A Critical Evaluation of PISA's Measurement of Mathematics

Mary M. Lindquist

May 11, 2009

Send all correspondence to:
Mary M. Lindquist
Fuller E. Callaway Professor of Mathematics Education (Emeritus)
220 North Jefferson Street
Lewisburg, WV 24901
Voice: (304) 647-3150
Fax: (304) 647-3150
E-mail: mlindq@suddenlink.net

Ross Turner (2009), presenting an insider's view of the Programme for International Student Assessment (PISA), gives an overview of the program, so I have not repeated that background information here. I have been asked to do a critical evaluation of PISA. In the past, I have reviewed items and frameworks for PISA and served on the Mathematics Expert Group for the 2003 PISA assessment, but I have not been involved formally with PISA since then. To give this evaluation structure, I have chosen to use a framework developed by the National Council of Teachers of Mathematics (NCTM). This tool, based on the *Assessment Standards for School Mathematics* (NCTM, 1995) and other recommendations and research, allows individual or group evaluation of assessments or assessment systems. I made the evaluation considering three different uses: (1) PISA used as an international instrument, (2) PISA as used presently in the United States, and (3) PISA if used as a high-stakes instrument in the United States.

ASSESSMENT FRAMEWORK

The National Council of Teachers of Mathematics developed a framework for evaluating large-scale mathematics assessments (http://www.nctm.org/resources/assessment/). The framework presents six categories, each with several criteria. The categories are briefly described below and the criteria for each category are given in Appendix A. Further explanation of each criterion can be found at the website.

- **Quality—**A high-quality assessment system effectively and validly measures the mathematics that students know and are able to do.
- **Alignment—**The assessment system is aligned with other elements of the system, measuring the learning of important mathematics.

- **Fairness—**The assessment system is fair to students, giving them every reasonable opportunity to demonstrate the mathematics they know and can do.
- **Transparency and Openness—**The components of the assessment system are widely available and broadly communicated to all stakeholders.
- **Reporting and Communication—**The results of the assessment system are reported and communicated widely and accurately, maximizing the likelihood of their being used validly to improve performance.
- **Results—**The results of the assessment system are used to guide decision-making.

The electronic tool allows the user to rate each criterion as fully met, partially met, not met, or don't know whether the criterion is met. I have one caveat: nothing is ever fully met. One can always improve the methods and procedures used in creating, administering, and reporting assessments. I think of the criteria on this evaluation tool as being satisfactorily met for the purpose of the assessment in question rather than fully met. The tool gives a report of the ratings made by the individual or group, but makes no other judgment.

My judgments were based on my previous work with PISA and reading the latest reports. They are also colored by my extensive involvement with Trends in International Mathematics and Science Study (TIMSS) and National Assessment of Educational Progress (NAEP) mathematics assessments. They are definitely made as an outsider who may not be familiar with all the nuances of PISA's development, administration, and reporting.

PISA: USED AS AN INTERNATIONAL INSTRUMENT

The original outcomes proposed by the Organisation for Economic Co-operation and Development (OECD) in 1999 were to provide

- a basic profile of knowledge and skills among students at the end of compulsory schooling,
- contextual indicators relating results to student and school characteristics, and
- trend indicators showing how results change over time (OECD, 1999, p. 2).

The knowledge and skills were defined for reading, mathematics, and science literacy—what students can do with the knowledge and skills they are expected to obtain by the end of compulsory schooling. The original outcomes alone would be worthwhile, but I believe that PISA has been able to accomplish more than this as it has been made operational. Turner (2009) has described many of the positive aspects of comparative studies in terms of the mathematics portion of PISA.

It is, however, important to evaluate all assessments periodically. Table 1 gives a summary of my evaluation of PISA as an international assessment according to the criteria given in the NCTM framework. The remainder of this section gives the rating for each criterion and discusses issues that the criteria raise.

Quality. If the criteria in the quality category were not met, there would be no need to examine the other categories. In my judgment, the PISA mathematics assessment fully meets five of the criteria, partially meets one other, and the remaining criterion is not applicable to a single assessment such as PISA.

It fully meets the criteria about the development process (1)*; sound psychometric principles (3); accuracy in developing, administering, and scoring (6); and public statement of purpose (7). The other criterion (2) that is fully met is one concerning the measurement of

* Numbers in parentheses are the criteria numbers in Appendix A. See the appendix for a fuller description of the criteria.

important mathematics based on a reasonable set of standards. This criterion is central to the discussion of PISA for the mathematics community, so I will discuss it more fully below. The criterion (4) that is only partially met concerns adequate resources to construct, monitor, score and use the information. The phrase "use of information" triggered my rating. I do not think enough resources have been allocated to help all stakeholders use the information. Insiders probably have a different opinion about whether there are enough resources to carry out the other tasks. Criterion (5) focuses on systems of assessment and is not relevant to PISA.

The criterion (2) about the mathematics being assessed states, "The system measures the learning of important mathematics, and is based on a clear, reasonable set of content standards." There is no doubt that countries differ on the view of what mathematics is important, the time topics should be taught, and how they are taught. A quick look at the TIMSS survey of what topics are included in curricula of different countries shows the diversity among countries (Mullis et al., 2008). Many countries are wrestling with conflicting views of mathematics partly initiated by international studies. For example, researchers from Norway (Gronmo and Olsen, 2006) argue that school mathematics does not communicate the same meaning to all; sometimes it is used more in the sense of pure mathematics focused on skills and procedures and at other times on applied mathematics. The paper by Turner (2009) clearly sets forth the literacy view of mathematics that PISA used in the assessment.

Alignment. This category has only three criteria, each of which I rated as fully met. The first criterion (1) is about measuring the alignment among the assessment parts such as between the items and the content standards. I do not think there is a formal measurement procedure for alignment conducted independently from the developers such as the ones used in many of the

states. There is a careful matching of items with the content domains, the contexts, and the competencies described in the OECD framework for PISA.

The second (2) criterion states, "Each item deliberately targets one or more elements of the content standards." Because of the broader conceptualization of the content standards in PISA, the targets are more those that can be probed, rather than those that require specific deterministic answers or actions. This requires designing more flexibility into the items than is found in an assessment requiring results about specific skills. I believe this is one of the strengths of PISA as an assessment and would be one of its weaknesses if used to examine specific curricular outcomes (not the intent of PISA).

The third (3) criterion speaks to the sufficiency of the items to assess the breadth and depth of the mathematics framework. This criterion, considering the time constraints of testing in schools, is met especially when mathematics is the focus content domain of the PISA assessment, such as in 2003. I question whether the same statement can be made during the PISA cycle years when mathematics is not the main focus. On the other hand, the item sampling allows for a much broader set of items than many of our states' assessments. A true assessment of students' mathematical literacy would contain items without time restraints and with access to all the tools available (computers, books, and other resources).

Fairness. Two of the criteria are fully met and four are partially met. The two that are fully met focus on the opportunity students have to learn the mathematics required (1) and on the students' use of the same technology used in their classes (4). The mathematics required is based on content, processes, and skill levels that students ending compulsory schooling should have had the opportunity to learn. Given the diverse treatment of mathematics around the world, this

is an assumption that should be verified by the participating countries. The policy on technology allows the use of calculators that students normally use in their classes.

I rated criterion (2) as partially met. It states, "Assessment items are constructed to maximize the likelihood that students can demonstrate the mathematics they know." Many educators and researchers echo my concern about the reading level of the mathematics (and science) items. From England, Ruddock and colleagues (2006) speak of the high reading demand. From Norway, Sjoberg (2007, p. 9) raises the question of whether "real-life challenges can be assessed by wordy paper-and-pencil items."

Two of the other criteria that I rated as partially met reflect my own position relative to the bias criteria and perhaps an unrealistic personal expectation about time. Although statistical analyses are done on the items to check for bias (3), many items strike me as catering to masculine interests. Statistics may show that the items do not negatively discriminate against female students, but that still does not give me the confidence that the whole set of items interacts with female test-takers in the same way as it does with male test-takers. The other criterion (5) deals with adequate time. Most students probably had the time they needed or wanted. If we were to improve students' persistence in solving mathematical problems, I would hope that time would not be limited. If we are truly going to measure mathematics literacy, we need to have some items that take a great deal of time.

The remaining criterion (6) speaks to accommodations. There are policies, but it is not clear that they are uniform from country to country. There are international guidelines for exclusion of students and possible accommodations for special needs students. "There is also provision for cases where a school caters exclusively for students with special educational needs to administer a shortened (1-hour) version of the test to all sampled students. However, more

recently procedures have been adapted that would enable individual students to use this form [the 1-hour booklet] under certain defined conditions, and would also allow more flexible administration conditions. For example, the number and duration of breaks during the assessment session can be changed to suit the needs of individual students using the *Une Heure* (UH) Booklet, and the provision of extended time to complete the assessment in particular situations is possible" (correspondence from Ross Turner).

Transparency. My ratings for this category split, with three criteria being fully met and three being partially met. Criterion (3) concerning the number of released items, criterion (4) about guidelines for scoring constructed-response items, and criterion (5) about setting performance levels were rated as fully met. One of the strengths of the PISA mathematics assessment is the description of the performance levels based on the competencies.

The remaining three criteria are partially met. Criterion (1) states, "All stakeholders are familiar with the background, purpose, and consequences of each assessment." Because there are no consequences for students, I doubt that the lack of knowledge about background and purpose matters to them. The lack of consequences for students is both a positive and a negative. It is a paradox that we have with comparative assessments, including NAEP and TIMSS. What is the motivation for students to do well? How can there be consequences when it is not an assessment of individuals? On the positive side, the lack of consequences for individuals allows for innovation and for pushing the boundaries of the assessments themselves.

Stakeholders in many countries have a lack of familiarity and understanding of PISA. This may not be true in countries such as Germany, whose citizens were shocked by the results of the 2000 assessment. This prompted widespread discussion and reform, partly in the form of

more testing. An article, *PISA Education Studies Come under Examination* (Deutsche Welle, 2007), discusses the recent controversy about reform raised by the 2006 results.

My concern centers more on one of the stakeholder groups that I believe has been rather ignored—the teachers of school mathematics. They have much to offer in the development of the assessment and can be ambassadors for positive use of results. I do not know of any effort to elicit their evaluation of PISA.

Criterion (2) concerns the familiarity of the students with the content and format of the assessment. This is partially met because the content is underlying the mathematics with which students at this level should be familiar. Many are not familiar, however, with how items are presented— in contexts that are often novel and with a slightly different format in some of the multiple-choice items. The amount and type of testing in other countries varies, so there is a question of fairness, especially with the demands of large-scale assessment items, among countries that needs to be monitored. On the other hand, we do not structure the problems we meet in everyday life and, hence, we should expect students to be prepared to deal with the unfamiliar.

Criterion (6) focuses on evaluation and modification of the assessment process. There have been concerns about the openness of the process. For example, Schagen and Hutchinson (2006) call for more openness: "Both IEA and OECD should maintain open access to their methodologies and encourage criticism and debate from the wider academic community. This should be one in a spirit of openness and willingness to learn and improve, recognizing there is not necessarily a 'correct' answer to each technical problem."

Communication. Only the international reports (OECD, 2003, 2007) were considered in this part of my evaluation. Some could argue that the reports (1) are not timely, but I am not sure

how much hurry we should be in for such a massive undertaking. I rated this criterion as fully met, especially considering that the criterion included statements about the meaningfulness and completeness of the reports. Likewise, I find the reports to be useful and correctly interpreted by the intended audiences (2). I commend those responsible for the most recent discussion of the mathematics results in the international report (OECD, 2007) for placing more emphasis on the items and structure of the assessment than on the rankings.

On the other hand, the criterion (4) stating that reports should connect "how did we do" with "how can we do better" is weak (partially met). What has become known as the First International Mathematics Study, The International Study of Achievement in Mathematics, stated clearly that more than "how did we do" must be studied (Husen, 1967). PISA does collect much more information about students and schooling, but does not take the next step of suggesting how to do better. There are some efforts by individual countries and individual authors to make some interpretations of steps toward improvement of mathematics education. This is probably the proper level for such interpretations, but the international study should consider collecting the data that make this possible or encouraging other research studies to complement their efforts.

The remaining criterion (3) relates to the evaluation of reports. No formal procedure is in place, but certainly the international reports are open to scrutiny. There is a survey about the quality and presentation of the electronic version (OECD, 2007). I rated this criterion as partially met and believe that until this stage of the development of PISA it was probably more productive to spend the limited resources and time on other phases of the assessment than on this criterion.

Results. Four criteria were not appropriate to rate because they dealt with intervention at the individual student level, professional development, stimulating curricular and instructional

modifications, and systems of assessments with multiple measures to make high-stakes decisions. These are either not appropriate to the purpose and design of PISA or should be the responsibility of individual countries.

The criterion (4) that was fully met dealt with valid subgroup comparisons. Partially met was criterion (5), which dealt with appropriate interpretations of growth and improvements that could be made. Certainly change has been addressed in PISA, but it will not be until 2012 that another full mathematics assessment is given. There is some information about change in mathematics performance in the OECD report (OECD, 2007), but also a caution about making any inferences because there are only two data points (p. 319). In years to come, I would expect this criterion to be fully met.

PISA: AS USED IN THE UNITED STATES

This section looks at PISA from the perspective of what the United States has done with the assessment in addition to participating and supporting the effort. The National Center for Education Statistics has produced a U.S. report for each assessment, has provided databases for other explorations, and has done comparisons among PISA, TIMSS, and NAEP mathematics assessments. After each administration of PISA, there is a flurry of press reports, but little more. Recently, there has been a renewed interest as shown by this forum.

If the criterion is targeted at the assessment itself, the rating was the same as in the previous section and most often is not discussed here. Additional issues about a criterion that should be raised by the United States as it continues to participate in PISA are addressed. The evaluations are given in Table 2.

Quality. Criteria (1)–(3) and (6) are concerned with the instrument, the framework on which the assessment is based, and the quality of developing, administering, and scoring. Thus,

the ratings for these four criteria are the same as for the international use or rated as fully met. Criterion (7) is concerned with the purpose of the assessment and the clarity of the public statement of this purpose. I rated this as fully met for the international use and at first rated it as partially met as used in the United States. Then I realized that this judgment was based on the fear that PISA may be used in the United States for other purposes. Until now it had not been, so it should have the same rating (fully met). Criterion (5) is about a system of assessment, which is not applicable to a single assessment such as PISA.

I rated only one criterion differently in United States use from international use. Criterion (4) states, "Adequate resources are available to construct, administer, monitor, score, and use the information." It was because of "use the information" that I rated it partially met in the international rating, but lowered this to not met in the United States. There was little effort, other than the usual publicity of the horse race, to make use of the information in the United States. Recently, there has been renewed interest in PISA by policy makers. It is important to have policy makers informed. It is equally as important to have other stakeholders informed, especially the broad mathematics community, including those who teach school mathematics.

Alignment. The alignment category is concerned with the alignment of the assessment with the framework of the assessment. This does not change when considering the use in the United States. Thus, the criteria in this category are rated the same (fully met) as they were in the international use. A word of caution is in order here, because the alignment considered here, as it should be for this rating, is between the PISA assessment and its framework. If one looks at the alignment between PISA and frameworks or standards used in the United States, then another rating may be made. The Council of Chief State School Officers (CCSSO) examined the alignment of PISA with one state's 10th-grade standards. It found "a number of areas where

math content is similar, but there are also several areas of differences. The coarse grain alignment level is .37 and overall alignment is .04" (Blank & Smithson, 2009, p. 4). The comments by those who participated in the entire alignment study were as enlightening as the alignment results. For example, the math group noted "that PISA has very little algebra or advanced algebra (considering the 15-year target), but the level of reasoning and numeric literacy required is what gives the test validity and/or its difficulty" (Blank & Smithson, 2009, p. 8). The conference for which this paper is written includes a session by Smithson on this alignment process, so we should hear more about it and the findings.

Fairness. The ratings for fairness are the same as for the international use of PISA except for two criteria. One difference is for criterion (1): "Students are given the opportunity to learn the mathematics on which they are being assessed." Because it is clear that the underlying knowledge and skills needed to answer the items on PISA are included in the standards of most states, there must be other reasons that U.S. students do not perform well. The reason cannot only be that there are no consequences for our students, because this is true of all countries. An examination of most state assessments gives insight into some of the reasons. The high-stakes nature of the state assessment programs, the limited budgets that forces mainly multiple-choice items, and the need for a large percentage of our students to reach set performance levels have turned most of our testing instruments and performance level settings into a vicious circle of minimal competency on routine problems. This is a far cry from the PISA items that do not look like what most students see on their assessments. Even college students complain if a test item does not mock something they have done for the class. I claim that we are not preparing our students (there are exceptions) to expect that they need to pull together their knowledge and

skills in a unique situation. If we continue to participate in PISA, then we should seriously address how our curriculum, instruction, and assessment practices should change.

The other difference in my ratings is on the criterion (6) involving accommodations. The United States chose not to use the special 1-hour test booklet that was available for countries (Baldi et al., 2007, p. 29). Thus, I lowered my rating from partially met at the international level to not met in the United States. There probably need to be more accommodations available if this is to truly measure all students.

Transparency. There were only two ratings of the six criteria that were different for this use of PISA than for the international use. I lowered my rating on criterion (1), which focuses on the familiarity of stakeholders with the background, purpose, and consequences of PISA from partially met to not met. There is little knowledge in the United States—among students, teachers, and others involved in education—about PISA. I would hope as PISA becomes more institutionalized that groups of educators and mathematicians would become more involved with all phases of the assessment. Turner (2009) cites one example of an article published in the *Mathematics Teacher*, but there has been little other information for teachers. This is quite a contrast from the huge public relations and educational effort that was made when the results of The Third Mathematics and Science Studies (1995) were released. I have volumes of presentations and in-service sessions that were developed in the United States around that assessment.

The other criterion (6) that I lowered deals with evaluating and modifying the assessment process. Partly because PISA has not been on the front burner, there has not been a lot of input from those in the United States. There are a few individuals from the United States that have been central to the development, but this criterion includes more than development. I think of all

the involvement today in the NAEP assessments. With all the input and checks and balances, I often wonder how a NAEP assessment ever comes to fruition. The international input mirrors some of these inputs, but I do not think the United States has involved sufficient or various expertise in developing or evaluating PISA.

Communication. Not surprisingly, I rated the United States on communication much lower than I did the international effort. I rated two criteria as partially met as I did the international communication, but these felt like "less than partially met." The criterion (2) dealing with reports being correctly interpreted by intended audiences and results are disaggregated. The reports do disaggregate data in ways that are possible with the design and sampling, but they do not reach a wide enough audience. The brief press releases place too much emphasis on the rankings. The other criterion (3) is similar to the previously discussed one, and deals with accessibility and evaluation of reports. Limited accessibility to limited reports and limited evaluation (perhaps none solicited other than initial reviews) led to a partially met rating.

My ratings for the other two criteria went from partially met to not met. Criterion (1) states, "the reports about mathematics achievement are meaningful, timely, and complete." There are so few reports and such a limited audience that I could not even say that this is partially met. This is probably because I feel so strongly that criterion (4), which connects "how did we do" to "how can we do better," is missing in the reports (not met).

Results. Although four of the criteria did not seem applicable when considering PISA as an international report of results, only two of these are not appropriate for the United States. Criteria (2) and (6) are not applicable because criterion (2) deals with individual students and criterion (6) deals with using multiple measures to make high-stakes decisions. I rated the other two criteria, (1) and (3), as not met. Criterion (1), "the interpretation of assessment results

stimulate and inform curricular and instructional modifications," has not been seriously considered by the United States. Likewise, professional development (3) spurred by PISA is almost nonexistent in the United States.

Criterion (4) focuses on subgroup comparisons; I rated this partially met. Turner (2009) argues that this could be an expansion of the use of PISA in the United States. Although the usual groups (gender and race) are compared, there are many other comparisons that could be made. Whether this is done by increasing sampling in future administrations of PISA or by other research studies is a question to be answered.

The question of growth over time, criterion (5), was rated partially met in my international rating and the same rating holds here. No more can or should be done with the present, limited set of data points. However, PISA can provide a valuable resource in the future to examine change when a larger set of data points is available.

PISA: IF USED AS A HIGH-STAKES INSTRUMENT IN THE UNITED STATES

This section deals with the future, but builds on the ratings that were given in the last two sections. A picture (Table 3) emerges if the United States were to use PISA in other ways; there would be many questions to be answered and many changes made. Even if it were used as a benchmarking instrument as some states and other educational entities use TIMSS, many of these criteria would have to be examined carefully. If the purpose were carried to the extreme use as a measure of individuals' attainment, then serious decisions would need to be made and implemented.

Quality. The quality of any assessment depends on the effectiveness and validity of the measures. In the case of PISA, I am not questioning the development procedures (1) or the

psychometrical procedures (3), but the other five criteria would need to be examined carefully and would need to be met.

The first and foremost question is about the mathematics. Is the view of mathematics (2) described in PISA one that we, the nation, hold as important? Let's examine the item titled Rock Concert. Turner (2009) discussed the underlying mathematical thought in an item that involves a relatively simple computation. We have long known from NAEP results that the difficulty level of an item increases greatly whenever students have to bring additional information to a problem or have to do more than one step.

ROCK CONCERT

For a rock concert a rectangular field of size 100m by 50m was reserved for the audience. The concert was completely sold out and the field was full with all the fans standing.

Which of the following is likely to be the best estimate of the total number of people attending the concert?

A 2,000
B 5,000
C 20,000
D 50,000
E 100,000

The Rock Concert item is seductive to those who are taught to look for key phrases or who have little motivation to read. There are two numbers in the problem and their product is one of the choices. Although U.S. curricula emphasize word problems more than when I taught school mathematics, we still have a tendency for quick responses to short items on tests. This is certainly an item that should be within the reach of all students, and they should need neither paper and pencil nor calculator. What about the metric measurements? Would the response be

different if this item were expressed in yards? Are students in the United States motivated to answer such questions? Are we preparing our students to be successful on items like the ones on the PISA assessment? Is this item illustrative of the type of mathematics that you want students to be able to do? Do we value mathematical literacy?

Criterion (4) deals with resources to construct, administer, monitor, score, and use the information. While there may be sufficient resources for the development and administration of PISA at the level that it is being used today, there have not been the resources to expend on using the information. If the United States were to use PISA as a high-stakes assessment, there would need to be substantially more resources to increase the sample size and to make use of the resulting information.

Criterion (6) is about accuracy of developing, administering, scoring, and reporting. At present, we are wrestling with the compatibility of the various states' assessments and performance levels. Do they give an accurate measure of students' knowledge and skills? Would using PISA and/or TIMSS give assistance with this problem? Could this mean that we would stop overtesting and consider sampling students at only a few junctures in their education? Would we gain enough information to make sound judgments about the curriculum and instruction in our schools? One of the strengths of NAEP, TIMSS, and PISA is the breadth tested because of the design of item sampling. We need to reconsider our state assessment programs in total, not just add one more test to the mixture. This speaks also to criterion (5) that looks at high-stakes assessments as systems. How could all of our assessments complement each other?

The remaining criterion (7) is about understanding the purpose of the assessment by the stakeholders. This is probably the first question that should be answered. What is the purpose of using PISA as a benchmark or in other ways? What benefits could the United States derive? How

would students become an integral part of this if it were not high stakes for them? (If it were high stakes for individuals, then we would have misused PISA.) How would we "educate" the public and those responsible for mathematics curriculum and instruction?

Alignment. As alluded to in the discussion of quality, PISA is not (and was not intended to be) aligned with standards in the United States. Would we change our view of mathematics? Would we change our standards or the interpretation of our standards? Would we change our assessments? Would we change our instruction? I gave the alignment category high marks when considering PISA as an international assessment, but strongly suggest that we would need to make fundamental changes in the United States before it would meet the alignment criteria for high-stakes testing.

Fairness. It is already clear that I do not think this use of PISA would be fair to our students, our teachers, and our school systems at this time. Would we give our students the opportunity to learn the mathematics (1) in a way that enables them to do well on PISA? Would we adapt our teaching and assessing to maximize the likelihood that students (2) could demonstrate the mathematics that they know? Would we allow the proper use of technology (4) in our mathematics classrooms (K–16) so students could use technology wisely? Would we be willing to consider the time (5) that should be devoted to such testing that PISA could provide? What accommodations (6) would be provided? Would we address bias issues through additional research studies?

Transparency. The needs in this category can be addressed by the following question. What procedures would we put into place so all the components of PISA are widely available and broadly communicated to all stakeholders? One of the components that may not be evident when we think of transparency is the availability of released test items. There are a good number

of released test items now, but with state assessment programs we have raised the expectations. Would testing every 3 years be sufficient? The PISA items are much more complicated to develop than much of our state assessment items. Would we be willing to fund massive development efforts so that items like this would be available not only for our assessments, but also so more could be released that were actually used on PISA?

Communication. The description of this category states, "The results of the assessment system are reported, and communicated widely and accurately, maximizing the likelihood of their being used validly to improve performance." The first three criteria are about reports being timely and complete, ensuring correct interpretation by different audiences, and accessibility. Are we willing to spend the resources—financial and human—to ensure that we provide such reports of the U.S. performance? If we decide to this, then the crucial and most difficult task would be to satisfy the last criterion in this category. Are we willing to connect "how did we do?" with "how can we do better?" This would require additional research, larger sampling of students, and a rethinking of the issues raised when considering quality.

Results. If we are to use PISA as a benchmark, are we willing to use the results to improve the mathematics education of all our students? Will we be willing to test conjectures rather than make casual conclusions? Will the results (1) stimulate and inform curricular and instructional modifications? Will we be willing to spend resources on professional development (3)? Will comparisons (4) made of groups give us insight into different paths for improvement? We will be able to support such testing for a long enough period to interpret growth over time (5)?

SUMMARY

I chose to use the structure of the NCTM framework to make my evaluation. In so doing, I could have overlooked other crucial issues. Although I am familiar with the development of PISA, I made the assumption that the development and psychometric procedures were sound, knowing full well that others would question this assumption. In all my years of working with NAEP, I have been disappointed with the lack of emphasis placed on developing and piloting of items. Yet, I also would rate their procedures as defensible.

Before closing, I would like to return to some challenges that I posed to OECD at an early meeting of international representatives. I called one group of challenges "present challenges" in that they were recommendations that should be addressed from the beginning of PISA. These included involving more teachers, answering how to motivate students to do well on PISA assessments, and probing aspects of PISA's use of situated contexts. I believe these challenges are still present today. Another group called "interpretation challenges" focused on the lack of information about the curriculum and instruction and on the need to balance research questions with policy issues. Lastly, I raised future challenges. How is PISA going to take advantage of technology, both in testing and in the changing demands of the world? I recommended (and still think it should be considered) assessing young adults. Do they really need this mathematics in their everyday life and in their work? At what level do they perform on PISA?

In making my ratings for this paper, I yearned for a set of colleagues to discuss the ratings and issues that surround them. I was uneasy making the judgments alone. So I encourage you to make your own ratings and raise issues and questions that will help us move ahead. Table 4 shows the comparison of my ratings for the three uses of PISA described in this paper.

The not met rating for the last use (and the last rating in the triplet) should be interpreted as "would need to be met."

I have argued that PISA as an international assessment, considering its purpose and framework, has met most of the criteria in the rating scheme. There is still need for improvement, especially in the areas of fairness, transparency, and communication. I firmly believe that the United States has not taken advantage of the wealth and type of information that PISA provides. It is my hope that this sudden urge to use PISA as a benchmark or in other ways will encourage us to address some of the present uses of PISA. My recommendation would be to put our efforts in seeing PISA use wisely in the United States. It should be clear that I think we need a great deal of serious thought and from a wider group of stakeholders before we make PISA a benchmark instrument for all. As I said to a colleague, "I don't think we are ready for PISA to be a benchmark." And she wisely answered, "That is probably true, but when will we be ready?"

REFERENCES

Baldi, S., Ying, J., Skener, M, Green, P. J., & Hengel, D. (2007). *Highlights from PISA 2006: Performance of U.S. 15-year-old students in science and mathematics literacy in an international context*. NCES 2008-016.

Blank, R. K., & Smithson, J. (2009) *Alignment content analysis of TIMSS and PISA mathematics and science assessments using the Surveys of Enacted Curriculum Methodology.* Prepared for the National Center for Education Statistics and American Institutes of Research.

DW-World.De Deutsche Welle (2007). *PISA education studies come under examination.* Article in Deutsche Welle, 04.12.2007. Retrieved from http://www.dw-world.de/dw/article/0,2144,2984512,00.html.

Gronmo, L. S., & Olsen, R. V. (2006). *TIMSS versus PISA: The case of pure and applied mathematics.* Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC2006/IEA_Program/TIMSS/Gronmo___Olsen.pdf.

Husen, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries. (Volumes I & II).* Stockholm: Almqvist & Wiksell.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College, TIMSS 7 PIRLS International Study Center.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics, (2008) *Assessment tool.* Retrieved from http://www.nctm.org/resources/assessment/.

Organization for Economic and Cultural Development. (1999). *The PISA assessment frameworks monitoring student knowledge and skills in the new millennium.* OECD: DEELSA/PISA/BPC (99) 8.

Organization for Economic and Cultural Development. (2003). *Learning for tomorrow's world—First results from PISA 2003.* Paris: OECD.

Organization for Economic and Cultural Development. (2007). *PISA 2006 science competencies for tomorrow's world. Volume I Analysis.* Paris: OECD.

Ruddock, G., Clausen-May, T., Purple, C., & Ager, R. (2006). *Validation study of the PISA 2000, PISA 2003 and TIMSS 2003 international studies of pupil attainment* (DfES Research Report 772). London: DfES.

Schagen, I., & Hutchison, D. (2006). Comparisons between PISA and TIMMS—We could be the man with two watches. *Education Journal*, *101*, 34–35.

Sjoberg, S. (2007). PISA and "real life challenges": Mission impossible? Contribution to Hopman (Ed.), *PISA according to PISA*. Retrieved from http://folk.uio.no/sveinsj/Sjoberg-PISA-book-2007.pdf.

Turner, R. (2009, June). *The PISA mathematics assessment—An insider's view.* Paper prepared for PISA Research Conference: What can we learn from PISA? Washington, DC.

APPENDIX A:

THE NCTM ASSESSMENT FRAMEWORK

Quality

1. Assessments are developed on the basis of a coherent process beginning with content standards and culminating in test instruments. The elements of a quality assessment include item specifications, item development, field testing, rubric development, and scoring materials.
2. The system measures the learning of important mathematics, and is based on a clear, reasonable set of content standards.
3. The assessment system is psychometrically sound and based on sound principles of measurement.
4. Adequate resources are available at the state/provincial and local levels to construct, administer, monitor, score, and use the information from the assessments.
5. High stakes summative assessments exist within an assessment system that includes formative assessments and other measures of student accomplishment.
6. Quality control procedures are put in place to ensure accuracy in developing, administering, and scoring the assessment, and in reporting the results.
7. There is a clear, reasonable and public statement of purpose for the assessment program, and the program is used for those purposes only.

Alignment

1. A valid and reliable method is used to measure alignment among the system's parts (assessment/ content standards).
2. Each item deliberately targets one or more elements of the content standards.

3. There are sufficient test items to adequately assess students on the breadth and depth of the mathematics content standards.

Fairness

1. Students are given the opportunity to learn the mathematics on which they are being assessed.
2. Assessment items are constructed to maximize the likelihood that students can demonstrate the mathematics they know.
3. Assessment items are free of bias, pilot tested and reviewed with appropriate statistics to validate item fairness.
4. Students have access to the same technology on the assessment as they do in class.
5. Adequate test time is given to students.
6. Modifications and accommodations in assessment conditions are available to provide students maximum opportunity to demonstrate the mathematics they know.

Transparency

1. All stakeholders are familiar with the background, purpose and consequences of each assessment.
2. Students are appropriately prepared to be successful on high-stakes assessments and are familiar with the content and format of the assessments.
3. For each assessment in the system, a sufficient number of test items are released annually and are easily accessible.
4. Scoring guidelines and anchor papers for constructed response items account for unanticipated, but reasonable, responses.
5. A valid and transparent process is used to set performance levels.

6. The assessment process itself is open to evaluation and modification.

Communication

1. The reports about the mathematics achievement students demonstrate are meaningful, timely and complete.
2. The reports are useful and likely to be correctly interpreted by intended audiences, and the results are disaggregated.
3. The reports are accessible to all stakeholders and are evaluated annually to ensure their usefulness.
4. The reporting connects "how did we do?" to "how can we do better?"

Results

1. When appropriate to the design and purpose of the assessment, the interpretation of assessment results stimulates and informs curricular and instructional modifications.
2. When appropriate to the design and purpose of the assessment, results are used to make instructional and intervention decisions about individual students.
3. Professional development and other support are provided to stakeholders in the areas of assessment literacy, interpretation of assessment results, and the appropriate use of assessment results.
4. Comparisons that are made about groups on the basis of assessment results are valid (including with other tests and subscales) and made with great care.
5. The assessments are appropriately used to make interpretations about growth and improvement.
6. High-stakes decisions are always based on multiple measures of students' mathematics achievement.

Table 1. Rating of criteria for PISA as an international assessment

Category (number of criteria in category)	Fully Met	Partially Met	Not Met	Don't Know	Not Applicable
Quality (7)	5	1	0		1
Alignment (3)	3	0	0		
Fairness (6)	2	4	0		
Transparency (6)	3	3	0		
Communication (4)	2	2	0		
Results (6)	1	1			4

Table 2. Rating of criteria for PISA as used in the United States

Category (number of criteria in category)	Fully Met	Partially Met	Not Met	Don't Know	Not Applicable
Quality (7)	5	1	0		1
Alignment (3)	3	0	0		
Fairness (6)	1	3	2		
Transparency (6)	3	2	1		
Communication (4)	0	2	2		
Results (6)	0	2	2		2

Table 3. Rating of criteria for PISA if used as a high-stakes instrument by the United States

Category (number of criteria in category)	Fully Met	Partially Met	Not Met	Don't Know	Not Applicable
Quality (7)	2	0	5		
Alignment (3)	0	0	3		
Fairness (6)	0	2	4		
Transparency (6)	0	0	6		
Communication (4)	0	0	4		
Results (6)	0	0	4		2

Table 4. Summary of ratings

Category (number of criteria in category)	Fully Met	Partially Met	Not Met	Don't Know	Not Applicable
Quality (7)	5,5,2**	1,1,0	0,0,5		1,1,0
Alignment (3)	3,3.0	0,0.0	0,0,3		
Fairness (6)	2,1,0	4,3,2	0,2,4		
Transparency (6)	3,3,0	3,2,0	0,1,6		
Communication (4)	2,0,0	2,2,0	0,2,4		
Results (6)	0,1,0	1,2,0	0,2,4		4,2,2

** The three numbers refer to PISA international, PISA in the United States today, and PISA as a benchmark.