Report of the Large-Scale Assessment Task Force
March 7, 2006

Cathy Brown
Michael Brown
Steve Leinwand
Linda Dager Wilson (chair)

I. Background

NCTM's vision of the role and characteristics of large-scale assessment that supports high-quality teaching and high levels of student performance in mathematics was initially described in the Council's 1989 *Curriculum and Evaluation Standards* and further fleshed out in the 1995 *Assessment Standards*. We envisioned a system in which there is a deliberate alignment among clear learning expectations, instructional practices, and high-quality assessments. We argued that in such an aligned system, accountability was but one part of a system where assessments were also used for program evaluation and improvement and to help design appropriately targeted interventions for students. Moreover, we advocated for a system of assessments that was, simply stated, of high quality, useful, meaningful, coherent, and fair.

Much has been accomplished since 1989 to move us toward this vision:

- Clearer and more specific curriculum standards have been developed at the state and district levels.
- State-level assessments have been more closely linked to these standards.
- State-level assessments began to include a greater percentage of items designed to assess NCTM's Process standards in addition to the content standards.
- There has been significant research and development, as well as professional development, in the area of alternative assessment practices.
- NCTM has published excellent materials that provide support and examples of more effective classroom assessment practices.
- Assessment data are now, by law, widely disaggregated by subgroups.

In the past few years the No Child Left Behind (NCLB) Act has, by design, raised the stakes. More students, teachers, schools, districts and states are rightfully being held accountable for student achievement. But the basis of this accountability is a set of high-stakes large-scale assessments that are having an increasing powerful—and not always positive—impact on what is taught and how it is taught. Moreover, NCLB has engendered a set of pressures and activities not entirely aligned with NCTM's visions. This high-stakes, test-driven environment has, in fact, revealed a broad array of practices, policies, and consequences that widen the gap between standard practice and NCTM's vision:

- The costly shift to more frequent testing has resulted in a reduction in the number of constructed response items included on the assessments and the number of items released after each administration.
- The pressure to raise test scores has resulted in far more instructional attention being paid to students just below cut scores than those far below these scores.
- The format and quality of the tests (and the standards to which they are more or less aligned) vary tremendously from state to state resulting in a significant diversity of opportunity across the country.
- Many of the results are reported in ways that provide teachers and parents with little useful information about how to interpret and productively use the results.
- Time devoted to testing and test preparation increasingly comes at the expense of instruction.
- Curriculum and instruction in some states are narrowing to that which is most easily testable in multiple-choice format.

As a result, current realities in states, districts, schools, and classrooms often fall well short of our vision and raise a number of important challenges that NCTM and its members face. To address this situation, this report describes what the task force views as the five core challenges, reviews current initiatives, both internal and external, and then proposes a set of recommendations for consideration by the NCTM Board of Directors.

II. Challenges

Challenge 1: Alignment

The original NCTM vision called for alignment among learning expectations, instructional practices, and assessments. When these components of education are not aligned, there are serious consequences. The assessments may test ideas, concepts, or skills that are not being taught. Instruction may focus on facts or skills and omit such important mathematical notions as grappling with solutions that are not obvious, communicating mathematical ideas, making connections, representing mathematical ideas in multiple ways, discovering effective approaches to solve problems, or testing the reasonableness of a solution. For alternate assessments, there are issues of whether the tests are aligned to content standards of age peers for those students or to instructional standards for students of comparable ability. Accommodations that are meant to allow greater access to more students sometimes lower the level of rigor of the tests.

Another type of misalignment occurs in situations involving a high measure of alignment between tests and standards, but the standards are not high quality. This can create another set of negative consequences. An assessment that is written to poor quality standards may not be a valid measure of students' knowledge of important mathematics. The results of the assessment may be of little use to teachers, parents, students, or other potential constituencies. For example, if the standards are vague the results will not provide useful indicators of how to differentiate instruction. If the standards do not include the NCTM Process standards, those are not likely to be emphasized in the classroom. If they are too narrowly conceived the results of the assessment may

encourage poor judgments, such as retaining students until their arithmetic fluency meets an established level, denying those students access to broader mathematical skills and concepts. Poor quality standards also invite loopholes for alternate assessments, where the loopholes become the norm rather than the exception.

Challenge 2: Uses and Consequences of Large-Scale Assessments

The single greatest misuse of assessments for high-stakes accountability comes from the common practice of using one test for purposes other than the purpose(s) for which it was developed. A test designed to broadly measure eighth-grade level content at the state or district level, for example, is inappropriately used to make decisions about promotion or retention of students to secondary-level mathematics. The pressure to do so can be fierce, fed by insufficient resources to develop or use multiple measures for high-stakes decisions. However, the fallout from such practice is troubling. First, such practice is a huge threat to the validity of the inferences made. This cannot be overstated. Because the content covered in standardized tests is nearly always too narrow to reflect the complexities of teaching and learning, such tests can at best be considered a weak indicator of the mathematics the students have learned. Thus they should be used as only one component in an array of data.

As the stakes attached to a given assessment rise, so do the influences (often negative) of that assessment on curriculum and instruction. At the same time, the need for professional development and support for teachers also rises. Sadly, this is often not a high priority for states, districts, or schools. Or, if it is, the professional development is misguided, in that it only reinforces the narrowing of the curriculum. For example, if teachers were provided rich professional development that encouraged them to engage their students in deep conceptual understandings of mathematics, the students would more than likely perform better on narrow standardized tests. But often, in an atmosphere of "test prep" fear, the only professional development provided is geared toward teaching students only the narrow slice of mathematics that is likely to be covered by the test.

In an effort to improve scores on the large-scale assessment, some schools are reported to have engaged in questionable practices. These range from retaining students one or more years so they don't enter the pool of students taking the examination, suspending low performers, or encouraging them to stay away from school during the testing day.

Challenge 3: Reporting of results

The challenge of reporting is to make the results of assessments useful, informative, and the basis for making improvements. It also involves making inferences that are as valid as the assessments upon which they are based. If the results meet none of these criteria, if they are not used to support teachers and administrators, if they are not timely, or if they involve inappropriate comparisons, then the potential benefits of the assessments will have been lost. Especially challenging is providing a meaningful context for these data to the press and general public.

Challenge 4: Equity

Each of the three challenges discussed have the potential of creating threats to equity. Assessments that are misaligned or aligned to poor quality standards can deny students the opportunities to learn the mathematics that all children should have access to. When decisions are based on a single measure there is a high risk of inequitable outcomes. And when results are not reported in an informative way to all constituencies, some students may suffer as a result. Furthermore, when test items are written without concern for giving all students broad access to show the mathematics that they know and can do, there are equity challenges. Equity and equality are not the same. A high-quality assessment system would be designed with the aim of reducing the achievement gaps that exist among many different subpopulations.

Challenge 5: Quality

While the four previous challenges present the negative consequences of certain features of assessment systems, this challenge delineates the primary features that encompass a high-quality assessment system. Such a system should:

- include multiple measures
- involve classroom teachers and mathematics specialists throughout the development process, including piloting and field-testing
- have test specifications that focus on a small number of significant curricular aims, including the NCTM Process standards, and provide clear guidance about the emphasis of the strands of mathematics at each grade level as well as the various item types used to assess the standards
- clearly describe the cut scores, the use of technology, the testing conditions including accommodations and modifications, and how to read and interpret the reports
- have sample tests accompanying the test specifications that include exemplars of items with high and low complexity, are updated annually, and include each item type used to assess the standards along with information about scoring student responses (and student work with scoring descriptions for any performance item types)
- annually release a significant proportion of the test items previously administered along with percent correct and samples of student work on constructed response items
- provide guidance to teachers on the development of quality formative assessments aligned to the high-stakes testing, but *not* simply imitations of the high-stakes testing—instead, including a variety of item types, providing additional information to the teacher, student, and parent about each student's understanding and application skills of significant mathematics
- establish quality controls for each component of the system, as well as the system as a whole, that are clearly described
- annually update the development process, test specifications, sample tests, and quality controls, and make these readily available to the public
- be robust enough for valid and reliable inferences to be made about each student's abilities while being sensitive enough to detect the impact of instruction

- be supported by sufficient resources for development, administration, monitoring, scoring, and interpretation of the results.

III. Current Initiatives

A. External to NCTM

Organizations other than NCTM have produced a wide range of materials and information related to the challenges discussed above. Much of this work is generic and not specific to mathematics. Nevertheless, there are elements from this work that can serve as useful resources to future initiatives at NCTM. Included in these materials are various sets of criteria for quality assessment systems. (See Appendix 1 for details.) In summary, there is a lot of material relating to high-stakes assessment, but very little is specifically directed at mathematics or to teachers of mathematics.

B. NCTM (see Appendix 2 for a complete list of NCTM materials and activities)

The four "flagship" publications of NCTM are the standards documents, beginning with the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989) and ending with *Principles and Standards for School Mathematics* (NCTM, 2000). The first standards document addressed assessment issues under the "evaluation standards," but the message of assessment was somewhat lost in the zeal of curricular reform. The third document, *Assessment Standards for School Mathematics* (NCTM, 1995), brought the spotlight back to assessment. Both the *Curriculum and Evaluation* Standards and the *Assessment Standards* called for changes in external testing, moving away from discrete objectives and the sole use of "objective" items and toward the use of more contextual, constructed-response items that would require human scorers. Both documents were written before the current era of pervasive high-stakes tests; consequently, neither directly addresses some of the issues facing teachers, students and other constituencies. The most current standards document, *Principles and Standards for School Mathematic*, attempts to treat assessment as a cross-cutting theme, threading it throughout discussions of the teaching and learning of school mathematics. The major assessment messages concern classroom assessment, but there is not a direct attempt to discuss issues of high-stakes testing. The result of all these efforts is that the *Assessment Standards*, while broadly conceived, come closest to encapsulating messages that may be relevant to the current high-stakes testing environment.

Each of the standards projects was followed up by other efforts at NCTM that were designed to sustain the momentum of reform and render support for classroom teachers. The *Addenda* and *Navigations* series are the premier examples of this. In assessment the primary efforts have been aimed at classroom assessment, starting with the highly popular *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions*. The grade-band classroom assessment series *Mathematics Assessment: Cases and Discussion Questions*, as well as *Mathematics Assessment: A Practical Handbook*, are also primarily aimed at improving assessment within the classroom. In

addition, the *Mathematics Assessment Sampler* series provides a compilation of items aligned with *Principles and Standards for School Mathematics.* As with the others, the primary audience is classroom teachers or those who provide professional development to teachers. Other documents, such as the executive summaries of the Standards, the *Advocacy Toolkit*, and *A Family's Guide: Fostering Your Child's Success in School Mathematics*, make little or no mention of assessment or high-stakes testing. One exception is the *Administrator's Guide: How to Support and Improve Mathematics Education in Your School* (NCTM and ASCD, 2003), which includes ten "actions" that school administrators might take in the realm of assessment, including the following:

- Ensure that assessments are aligned with the curriculum
- Examine the impact of high-stakes assessments on the instructional climate in schools
- Ensure that decisions about placing students in mathematics classes and evaluations of teachers' effectiveness are not based on a single test
- Ensure that teachers are using a variety of classroom assessment methods that measure conceptual understanding along with factual and procedural understanding
- Ensure that teachers rely on daily formative assessment to plan and evaluate instruction

Other forums for communication concerning high-stakes tests emanating from NCTM include the May/June 1998 *Mathematics Education Dialogue* that focused on the question, "Should high-stakes tests drive mathematics curriculum and instruction?" In November 2000, the NCTM Board issued a position statement on high-stakes testing. The primary points made in the statement are that no single measure should be used to make significant education decisions, that high-stakes tests can have negative impacts on instruction and curriculum, and that assessments should first and foremost advance students' learning. The more current position statement of January 2006 reiterates the earlier statement in a more succinct format. The professional development focus for the current year is "Assessing to Learn and Learning to Assess." Teachers have been encouraged to attend sessions at regional meetings that address assessment issues (only some of which are about high-stakes testing).

An Assessment Task Force submitted a series of recommendations to the NCTM Board in September 2002. Subsequent Board actions led to some of the publications noted above, as well as the "Assessment 101, 102, 103" series, some of which are in press. The first in this series, *The Language of Mathematics Assessment: Concepts and Terms,* is a general introduction to assessment. The third of these, *Interpreting Large-Scale Assessments,* is a solid primer on large-scale tests.

IV. Action Plan Recommendations

In light of these challenges and NCTM's efforts to date, the task force offers the Board of Directors four specific sets of recommendations.

Enhance the assessment knowledge base of our members by developing print and Web-based resources that provide:

1. Ongoing information and updates about assessment activities, reports, etc. to the entire membership, perhaps as blast e-mails.
2. Online professional development assessment institutes based on literacy materials
3. Easy-access portal (through the Web site) to released items and item banks
4. Questions about high-stakes accountability tests to which every teacher of mathematics should have answers
5. Focused news bulletin pieces on high stakes testing issues

Develop materials that support NCTM's high-stakes tests position statement and NCTM's ongoing advocacy and political action activities. These could be Web-based or inserted into journals or the *NCTM News Bulletin* to make them highly accessible. For example:

1. A simplified, focused Standards for Effective High-Stakes Testing Systems – based on *Assessment Standards*. This could be done by an NCTM staff person within 6 months, with advice and consent from the task force. It could be incorporated into the *Advocacy Toolkit*.
2. Examples of best practices in assessment designed to help our members address the challenges we've noted above
3. A game-plan developed by staff at NCTM for state Affiliate action to improve the positive impact of high-stakes large-scale assessments

Develop a set of criteria for rating state testing systems. This would entail developing a set of critical characteristics of an effective high-stakes testing system in mathematics that could be used for evaluation. (See Appendix 3 for initial draft). The task would involve, perhaps, members of the Large Scale-Assessment Task Force developing a draft, piloting, and revising. It would require five people, a year of work, and two meetings. The product would be Web-based. (See Appendix 4 for budget)

Strengthen collaborative ventures with other organizations to undertake work, such as reviewing and rating state testing systems in mathematics (e.g., Achieve, CCSSO, CEC, TESOL, NAEP). Reports could link ratings with NAEP results.

V. Conclusion

Our members face a range of challenges as they try to adapt to the increasingly test-driven educational environment. NCTM has provided a wealth of materials that address assessment of mathematical understanding—most focused on classroom assessment. Our analysis finds, however, that there are large gaps in the resources and materials NCTM has developed that specifically address the issues surrounding high-stakes assessment. Moreover, our analysis finds that what is needed is *not* more manuals or guidebooks or typically large publications. Instead, what we believe is most helpful and needed are the smaller, more focused materials delineated above.

References

Olson, Lynn; State test programs mushroom as NCLB mandate kicks in; *Education Week;* November 30, 2005; p10-14

Popham, W. James; Living (or Dying) with your NCLB tests; *The School Administrator*; December 2003.

Wilson, Linda D.; High-stakes testing in mathematics; draft chapter for NCTM research handbook.

Herman, Joan; Making accountability work to improve student learning; CSE Report 649; March, 2005; Center for the Study of Evaluation; UCLA, Los Angeles, CA.

Linn, Robert; Issues in the design of accountability systems; CSE Report 650; April, 2005; Center for the Study of Evaluation; UCLA, Los Angeles, CA.

Linn, Robert; Accountability: Responsibility and reasonable expectations; CSE Report 601; July, 2003; Center for the Study of Evaluation; UCLA, Los Angeles, CA.

Appendix 1: Materials External to NCTM

National Commission on Testing and Public Policy (2002). *From gatekeeper to gateway: Transforming testing in America.* Boston: National Board on Educational Testing and Public Policy, Lynch School of Education. The commission issued a statement that calls for less reliance on multiple-choice tests, greater use of multiple sources of evidence for decision-making, critical evaluations of tests for fairness and accuracy, and a separation of testing that is used for instructional purposes and that used for accountability purposes.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association. Among other standards these groups emphasize that when stakes are high, it is particularly important that the inferences drawn from an assessment be valid, fair, and reliable.

The Commission on Instructionally Supportive Assessment (October, 2001). *Building tests to support instruction and accountability: a guide for policymakers.* The report calls for states to follow nine requirements as "steps to create responsible state assessment systems." This commission was convened by American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, and National Middle School Association. (http://www.testaccountability.org)

Crafting Curriculum Aims for Instructionally Supportive Assessment (2003) (http://education.umn.edu/nceo/Presentations/CraftingCurricula.pdf) – practical ideas for how state curricula can be optimally configured to foster instructionally supportive assessment; that is, assessment intended to promote more effective classroom instruction.

Selected articles and policy briefs from the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA (http://www.cse.ucla.edu/) including:

- Making Accountability Work to Improve Student Learning, March 2005
- Accountability: Responsibility and Reasonable Expectations, July 2003
- Issues in the Design of Accountability Systems, April 2005

Appropriate Use of High-Stakes Testing in Our Nation's Schools (http://www.apa.org/pubinfo/testing.html) – a policy brief from the American Psychological Association

High-Stakes Testing, Uncertainty, and Student Learning (http://epaa.asu.edu/epaa/v10n18) - a provocative article that examines whether there is a link between high-stakes graduation tests and increased achievement.

High-Stakes Assessments in Reading (http://www.reading.org/resources/issues/positions_high_stakes.html) – a position statement from the International Reading Association

Appendix 2: List of NCTM Materials

ASSESSMENT PUBLICATIONS
January 2006

From the NCTM Catalog

Assessment Standards for School Mathematics (p. 17)
Exploring Classroom Assessment in Mathematics (p. 38)
How to Evaluate Progress in Problem Solving (p. 19)
Learning from NAEP: Professional Development Materials for Teachers of Mathematics (manuscript now being edited)
Mathematics Assessment: A Practical Handbook for Grades K-2 (p. 10)
Mathematics Assessment: A Practical Handbook for Grades 3-5 (p. 10)
Mathematics Assessment: A Practical Handbook for Grades 6-8 (p. 10)
Mathematics Assessment: A Practical Handbook for Grades 9-12 (p. 10)
Mathematics Assessment: Cases and Discussion Questions for Grades K-5 (p. 10)
Mathematics Assessment: Cases and Discussion Questions for Grades 6-12 (p. 10)
Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions (p. 20)
Mathematics Assessment Sampler, Grades PreK-2: Items Aligned with NCTM's *Principles and Standards for School Mathematics* (manuscript not received, expected 1 March 2006)
Mathematics Assessment Sampler, Grades 3-5: Items Aligned with NCTM's *Principles and Standards for School Mathematics* (p. 11, 29)
Mathematics Assessment Sampler, Grades 6-8: Items Aligned with NCTM's *Principles and Standards for School Mathematics* (not yet in catalog)
Mathematics Assessment Sampler, Grades 9-12: Items Aligned with NCTM's *Principles and Standards for School Mathematics* (not yet in catalog)
Results of the 8th Mathematics Assessment of the NAEP. (Manuscript to arrive by 13 January)
Results and Interpretations of the 1990 through 2000 Mathematics Assessments of the NAEP (p. 42)
Results from the Seventh Mathematics Assessment of the NAEP (p. 42)
Using Assessment to Improve Middle-Grades Mathematics Teaching and Learning (p. 34)

President's and Past President's Chats

Assessing to Learn and Learning to Assess
An Educational Crash Diet

Focus of the Year

Main Page:
http://www.nctm.org/focus/

Elementary School Resources
http://www.nctm.org/focus/assessment/elementary.htm

Middle School Resources
http://www.nctm.org/focus/assessment/middle.htm

High School Resources
http://www.nctm.org/focus/assessment/high.htm

Appendix 3: Draft Criteria for Evaluation of High-Stakes Tests

ALIGNMENT

Is the testing system aligned with other elements of the system, measuring what is important and what should be taught?

1. Are there materials that support teachers in aligning curriculum and instruction to the assessments?
2. Do alternate assessments align with instructional expectations or age of students?
3. Can results on alternate assessments be compared with results on standard assessments?
4. Is there evidence that the standards describe the mathematics that is important for students to learn?
5. Are comparisons made with other tests valid?

EQUITY

Is the testing system fair to students, giving them every reasonable opportunity to demonstrate what they know and can do?

1. Are students given the opportunity to learn what they are being assessed on?
2. Are there opportunities for teachers and students to become familiar with the content and format of the tests and examples of "good work" on constructed responses items?
3. Are test items free of bias and piloted and reviewed with appropriate statistics, such as differential item functioning analysis?
4. Do scoring guidelines and anchor papers for constructed response items account for unanticipated, but reasonable, responses?
5. Are items constructed to maximize access, giving all students the opportunity to demonstrate the mathematics they know?
6. Are results disaggregated?
7. Do students have access to the same technology on the test as they do in class?
8. Are tests timed? If so, do students have adequate time?
9. Can students write on the test booklet?

USES and PURPOSES

Are the uses and consequences of the testing system compatible with the design and original purpose of the tests?

1. Do the assessment uses match the purposes?
2. Is classroom assessment data used as one measure of student achievement?
3. When test results identify intervention needs, are long-term implications considered?
4. Are all uses made public?
5. Is appropriate data available to teacher and other stakeholders?

6. Is professional development and other support provided so stakeholders can interpret the data?
7. Does the system motivate teachers and students to learn rich and important mathematics?

REPORTING and COMMUNICATION

Are the results of assessment reported and communicated widely and accurately, maximizing the likelihood of their being used to improve performance?

1. Is the data meaningful?
2. Is the data useful?
3. Is the data accessible?
4. Is the data easily interpreted by targeted audiences?
5. Does the reporting connect "how did we do?" to "how can we do better?"?
6. Are reports timely?
7. Does the reporting allow for a variety of appropriate comparisons?
8. Do subscale results permit valid inferences?
9. Does the data allow for interpretations of growth?
10. Are items released and accessible?
11. Are descriptions of cut scores clear and meaningful?

QUALITY

Does the testing system include the elements of a high-quality system?

1. Does the system include multiple measures of student achievement?
2. Is the test balanced and comprehensive in covering standards?
3. Are teachers involved in all phases, from development to scoring?
4. Does the test assess both content and process standards?
5. Is the variety of item types adequate?
6. Are test specifications clear?
7. Are scoring procedures public and do they include samples of student work?
8. Is the scoring technically sound?
9. Are there quality control procedures in place to reduce error?
10. Are cut score procedures public and replicable?
11. Are there sufficient resources ($ and human) to construct, administer, monitor, score, and interpret results?