically justified? How much emphasis is given to non-achievement factors?

Reporting

What reports are disseminated? Do they contain the most useful information? Are they received when they are most useful? Are they valuable for school improvement? Do they contain enough information to be useful for teachers?

OVERVIEW

General Education Statistics

According to the Missouri Web site, the state's public school enrollment (K-12) is 893,350, and the private-school enrollment is estimated at 95,701. There are 524 public school districts with a total of 2,138 public school buildings. These data are for school year 2000-2001.

Accreditation and The Missouri School Improvement Program

State accreditation is by school district. Each district completes a formal review every five years. As of the 2000-01 school year, there are 477 (91 percent) accredited districts, of which 109 are accredited with distinction. Another 46 (9 percent) are provisionally accredited, and one is unaccredited.

State law mandates the process of accrediting school districts. The Missouri School Improvement Program reviews and accredits districts using a five-year review cycle.

The reviews cover the areas of resources, processes and educational performance. Under resources, five areas are covered: areas of instruction, class size, support staff, administrative staff, and teacher certification and planning time. Under processes, three areas are covered: instructional design and practices, differentiation of instruction and supplemental programs, and school services. Under educational performance, student achievement and persistence are covered and, for districts without high schools, post-elementary school status.

Data are gathered by districts, by outside teams and through formal assessments. A manual describes standards in terms of "minimum" and in some cases "desirable" criteria for the resource standards. The process standards are assessed through observer teams. Statewide assessments are used to evaluate the educational performance standards using a "points" system based on levels and changes in percents of students at various achievement levels. Once the data are gathered, teams summarize the data. The DESE School Improvement Committee makes accreditation recommendations to the State Board of Education, but how the committee uses all the information from the teams is not specified.

A summary of each report and recommendations for accreditation are presented to the State Board of Education for action. Each district submits a School Improvement Plan that addresses concerns in the review report and may request subsequent reviews.

The Role of Testing in Accreditation

The MAP provides part of the evidence used in the MSIP. Districts receive points based on either high overall achievement or improving achievement.

The Outstanding Schools Act of 1993 requires a series of tests, which the MAP satisfies. The state began requiring the MAP tests in 1998. Five content areas are tested in staggered grades. Testing begins in the third grade with Communication Arts and Science. In the fourth grade, students are tested in Mathematics and Social Studies. Fifth graders are tested in Health

and Physical Education. An optional Fine Arts assessment was developed for the fifth grade, but it is not being administered due to state budget cuts. In the seventh grade, students take Communication Arts and Science assessments. Eighth graders take Mathematics and Social Studies tests. In the ninth grade, students take the Health and Physical Education assessment. Tenth graders are tested in Mathematics and Science, and eleventh graders are tested in Communication Arts and Social Studies.

The subject-area tests have been phased in. The Health and Physical Education assessment was required for the first time in Spring 2001. Spring 2000 was the first time that the Social Studies assessment was required in all public schools. The Science and Communication Arts tests were required for the first time in Spring 1999. The Mathematics test has been required since 1998.

Test Formats

In a typical subject and grade, testing consists of three separate sections, each taking about one hour to complete. One portion is the Terra Nova, a commercial, multiple-choice, machine-scored, nationally normed test. The same form is used each year. The second part consists of constructed-response items. The third part is a "performance event." Students are asked to show work, write paragraphs, create charts and graphs, and explain how they arrived at their answers. Forms on sections two and three are new each year; there is some repetition of items, but items are generally not re-used unless they have “rested” for at least three years.

Item Development

The Terra Nova items are part of the Terra Nova national assessments. Missouri teachers who have completed workshops in item writing, along with outside contractor staff, write all the other items. These are pilot tested in small studies before field testing.

Content Domains and Communication of Domains

The MAP tests are built around specifications derived from the “Show-Me Standards,” which identify what students should know and be able to do upon graduation from high school. The Show-Me Standards consist of 40 knowledge standards and 33 performance standards. The knowledge standards are separated into six content (subject matter) areas: Communication Arts, Mathematics, Science, Social Studies, Fine Arts, and Health and Physical Education. These standards are broad statements such as “data analysis, probability and statistics” (Mathematics, Standard 3). Specificity is found in the frameworks and assessment annotations. The performance standards are divided into four goals: gathering, analyzing, and applying information and ideas; communicating effectively; recognizing and solving problems; and acting as responsible members of society. As with the knowledge standards, these are broad statements such as “design and conduct field and laboratory investigations to study nature and society” (Goal 1, Standard 3). The text accompanying the standards indicates that they are to serve as a blueprint for districts to write curricula. Since the term “performance standard” is generally used with a different meaning (e.g., an achievement level) among assessment professionals, the term “process standard” will be used here, instead. This is consistent with the usage in the MAP Technical Manual.

Each content area has an accompanying framework for curriculum development. The framework document is intended to describe for local districts what specific student outcomes are expected at the state level at each tested grade level for that content area. For example, in mathematics, content expectations for data analysis, probability and statistics for fourth graders are expanded with five knowledge statements (e.g., "strategies to collect data") and four performance statements (e.g., "explore concepts of chance"). Sample learning activities are also described. Missouri teachers write or review items based on these curriculum frameworks.

Each curriculum framework is accompanied by an assessment annotations document in which the process standards from the framework are identified as to whether they are in the state assessment or left to local assessments. For example, the standard "explore concepts of chance" is to be assessed on the state assessment in the fourth grade. Booklets of released items for the constructed response and performance event sections are available. These also show scoring guides and, for the constructed response items, anchor papers.

The technical report contains maps of specific items (and types of items) identified with each of the content standards (Missouri Assessment Program Technical Report 2000 Supplement pp. 27-36) as well as types of items and numbers of points identified with each of the content standards (pp. 38-45) and each of the process standards (pp. 47-55). The degree of specificity of the standards is at the broadest possible level: the Show-Me Standards. There are some examples where an item is identified with two or more content standards as well as two or more process standards.

Administration

Students take all of the MAP assessments in the spring of the year (April through mid-May), depending on which grade-level assessments are scheduled for their grade. There is no out-of-level testing. Students usually take all parts of a given assessment in sequence during the same week; district administrators decide the week, unless the district is designated to be in the early returns sample (these data are used for item calibration). The examiner's and test coordinator's manuals give the specific information about these procedures.

Procedure manuals describe the assessment activities to be carried out in each school. The manuals also describe allowable accommodations. The basis for the accommodations is a series of studies carried out at the National Center on Educational Outcomes at the University of Minnesota.

Scoring

The constructed-response and performance-event sections are hand-scored by trained raters in three states, including Missouri (using Missouri teachers), Indiana and California. Missouri classroom teachers assist in the scoring process in both external states. Each student-constructed response is read by one rater with quality control during the rating study implemented through read-behinds and check sets.

Approximately 10 percent of the responses are scored twice as a check for reliability. Scoring consistency is evaluated through records of exact and adjacent agreement between rater pairs.

Scaling and Calibration

Student scale scores are based on the combined results of the three sections. (There is no Terra Nova component in Health and Physical Education or Fine Arts.) For all assessments, a student's item scores on the multiple-choice and rater-scored items are used together to compute a scale score, which is then used to determine an achievement level. In Communication Arts, Mathematics, Science and Social Studies, the Terra Nova items are combined with "Missouri" items to obtain the scale score. Item Response Theory methodology is used to arrive at the overall scale score. The models used are the three-parameter logistic for multiple-choice items and the two-parameter partial credit for the rater-scored items. In any given content area, forms are calibrated to the same proficiency score scale using a common-items model.

The scale used for calibrating the scores is the Terra Nova score scale. All Terra Nova items have parameter estimates based on national norms, and all Missouri item parameters are placed on the same scale through field testing them as embedded items on booklets spiraled within classes on both the constructed-response and performance-event sections. The Missouri scale scores increase with grade just as the Terra Nova scale scores do. Each year's MAP tests are developed and scored using the existing item parameters from prior data.

Achievement Levels

There are five achievement levels for each tested area. These are Step 1, Progressing, Nearing Proficiency, Proficient and Advanced. For each student, the level is determined by comparing the student's scale score with the four cut points that differentiate the levels.

The cut points for any one subject and grade level were set at a conference jointly organized by the contractor and DESE. Participants at the conference were representatives from the education, professional and business communities selected through a statewide nomination process involving school superintendents, mayors and business leaders. The panels implemented the Bookmark procedure.

Following cut-point determination, the panels wrote full and condensed achievement-level descriptors. The full descriptors are published in teachers' guides, and the condensed descriptors appear on score reports.

Accommodations, Exemptions and MAP-A

A student who receives special educational services has an Individualized Education Program (IEP) that is developed by a team and describes how the student is to be taught (e.g., what instructional services the student receives). Such a student takes either the MAP with or without accommodations as indicated by the IEP, or is exempted from the MAP and takes the MAP-Alternate Assessment (MAP-A) as per the IEP.

The MAP-A is a portfolio system that is appropriate for up to two percent of a district's students (i.e., the lowest functioning students). The student's IEP is used to determine which assessment (MAP or MAP-A) the student will take.

DESE does not monitor these decisions on a case-by-case basis, but the data on proportions of students tested using MAP-A are examined as part of the MSIP review process. MAP-A results will eventually be reported separately but not used in the accountability system. At this time, no MAP-A data have been reported.

Operationally, MAP-A is provided to districts with an expectation that they will use it to assess students who are not assessed using the MAP because they exhibit the lowest levels of cognitive functioning. Scoring will be done by the state, and results will be reported in terms of meeting goals selected for portfolio assessment (which should be explicitly associated with the Show-Me Standards), but no accountability expectations for these data are planned for the near future.

Level Not Determined

A student is placed in Level Not Determined (LND) when he or she does not complete three valid sessions on the MAP Communication Arts, Mathematics, Science, or Social Studies assessments, or, in the cases of Health and Physical Education and Fine Arts, two valid sessions. Being absent for or refusing sessions will cause students to be assigned to LND, as will being exempt for one year (in-state) due to Limited English Proficient (LEP) status. Students who do not attempt one item in session 1 and one item in session 2 and (for Communication Arts, Mathematics, Science, or Social Studies assessments) five items in session 3 will be assigned to LND. In addition, students who take MAP-A are counted as LND on MAP.

Reporting Schedule and Formats

MAP results are reported each year in the late summer. For an individual student, the contractor reports: achievement level (there are five levels: Step 1, Progressing, Nearing Proficiency, Proficient, Advanced) for each content; scale score for each content (which determines achievement level); percent of items answered correctly for each content and process standard measured; and a national percentile based on the Terra Nova items. Achievement level summaries (percent of students in each level), scale score, content/process standard data, and percentile results are also reported for buildings and districts, and disaggregated results are reported as well. The contractor sends hard copy reports for all of the above. Student-level information is released only to the student's district.

Individual student reports are distributed for each test. The level of the student's scale score is compared graphically with the five achievement levels along with the level descriptions. Within each of the Show-Me Standards tested, separately for content and process standards, the number of possible points and the percent of points earned by the student are presented. The Terra Nova national percentile on that section of the MAP, if available, is also given. On the back of the report is a message to parents about how to interpret the scores.

Each school receives a report showing the number and percent of students at each achievement level and at Level Not Determined. For the Terra Nova portion, the median national percentile and the national percentile of the mean normal curve equivalent are given. A graphic display shows, for each of the content standards, the statewide mean percent correct and the school mean percent correct along with its confidence band (the mean plus and minus a multiple of its standard error; the multiple is usually calculated so that there is a certain probability, called the confidence level, that the true school mean is spanned by the band; however, the confidence level used is not given in the documentation). A table of disaggregated reports shows the numbers in each Achievement Level, combined Levels 1 & 2, combined Levels 4 & 5, and the median Terra Nova national percentile. Disaggregations are by gender, race/ethnicity, and special programs and are only tabulated when five or more students are in the category. Individual student reports and a brief summary report complete the building-level products.

Districts receive summaries that parallel the school-level reports. A Guide to Test Interpretation developed by CTB/McGraw-Hill and written for teachers and administrators is available. DESE has plans to revise this guide in the near future.

Districts also receive their results on a CD ROM that can be analyzed using a software program called Clear Access. Districts usually receive their Clear Access data in September. This program allows districts to produce detailed reports customized to meet particular needs. The data display student responses to individual items; the items have short descriptions attached to them. Concern has been expressed about the degree to which the descriptions are representative of the items as well as the degree to which individual student responses to items are helpful in any sort of decision making.

Score reporting to the public for the state as a whole and for districts is in terms of percents of students in the five achievement levels across years. For ease of interpretation, the top two category percents are summed, as are the bottom two in these reports. These data and other pertinent statistics, such as the median percentile rank for the Terra Nova portion of the test, are freely available (e.g., they appear on the DESE Web site at www.dese.state.mo.us/schooldata/direct.html).

Interpretation for District Accountability

Districts have been expected to make progress in moving students into the top two achievement levels and not to have more than about 2 percent of students in the LND category. The percents of students in the five categories are calculated based on the number of students tested (i.e., they sum to 100 percent). This excludes the students who were in the LND category (which may result from absences, refusals to be tested, etc.) as well as those who were exempted (and presumably have taken the MAP-A).

If a student is not exempted, then that student is eligible to take the MAP. Table 1 shows for each test the numbers and percents of students who were tested, were exempt, and were eligible to be tested but for one reason or another did not receive a score in 2000. The latter event is called "refused" in the table, but may be due to absence as well as refusal to sit for or to complete the entire test. All entries in this table were derived from three figures: the number tested and the percent tested were taken from the Web site, and number exempt was taken from a communication from DESE.

Table 1. Tested, Exempted, and Refused Patterns in 2000

Test and Grade	Enrollment	Number Tested	Number Refused	Number Exempt	Percent Tested	Percent Refused	Percent Exempt
Communication Arts - 3	71263	69638	1108	517	97.72	1.55	0.73
Science - 3	71181	69928	764	489	98.24	1.07	0.69
Mathematics - 4	70649	69554	581	514	98.45	0.82	0.73
Social Studies - 4	70620	69441	655	524	98.33	0.93	0.74
Communication Arts - 7	68790	66713	1634	443	96.98	2.38	0.64
Science - 7	68920	67121	1369	430	97.39	1.99	0.62
Mathematics - 8	69558	67527	1481	550	97.08	2.13	0.79
Social Studies - 8	69433	67364	1557	512	97.02	2.24	0.74
Mathematics - 10	62897	59979	2528	390	95.36	4.02	0.62
Science - 10	62963	59922	2682	359	95.17	4.26	0.57
Communication Arts - 11	56744	53396	3023	325	94.10	5.33	0.57
Social Studies - 11	56732	54105	2319	308	95.37	4.09	0.54

Note: A student is called "refused" if he or she was eligible to be tested but was not.

A draft section in the March 21, 2001, Procedures Handbook describes how accreditation is to be determined after July 1, 2001. According to the draft, a district must have the needed number of points in a combined Resources, Processes and Performance rating as well as the

required number of points in Performance to achieve accreditation. In K-12 districts, a total of 74 Resource and Process points and a total of 126 Performance points are available.

The fundamental elements of performance scoring are a MAP Index for each tested subject for each grade span (3-5, 6-8, 9-11). A MAP Index is calculated by multiplying the percentages of students in each of the five achievement levels by graduated point values. Performance Points are awarded by evaluating the district's MAP Index in each grade span according to each of four methods of calculation, and then using the method that yields the most points. The four methods are High (points for index above stated thresholds for stated numbers of recent years), Yearly Increase (points for recent yearly increases of at least four index points), Multiple-Year Average Over Base Year (points awarded if the average index over recent years is a stated threshold above the base year), and Rolling Average (points awarded for yearly increases of at least four index points in two-year rolling averages). No points are awarded in any subject if the LND percentage is greater than 10 percent. Thresholds exist for subjects and grades below which districts may not receive full points. A system of partial points based on the same four calculation methods also exists.

Technical Advice

The MAP has a relatively stable technical advisory panel consisting of nationally known outside professionals with psychometric and statistical expertise as well as practitioners who have experience using accountability systems.

Technical Quality

The assessment contractor, CTB/McGraw-Hill, has developed several reports to describe the psychometric characteristics of the various tests. There are three series: (1) documentation of achievement-level setting (using the "bookmark" method) separately for each content and grade level; (2) item-by-item tables of inter-rater agreement (perfect and adjacent) for constructed-response items; and (3) technical reports covering reliability and validity.

Reliability is documented through internal homogeneity coefficients and standard error plots across scale score values. The alpha coefficients are generally above .90, with a few exceptions. Standard errors are correspondingly low in the ranges where most students are expected to score (and near the scale-score cut points for achievement levels). Simulations to estimate classification consistency (into the five achievement levels) were performed and show that the standard errors do not imply unacceptable inconsistency.

Validity is addressed through content sampling designs, item fit, fairness and consequential benefits. The sampling designs identify the relationship between the test items and the Show-Me Standards. They appear in the technical manuals.

A statistical process that compares observed with model-based proportions of examinees in certain ability ranges earning the various scores is used to evaluate item fit. Items showing misfit are flagged. This is a standard procedure that indicates whether the statistical model being used is appropriate for the items that are on a test. Of the 13 tests given in 2000, all had 15 percent or fewer poorly fitted items except for eighth-grade math (24 percent), third-grade

science (21 percent), fourth-grade social studies (30 percent) and eleventh-grade social studies (20 percent). These percentages do not seem excessive and thus suggest that the psychometric modeling being used in the program is reasonable.

Fairness for demographic groups is evaluated at the item level using a statistical procedure that compares the group's observed scores with model-based predictions based on all examinees. The resulting DIF (differential item functioning) statistic is compared with a flagging criterion for each demographic group studied. Five DIF analyses, for five different groups, were performed for each item: Caucasian, African-American, Hispanic, Asian, female and male. Flagged items are referred to a follow-up committee for evidence of biased content. Very few items have been flagged, and no content bias has been found on follow-up.

At the total score level, fairness is evaluated by subgroup comparisons. The mean of each demographic group is compared with the score that is one standard deviation below the "majority group" mean. Of these, the African-American population scored below that criterion on tenth-grade mathematics, seventh-grade science, tenth-grade science, eighth-grade social studies and ninth-grade physical education. All other means across groups and tests were above the criterion. The program seems to have sufficient evidence to support its fairness.

Consequential benefits on classroom practices are being studied at the Center for Learning, Evaluation and Assessment Research at the University of Missouri-Columbia. This is a longitudinal study that analyzes statements of teacher respondents about their classroom practices and attitudes.

The Center has continued to collect similar data. A report will be available at the approximate time that the present report is distributed. In addition, the Center is collecting survey data on classroom practices in health and physical education and is evaluating videotaped classes according to a coding scheme used by the Third International Mathematics and Science Study (TIMSS).

EVALUATION

This section focuses first on the questions that prompted the evaluation study, followed by additional issues that were identified as the study progressed. Recommendations are in bold face in this section and, for convenience, are repeated in a summary at the end as well as in the executive summary.

Content and Performance Standards

How is the domain of each assessment in the program described? Do these products clarify effectively what is to be assessed?

The greatest degrees of specificity for domain descriptions are found in the frameworks and assessment annotations. However, these are not very specific. To continue the example from the previous section on Content Domains and Communication of Domains, neither the knowledge statement "strategies to collect data" nor the performance (process) statement "explore concepts of chance" documents for a curriculum writer or a teacher (or a test writer) what might and what might not appear on the assessment. As a consequence, whether students are tested on what they have been taught is to some degree a matter of whether teachers and test writers have the same understanding of the frameworks. A way to describe what is "fair game" for the assessments would be beneficial. A possible way to do that would be to develop "assessment limits" for each of the elements in the frameworks and then to show where and how each will appear (or be sampled) in the assessments. A "test map" showing how the framework elements appear on the tests could then describe each assessment for curriculum writers and for teachers. (The current mapping of the items to the Show-Me Standards in the Technical Manual is too broad to be helpful). Taken together, the assessment limits would become an "at-most list" for assessment developers and an "at-least list" for teachers. If this were done well, there would be no surprises when the tests appear since everyone would have the same understandings about what they will cover.

Recommendation: Develop assessment limits (specifications for the content that can be tested) for each sub-domain description in the state's curriculum frameworks and describe where and how each sub-domain is assessed in the MAP through test maps (i.e., blueprints) that are useful both to educators in planning instruction as well as to test developers in creating items and other test prompts. (The supplement to the curriculum frameworks provides a good start in mathematics.)

What methods are used to communicate the domain to local education professionals? Are performance standards communicated effectively?

The curriculum frameworks have been developed to communicate domain descriptions. Several process standards are left to local districts to assess, and these are described in the accompanying assessment annotations. In addition, the booklets of released constructed-response and performance-event items with scoring guides should be helpful to local professionals. Cutoff scores for the achievement levels were set using accepted procedures and descriptions of them exist in both full and condensed versions. It would be helpful to use

existing data to amplify these descriptions with samples of student work that is rated in the various performance levels (e.g., each anchor paper in a released scoring guide could be associated with the achievement level of the scale score at which the probability of its rating or better is, say, two-thirds). That would enable the use of actual student work to help understand what students at the various achievement levels know and are able to do.

Recommendation: Add clarity to the released scoring guides by identifying the achievement level (or scale score) that each anchor paper (example of student work) represents.

Alignment

How is the domain of each assessment sampled? Do the assessments remain consistent from year to year in their domain coverage?

While the curriculum frameworks and assessment annotations indicate which knowledge and process standards are sampled on the state assessments, the only map that is available to show the sampling is at the Show-Me Standard level, which seems too broad to be useful to local personnel. It is not possible to evaluate alignment more specifically. The map described in the discussion on the previous question could provide the needed detail.

How are the demands of the assessments chosen or developed to be appropriate for the ages and grades at which they are used?

Items are developed by practicing teachers and are pilot-tested and field-tested before actual use in decision making. Therefore, they should not be very different from what at least some students are asked to do in Missouri classrooms. Evaluating the procedures used in pilot testing is beyond the scope of this review, but it would be helpful in a further study to consider at least two features beyond traditional item-screening analyses. First, are data available from the pilots that show at least some students are able to perform at the highest levels on every item that survives into field-testing? Second, are anecdotal data collected from both the pilots and the field tests? If so, how are they used in item revisions?

Recommendation: Ensure that data from pilot and field testing are used in the revision of MAP assessments. Performance data should include frequencies and samples of student responses at all scoring-tool anchor points, especially the highest levels. Anecdotal data related to ease of administration and item validity should be collected from teachers as they administer the items.

Technical Aspects

What is the technical quality of the assessments? Does their reliability support their intended uses?

Reliability studies of the MAP assessments have generally shown coefficients to be at least .90. This level is sufficient for individual interpretation.

Is the validity evidence compelling? Are there important areas where validity evidence is missing (e.g., content evidence, convergent evidence, divergent evidence, consequential evidence, fairness)?

Validity is an area in which there is never enough evidence for any assessment program. However, as the assessments are used, they naturally generate multiple opportunities to study validity. The program appears to be generating appropriate evidence from a reasonable range of perspectives. These include content evidence (although more detailed assessment maps would be helpful), statistical evaluations of item fit and fairness, and commissioned studies of consequences. The one area in which there seems to be little evidence has to do with convergence (agreement with other achievement measures). Studies relating MAP performance with such easily obtained data as teacher assessments or local tests would enhance the validity evidence available to the program.

Recommendation: Conduct correlational studies relating MAP results to existing data, such as teacher grades and local tests, to provide evidence that scores resulting from MAP are consistent with those from other ways of measuring the same achievements.

How have performance standards been developed?

Performance standards for students were developed using the Bookmark procedure, a well-respected technique, and are well documented. The groups of people sampled appropriately included both educators and other members of the community who have important interests in education outcomes.

It is less well documented how performance standards for districts (or schools) were developed. Compared with those in other states, they seem unlikely to be understood well by the public at large. In particular, the multiple ways that exist to generate district performance scores (see the last paragraphs under Interpretation for District Accountability in the Overview) is a highly unusual feature and should be reviewed to determine whether it is necessary. At present, it appears without rationale. (For further discussion about use of district performance scores, see the discussion in the Consequences section below.)

Administration

To whom are the assessments given? Are they given under standardized conditions? Are inclusion and accommodation policies administered fairly?

Assessments are administered according to procedure, examiner and test coordinator manuals. These are well detailed and appear to be sufficient to ensure reasonable standardization. Inclusion and accommodation are both addressed.

There does not seem to be much, if any, monitoring of how carefully schools implement the procedures in the manuals. It would be reassuring to those who adhere to the specified procedures to know that violations are likely to be identified and addressed.

Recommendation: Monitor districts' and schools' administration of the MAP assessments to ensure that they comply with the procedures detailed in administration manuals, including procedures related to accommodations and inclusion.

Are accommodations effective?

The accommodations are based on studies conducted by the National Center on Educational Outcomes at the University of Minnesota. This is an active and highly respected organization that can be relied on for informed recommendations.

Scoring

How adequately are the assessments scored? What training do scorers receive? How is scoring quality maintained throughout the project?

The Terra Nova portion of the MAP is machine scored using CTB/McGraw-Hill protocols. The other two sections require hand scoring. This is implemented by CTB/McGraw-Hill and involves Missouri teachers. Although most papers are read only once, the contractor implements standard quality-control checks during the process. Scoring seems to be done according to accepted industry procedures.

Consequences

What are the information sources and procedures used to evaluate programs (e.g., schools) or individuals (e.g., students)? What are the consequences of these evaluations?

While individual students receive scores, there are no direct consequences for students. Districts are evaluated through MSIP on several factors, grouped within three general areas: resources, processes and student performance. All but a small number of districts receive accreditation.

How much emphasis is given to non-achievement factors?

In the three general areas considered by MSIP, points are accumulated and deducted based on criteria with differing degrees of objectivity. Districts obtain or do not obtain accreditation based on the points they receive, but more points and greater emphasis in the decision-making process is given to student performance. While the resource and process areas can lead to a decision not to accredit, no district can be accredited with a performance rating below the criterion. On the other hand, one can reasonably expect that better resources and improved procedures exist to foster student success. Diluting student performance, the

assessment of that success, with other factors in accreditation decisions can be questioned and therefore should be justified.

Recommendation: Monitor the success of districts with as broad an array of standardized student achievement indicators as possible—including such criteria as attendance and dropout rates—and, at the same time, minimize the impact of non-achievement factors in making decisions about district and school success.

To what extent are the procedures used statistically justified?

Particularly in the performance area, the accreditation procedures are quite complicated and exist in draft form only. There are four ways to calculate the points earned by the district and two approaches. The justification for these different ways to arrive at an index is not in the handbook. Why do they exist? What does each encourage and discourage a district, a school or a teacher to do or not do? Will one or two of them be the only ones that are actually used for the great majority of districts? These are fair questions that should be addressed in the near future. Leaving these questions open invites distrust through lack of clarity in both intent and implementation.

In summary, there are two ways in which the effects of MAP scores upon district improvement are compromised. First, points for accreditation are based to some extent on other-than-achievement indicators. That substitutes process for product. Second, there are multiple options for combining MAP results for interpretation. These issues raise questions about the most effective uses of MAP data for school improvement. The present system may leave districts and thus schools and teachers unclear about what their targets should be.

Further, that the present system is in draft form leaves districts and schools unclear about how they will be judged in the future. The accreditation criteria should be finalized so that educators have stable guidance in setting instructional-performance targets.

Recommendation: Reconsider the advisability of using multiple ways to generate district performance scores for accreditation. This is a highly unusual feature of MAP that is not commonly used in other states. This feature leads to different criteria for different districts and schools, is not likely to be understood by educators or the public, and is not conducive to setting effective district and school educational goals.

Recommendation: If the state decides to continue using multiple criteria, then it should develop and include in a finalized version of the MSIP scoring guide a justification for the different ways in which districts are awarded points for student performance. Studies to evaluate the consequences of these alternatives should be initiated as data accumulate.

Reporting

What reports are disseminated? Do they contain the most useful information?

Student reports contain both norm- and criterion-referenced interpretations. All seem reasonable with the possible exception of the standard-based percent of items answered correctly. This is not comparable across different samples of items and can lead to misleading interpretations across time and between standards. Interpretation is further complicated since items are allocated to both content and process standards and sometimes even to more than one content standard and more than one process standard. It would probably be better not to report the percent-of-items scores.

Recommendation: Discontinue reporting percents of items answered correctly for each content and process standard because differences in difficulty between the item pools have unknown impacts upon those percents.

Student scale-score reports currently are displayed graphically as point values against the five achievement levels. It would be helpful to include confidence bands based on the appropriate standard errors of measurement. Each confidence band might span the range of the student's score plus and minus one standard error of estimate, with the interpretation that the probability of the band spanning the score the student actually deserves is two-thirds. School-level and district-level reports show appropriate levels of aggregation and are available electronically as well.

Recommendation: Include confidence bands in reporting individual-student scale scores to convey the degree to which random factors contribute to score imprecision and to help educators avoid over-interpreting small score differences.

Are they received when they are most useful?

Reports are received in the late summer. This is reasonably quick since the constructed-response portions of the tests must be scored and the data scaled along with the selected-response portion in each subject area. The data are available when needed for accountability. They are also available in reasonable time to be used to make school-wide educational program decisions. They are not available when needed if routine instructional decisions about students in that school year are to be based on them. However, local assessments that are tailored to the local curriculum and designed for diagnosis are typically preferable to accountability assessments for that purpose.

Are they valuable for school improvement? Do they contain enough information to be as useful as possible for teachers?

Both teachers and schools need data to help them evaluate their instructional practices and emphases. At present, the data seem to support this only at the content-area level. Within contents, grouping items into instructionally relevant categories (perhaps these would be by content and process standards) and reporting scale scores (generated by scoring students for each item group with all other items "turned off") accordingly would provide data helpful to schools

and teachers in evaluating their instruction. Such data should not be reported at the individual student level. Since it would be generated for groups of students, however, it should be reliable enough for programmatic decision making at the school level. An appropriate format for reporting might be box-and-whisker plots by year showing trends over time. (A box-and-whisker plot is a visual display of percentiles, commonly the fifth, twenty-fifth, fiftieth, seventy-fifth and ninety-fifth. It is used to compare the locations of groups of students throughout their ranges. One can study central trends by focusing on the median and the quartiles, while examining extremes to see whether the least successful students are nevertheless being reached and if the most successful students are being challenged.)

Recommendation: Group items in each content area tested by MAP into instructionally relevant categories and report scale-score summaries on these item categories for groups of students, such as classrooms, schools, districts and the state, so that decision making about curriculum and instructional programs can be informed by meaningful student-outcome results.

Additional Issues

Terminology

The use of the term "performance standard" usually communicates a degree of achievement that meets some criterion. Missouri calls these "achievement levels." The Performance Standards in the Show-Me Standards are usually called "process standards" elsewhere. In communicating with others, especially outside the state, Missouri's unusual use of these terms should be emphasized to avoid confusion.

Recommendation: For clarity of language, communicate to educators who are used to other states' terminology that Missouri's "Achievement Levels" are what others commonly call "performance standards" and that Missouri's "Performance Standards" are what others commonly call "process standards."

District and School Targets

It would be helpful for districts and schools to have targets for their students' performance. For example, setting goals that all districts and schools should be expected to reach in terms, say, of percents of students at the top three and at the top two achievement levels would be useful. In this way, achievement expectations that are consistent for all schools and districts can be communicated to the public and easily compared with district and school performance in terms that already exist through data formats currently used on the state's Web site. Further, having school-performance targets can facilitate the setting of adequate yearly progress goals toward those targets. Several ways to set adequate yearly progress levels have been suggested. For example, one approach is to divide the difference between current- and target-performance levels by the number of years it should take for all districts or schools in the state to be at the targeted levels. Another approach is to tabulate the actual progress achieved by similar districts or schools and to set adequate yearly progress goals at some level (e.g., the 75th percentile) of actual progress achieved in the peer group. Yet another method is to involve

stakeholders such as parents and community leaders along with educators in setting adequate progress levels on a yearly basis.

Recommendation: To measure districts and schools against the same standards, establish and communicate school and district achievement targets, both for overall performance and for adequate yearly progress, based on percents of students at and above selected achievement levels for schools and districts.

Motivation

The MAP tests presently have few consequences for students. As they become older, some students show, or at least are commonly thought to show, less motivation to succeed on tests that have no or few implications for them. Attaching consequences, especially for older students, should be considered. They might range from "mild," such as showing the last grade level score for each tested content on student transcripts, to "strong," such as including test scores in promotion and/or graduation decision making.

Recommendation: Consider adding student incentives for test performance, especially for older students, to increase students' motivation to perform well on MAP assessments.

Scorability

It is not clear at present how blanks and non-readable responses are handled in scoring. Because these items provide no information about the student, a case could be made for ignoring them. On the other hand, the student had a chance to attempt to receive credit and did not (or could not) do so. That would support a case for a score of "zero" being assigned to the item for that student and then included in scoring. The latter approach is recommended here. In any event, what is actually done should be documented. It should also be disseminated since it affects how students can maximize their scores and therefore affects their test-taking strategies.

Recommendation: Develop and disseminate the rules that are used to score MAP items so that parents, students and school personnel clearly understand how all student responses, including blanks and non-readable responses, are used in determining student test scores.

Process Sub-scores

The value of reporting process sub-scores should be revisited. Process skills may not transfer from application to application very well. If it is decided that process scales are valuable, then describing and assessing them within some taxonomy, such as procedural and meta-cognitive knowledge categories of content knowledge, instead of as generalized cognitive processes, should be considered.

Recommendation: Discontinue reporting process sub-scores (i.e., scores for thinking skills and processes). They may not generalize well from one content area to another in terms of either what students are asked to do or their success in doing it.

Level Not Determined

The LND status is assigned for a broad range of students. LND is used for students who are absent, who do not complete enough items, who are exempt due to limited-English proficient (LEP) status, or who take the MAP-A. But these conditions have different implications for programs. Most would agree that schools should be held accountable for the achievement of students whether or not they are absent for or complete only a few items on tests. It is actually to a district's advantage in the current system to encourage absence or non-completion for the lowest performing students. On the other hand, schools should not be held accountable for LEP students until they have had a chance to address the students' English learning. And students who take the MAP-A are not pursuing the same outcomes as students for whom the MAP is appropriate. Other approaches to students in these groups should be considered. For example, counting students who are absent or do not complete sufficient items as Step 1 for accountability purposes would encourage districts to provide evidence for all students. It could also allow districts and schools to allow such students to be "excused" from an assessment if they would likely become frustrated and not be able to perform (such a judgment could be made at the level of the school). Students who have initial LEP status and who take the MAP-A could be exempted from the MAP for accountability purposes. Programs for MAP-A students could be evaluated separately using results from that assessment system, since they have different learning goals. There would be some implications for program accreditation decisions. There would no longer be a need to have a threshold percent for LND, and decision criteria would need to reflect revised goals since all, as opposed to only some, eligible students would be included in the accountability percentages. Criteria would need to be developed for MAP-A results.

Recommendation: Incorporate the Level Not Determined (LND) category into district- and school-accountability decisions since the mission of districts and schools is to educate all students, not just those who are tested.

MAP-A

Scoring and reporting mechanisms need to be developed and evaluated for MAP-A. Whether or not MAP-A students are included in the LND category (see previous recommendation), programs that are designed for these students represent expenditures of public funds, and thus student achievement on MAP-A should be part of the accountability program. The scoring and reporting mechanisms eventually developed should be evaluated against criteria for accountability decisions.

Recommendation: Develop scoring and reporting systems for MAP-A (an alternate portfolio system used for the lowest-functioning students) that support accountability. These systems should assess the education goals that students in these programs are pursuing and provide sound data that can form the basis for making outcomes-based judgments about program success.

High-stakes Use of the Terra Nova

The items on the Terra Nova are the same year after year. To the extent that the results are important for high-stakes decisions (e.g., district accreditation), re-use of items can cause a bias in sampling of content that is predictable and can affect what schools do instructionally. It would be preferable to generate statewide pools of selected-response items that can be sampled year to year, with release and augmentation programs as Missouri presently does with its

constructed-response sections. Because these pools of new items could be calibrated into existing items using well-known psychometric methods for the models currently in use, it would be feasible to continue using the present scale for student achievement-level decisions.

Recommendation: Replace the Terra Nova with sampling of statewide item pools as is done on the other two portions of the MAP to avoid the possibility that teachers are teaching to the specific items used on accountability measures.

SUMMARY OF RECOMMENDATIONS

All of the recommendations identified in the previous section are collected here and grouped according to four general themes: Clarification of Standards, Technical Considerations, Fairness and Consequences, and Reporting Results.

Clarification of Standards

Two types of standards are commonly used in an assessment program. These are content standards (what is tested) and performance standards (how well students are expected to do). These first recommendations are intended to clarify, especially for Missouri educators, precisely what Missouri's standards are.

1. For clarity of language, communicate to educators who are used to other states' terminology that Missouri's "Achievement Levels" are what others commonly call "performance standards" and that Missouri's "Performance Standards" are what others commonly call "process standards."

2. Develop assessment limits (specifications for the content that can be tested) for each sub-domain description in the state's curriculum frameworks, and describe where and how each sub-domain is assessed in the MAP program through test maps (i.e., blueprints) that are useful both to educators in planning instruction as well as to test developers in creating items and other test prompts. (The supplement to the curriculum frameworks provides a good start in mathematics.)

3. Add clarity to the released scoring guides by identifying the achievement level (or scale score) that each anchor paper (example of student work) represents.

Technical Considerations

Both assessment systems (how data are generated) and accountability systems (how data are used) should conform to rigorous professional standards. Especially important in a statewide program are the quality of the data for purposes of making decisions about students, districts and schools and the effects of the program on improvement of educational practices throughout the state. The purpose of this set of recommendations is to further these quality goals as well as to enhance the evidence about the degree to which they are met.

4. Replace the Terra Nova with sampling of statewide item pools as is done on the other two portions of the MAP to avoid the possibility that teachers are teaching to the specific items used on accountability measures.

5. Ensure that data from pilot and field testing are used in the revision of MAP assessments. Performance data should include frequencies and samples of student responses at all scoring-tool anchor points, especially the highest levels. Anecdotal data related to ease

of administration and item validity should be collected from teachers as they administer the items.

6. Develop and disseminate the rules that are used to score MAP items so that parents, students and school personnel clearly understand how all student responses, including blanks and non-readable responses, are used in determining student test scores.

7. Conduct correlational studies relating MAP results to existing data, such as teacher grades and local tests, to provide evidence that scores resulting from MAP are consistent with those from other ways of measuring the same achievements.

8. Reconsider the advisability of using multiple ways to generate district performance scores for accreditation. This is a highly unusual feature of MAP that is not commonly used in other states. This feature leads to different criteria for different districts and schools, is not likely to be understood by educators or the public, and is not conducive to setting effective district and school educational goals.

9. If the state decides to continue using multiple criteria, then it should develop and include in a finalized version of the MSIP scoring guide a justification for the different ways in which districts are awarded points for student performance. Studies to evaluate the consequences of these alternatives should be initiated as data accumulate.

10. Develop scoring and reporting systems for MAP-A (an alternate portfolio system used for the lowest-functioning students) that support accountability. These systems should assess the educational goals that students in these programs are pursuing and provide sound data that can form the basis for making outcomes-based judgments about program success.

Fairness and Consequences

Students and schools should be motivated to score well on assessments that are fair to everyone. To be fair, assessment results should have the same meaning for all. For example, they should be administered under the same conditions for everyone (i.e., be standardized). Schools should be motivated to educate toward and students should be motivated to demonstrate the highest possible levels of achievement. The recommendations in this section are oriented toward increasing motivation for students and schools, and toward ensuring that the assessments are as fair as possible.

11. Consider adding student incentives for test performance, especially for older students, to increase students' motivation to perform well on MAP assessments.

12. Monitor districts' and schools' administration of the MAP assessments to ensure that they comply with the procedures detailed in administration manuals, including procedures related to accommodations and inclusion.

13. To measure districts and schools against the same standards, establish and communicate school and district achievement targets, both for overall performance and for adequate

yearly progress, based on percents of students at and above selected achievement levels for schools and districts.

14. Incorporate the Level Not Determined (LND) category into district- and school-accountability decisions since the mission of districts and schools is to educate all students, not just those who are tested.

Reporting Results

Assessment results are most valuable when users understand them. Fundamental to effective use is knowing what content is being assessed and how to use the information for improvement. Scores should be available on education outcomes that cover the spectrum of important goals, but at the same time are focused enough to lead to informed educational change. Finally, the scores should not be over-interpreted beyond the quality of the data. These final recommendations are intended to enhance the impact of Missouri assessments upon schooling through the ways they are reported.

15. Monitor the success of districts with as broad an array of standardized student achievement indicators as possible—including such criteria as attendance and dropout rates—and, at the same time, minimize the impact of non-achievement factors in making decisions about district and school success.

16. Include confidence bands in reporting individual-student scale scores to convey the degree to which random factors contribute to score imprecision and to help educators avoid over-interpreting small score differences.

17. Group items in each content area tested by MAP into instructionally relevant categories and report scale-score summaries on these item categories for groups of students, such as classrooms, schools, districts and the state, so that decision making about curriculum and instructional programs can be informed by meaningful student-outcome results.

18. Discontinue reporting percents of items answered correctly for each content and process standard because differences in difficulty between the item pools have unknown impacts upon those percents.

19. Discontinue reporting process sub-scores (i.e., scores for thinking skills and processes). They may not generalize well from one content area to another in terms of either what students are asked to do or their success in doing it.

ABOUT THE AUTHOR

William D. Schafer, Ed.D., is professor emeritus and affiliated professor with the Maryland Assessment Research Center for Education Success, University of Maryland-College Park. Prior to his work there, he served as state director of student assessment for the Maryland State Department of Education (1997-1999) and as a faculty member in the Department of Measurement, Statistics and Evaluation at the University of Maryland (1969-2000). He began his career in education as a math teacher at West Irondequoit School District, New York (1965-67).

Schafer earned a doctorate in measurement, statistics and evaluation from the University of Rochester in 1969. He has authored numerous journal articles and book chapters on assessment throughout his career. Most recently, he and R.W. Lissitz edited *Assessments in Educational Reform* (2002, Boston: Allyn & Bacon). He currently co-edits *Practical Assessment, Research & Evaluation*, an electronic journal sponsored by ERIC (Educational Resources Information Center).

Schafer continues to consult and present on assessment issues throughout the country, including recent work with the U.S. Department of Education and several states. He can be contacted by e-mail at ws7@umail.umd.edu.

APPENDIX

Acronyms Used in This Report

DESE	Department of Elementary and Secondary Education
DIF	Differential Item Functioning
IEP	Individualized Education Program
LEP	Limited-English Proficient
LND	Level Not Determined
MAP	Missouri Assessment Program
MAP-A	Missouri Assessment Program-Alternate Form
MNEA	Missouri National Education Association
MSIP	Missouri School Improvement Program

Missouri National Education Association
1810 East Elm Street
Jefferson City, MO 65101
www.mnea.org