We live in troubling times. In January of this year, the Union of Atomic Scientists moved the Doomsday Clock two minutes closer to midnight, i.e. the end of the world, citing the apocalyptic threat of global warming and nuclear proliferation.

Since 1991, we've moved closer and closer to the end of humanity as we know it and now have only 3 minutes left. (I'll be happy to explain the qualitative scale to you younger folks later.) I realize it is hard to think, walking around on a bright sunny day, that we could well see the end of human civilization within our life time or, at least, within the lives of our children and grandchildren.

To avoid this cataclysm, many challenging problems must be solved in the coming years. We are straining the resources of our planet; we may have already passed a tipping point with respect to global warming. Population density makes threats from infectious diseases more problematic than ever. Wars of religious intolerance are raging more fiercely than at any time since the Crusades. We will never find solutions to our technical and social problem unless the youth that we are educating today learn to reason critically, to tell truth from lies, fact from fiction, and learn to work collaboratively, not just out of self-interest but for the betterment of all.
When I selected a title for this talk, I meant to cover key, policy-related measurement issues that must be solved in the near future. Several people pointed out, however, that there is a second meaning to “lies ahead”. And there surely will be many espousing, if not outright lies, at least clear untruths. Will the coming generations be able to tell the difference between truth and lies? To do so will require a high level of appreciation for and understanding of evidence-based reasoning. Thus, the goal must be to teach not merely facts and more mechanical skills, but higher levels of critical thinking and reasoning. And our goal, as educational measurement specialists, must be to build tests that do a better job of assessing the effectiveness of our efforts to teach such skills. Educational measurement must provide the compass that will tell us if we are heading in the right direction.

[Slide 6] As a field, we, ourselves, are not immune to contentious arguments that sometimes defy reason and evidence. One versus three parameter IRT models; Angoff standard setting – state of the art or fundamentally flawed; vertical scales, regression-based value added, versus student growth percentiles?

[Slide 7] Yet within NCME, we come together, across academic and organizational boundaries, to define and improve best practices, advance the field of measurement, and support the development of new measurement professionals. Our mission is not just critical to our own professional advancement, but the fate of humanity literally hangs in the balance.

[Slide 8] Overview

The focus of this presentation is on how educational measurement can be more useful to informing effective educational policy and ultimately improving student
achievement. I want to begin by describing how educational measurement has become increasingly important to policy makers over the past several decades. Then, I will describe four areas of current work where further advances are needed to make our measures more meaningful and informative to users, from students to policy makers to the public at large (who, by the way, elects key policy makers). These areas are – attaching meaning to our measurement scales; providing accurate and complete normative information; assessing higher-order skills; and providing more useful diagnostic information. They may not be the most mathematically sexy areas of innovation for our field, but they are key to promoting valid and effective interpretation and use of test scores.

[Slide 9] Understanding how educational measurement has become so important to policy-makers, provides an important perspective on where we need to go from here.

How We Got to Where We are At

[Slide 10] Back in the 1950’s, when I was starting school, we all worried about the SAT (or ACT), but until then, tests had few, if any consequences. Schools and districts selected from a few nationally normed tests, such as the Iowa Test of Basic Skills or the Differential Aptitude Test, which provided percentile and sometimes grade equivalent scores. I discovered I was better at math than spelling and my parents made sure I worked through the weekly spelling lists and now I thank God for spell checkers. Test results were used, now and then, to evaluate educational programs.

[Slide 11] There was general trust in school leaders and teachers and no perceived need to hold them accountable.
Then, on October, 4, 1957, the Russians launched Sputnik, the first man-made satellite to orbit the earth. Suddenly, it was much harder to believe that we were first in the world in math and science and policy-makers began to ask whether our educational system was good enough. John Flanagan obtained funding from a collection of federal agencies, including NSF, the Atomic Energy Commission, and the Office of Education for a survey of the skills of American youth. In 1960, Project TALENT administered three days of tests and questionnaires to a representative sample of nearly 400,000 high school students and then followed them one, five, and eleven years after graduation to see how their skills and interests were being put to youth (ref). Shortly after this, planning was begun for a National Assessment of Education Progress (NAEP). The first NAEP assessments were administered in 1969 (ref). Initially, NAEP results were reported primarily in terms of the percentage of students who could answer each of a set of illustrative questions correctly.

The publication of A National At Risk in 1983 signaled increased attention among policy makers to the need to improve our educational system. A key response was the creation of the National Assessment Governing Board in 1989, to provide policy oversight for NAEP. The Governing Board took two early actions that further increased the use and interpretation of NAEP results by policy makers. First, the “Trial State NAEP” was launched. Whereas, previously, NAEP was precluded from reporting results below the level of broad geographic regions, now results were reported for each participating state. State policy makers could not use their results to argue that
results from higher-achieving states meant that improvement was needed and also that it was feasible, as indicated by results from other states.

The second key innovation introduced by the Governing Board was defining levels of achievement – basic, proficient, and advanced – and using these in reporting. Finally, we were beginning to answer the question of how good was good enough, although many argued that the consensus about what constituted “proficient” performance was not, itself, good enough.

One limitation of NAEP was that there are not individual student scores. Thus, while NAEP answers many important policy questions about national and state achievement, it is of little direct help to teachers, parents, and students themselves. During the 1990’s, President Clinton proposed and funded development of a set of Voluntary National Tests that would report results for individual students that could be compared to and reported in terms of the NAEP achievement levels. Many doubted that Congress would provide ongoing support for this effort, thinking Republicans would oppose it because it was “national” and democrats would not support it because it was only a test. Indeed, while the VNTs were developed, Congress blocked administration of these tests to any students (which created difficulties for the evaluation of the VNTs that the NRC was working on). What none of us anticipated, however, was something that occurred during the 2000 presidential debates. Vice President Gore was touting the need for the Voluntary National Tests to provide consistent information on individual student achievement and candidate Bush criticized the VNT because it was “voluntary”.
[Slide 14] We could not believe Governor Bush meant this, but a year or so later, he managed to push through the No Child Left Behind Act making it mandatory for each state to test students in grades 3 through 8 and in high school and report the percent found to be proficient. Policy use of educational assessments was further increased by mandatory participation rate requirements and by requirements to also report school-by-school results for key demographic subgroups. This supported equity interpretations of educational measures as well as more global assessments of educational adequacy.

[Slide 15] One issue with NCLB is that each state had to develop its own assessments and set its own achievement levels. The next big increase in policy use of educational measures resulted from work by my mentor, Donald McLaughlin, then leader of the NAEP State Analysis Center. Don developed an approach for mapping state proficiency cut scores onto the NAEP reporting scale, thus allowing comparison of the relative rigor of proficiency expectations across the states. The graphical depictions of results from this study, such as the one shown here, made it clear that while some states had high expectations for student achievement, similar to the NAEP proficiency levels, other states were considering students to be proficient even though they were below the NAEP Basic level. [Slide 16] I believe that these pictures, more anything, led to the rapid development and adoption of the Common Core State Standards, although policy-makers did not fully distinguish between content standards and performance standards. While some states have now “de-adopted” the CCSS, it is interesting that they still want comparisons to other states. Most have not adopted standards much, if any, different from the CCSS.
[Slide 17] Other policy goals have emerged in recent years. We are seeking comparisons, not just across states, but also across nations though international benchmarks. The publication of the Next Generation Science Standards has expanded our focus beyond reading and mathematics.

[Slide 18] One final recent ratchet in the policy use of assessment results came from a desire for outcome-based evaluations of individual teachers, whether to help each teacher improve their practice or simply to identify and eliminate the really bad ones.

[Slide 19] Thus, over the past several decades, educational measurement has gone from a somewhat obscure profession to a highly visible endeavor at the core of key policy debates. As Congress continues to debate state and federal roles, it is clear that the importance of our work will not soon diminish. With this in mind, I turn now to a discussion of several, policy-related issues on which we are working and which will be critical to valid and effective use of our educational measures.

[Slide 20] Attaching Meaning to Score Scale Values

Psychometricians often think once a scale with metric properties has been defined, we are done. We can argue whether increments of the same size are really comparable in ways that matter to policy-makers, such as equal value or equal work to obtain a fixed gain, but we are pretty confident that more is better and a lot more is generally better than a little more. Policy-makers, however, aren’t sure what to do with statements like “There was a mean improvement of five scale score points from one
year to the next.” Is that a lot or a little? Did it bring students up to a critical level for further learning?

[Slide 21] In recent years, we have used performance levels set by educators to communicate score meaning, with proficient being the level that a consensus of educators believes a student should achieve. Year-to-year gains are understood in the increases in the percentage of students reaching this level – policy makers do understand percentages pretty well. There is now a great policy need to go beyond achievement levels in at least two significant ways.

First, we have begun to explain what it means to be at a given score level in terms of the likelihood of success at what comes next. All of the recent work on college and career readiness or preparedness is designed to communicate the level of mathematics and language arts achievement needed for a given probability of success in college or career training, as measured by a variety of college and career criterion measures – measures such as Freshman or four-year GPA at a “typical college” or the likelihood of being able to take and pass credit-bearing courses at a community college without needing further remediation. Efforts are also underway to define what it means for students at earlier grades to be “on track” to college and career readiness by the end of high school. Interestingly, early achievement level descriptors for NAEP included a predicted component – prepared for success at the next grade level. [get exact quote]. The Governing Board quietly abandoned this component when it became clear that it would be nigh impossible to establish the validity of such score interpretations for an assessment that did not provide individual student scores and only tested at every 4th grade level. We’ve made a good start, but much more work needs to be done to
establish and measure more specific criteria of success at the next level and to build an empirical base of the relationship of these criteria to scores on our assessments.

[Slide 22]

The idea of assessing career and college readiness is not a new concept. The Chinese administered a written test at far back as 605 under the Sui dynasty to select candidates for civil service. In this country, we developed the Army Alpha and Beta tests to select candidates for military service during World War I. John Flanagan led an Army Aviation Psychology Team during World War II, developing tests to select pilots, navigators, and other air crew members. While not necessarily a direct consequence of these testing programs, it may be noted that the Chinese society flourished for a very long time and that we won both World Wars.

[Slide 23]

Similarly, tests have been used for college admissions for some time, beginning as early as 1901 with the College Board’s predecessor of the current SAT. What is new, is the attempt to simultaneously assess mastery of the elementary and secondary curriculum and predict important post-high school outcomes, all with the same battery of tests.

[Slide 24]

While work is underway to define levels of readiness for college and careers, much more is needed to achieve the level of predictive information sought by policymakers. We need to define outcomes of interest more clearly and collect additional data to link them to test score levels. We need research to see if the Common Core
Standards cover what is needed for readiness and also, how readiness assessment should relate to high school graduation.

[Slide 25] A second approach to providing meaning to scale scores is to provide specific descriptions of what students at a given score level can or cannot do. Performance level descriptors have evolved along with standard setting processes to provide a general description of what students at a given score level can do. Note, however, that evidential support for these descriptors is still primarily judgmental.

The concept of learning progressions offer the possibility of combining specific descriptions of what students at a given level can do with predictive statements about readiness for the next step in the progression. The Common Core State Standards were designed around the idea of grade to grade progressions and work is underway to define and test hypotheses about specific chains of progression. One issue is that in mathematics, there many related skills and areas of knowledge. It may be that progressions are better defined within specific areas of content and or for specific skills. A single number or latent class may not fully summarize a student’s standing on each of these subareas. For reading, much of the progression from grade to grade concerns text complexity. While Lexiles are increasingly used to relate text complexity to grade level expectations, there is much work yet to be done on relationships between text complexity and the performance of key tasks in reading.

[Slide 26] Better Normative Information
[Slide 27] Recall that we have been producing normative information for a long time. My fifth grade Differential Aptitude Test results were reported in terms of percentiles. We know how to identify an appropriate reference population, draw a statistical sample of that population, administer the test or assessment to this sample, and compute percentiles. So what more remains to be done to satisfy policy-makers endless appetite for comparative interpretations of test scores.

[Slide 28] There are at least three areas where further work is needed. First, what really is the reference population to which individual and group results are compared? For educational measurement, we would like to think it is “all students” (in a certain grade, of a certain age, etc.). When NAEP reports the percentage of students estimated to be proficient, we think they mean “all students”. But really, we are limited to children in school, although you could argue that children not in school are not students, but the distinction between children and students is not always made clear. Sometimes, as with the NAEP state results, we mean all children in a given grade in public schools, and comparisons of results across states may be biased by differential rates of enrollment in private school. The final limitation of our reference population is that we really mean all public schools in a given grade who can be tested. In fact, between two and ten percent of the students in each state are excluded from NAEP testing and thus from the reference population because they require accommodations that NAEP does not provide. Many are most of these students are included in state assessments when alternate assessments are included, but results from alternate and regular assessments are rarely combined and usually not reported on comparable scales.
Here I want to insert another shout out to my mentor, Don McLaughlin, who developed an approach for imputing scores for students who could not participate in the regular NAEP assessments so that these students could be counted too. The “Full Populations Estimates” resulting from this approach are now listed in appendices to some NAEP reports, but are not yet widely accepted and used. NCES is working on a compromise, call Expanded Population Estimates, where the approximately one percent of students who have severe cognitive impairments are excluded, but scores are imputed or estimated for all other students.

[Slide 29] A second issue with our current norms is that we often want to generalize beyond available samples. This is certainly the case with PARCC and Smarter Balanced where assessments are being administered to samples of states that are not representative of the country as a whole. Yet the whole idea of Common Core and the Common Core Assessments was that we could compare student results and progress across all of the states. Another area where generalization is now desired is moving from national norms to world norms so that we can compare American student achievement to achievement around the world. TIMSS, PIRLS, and PISA provide such comparisons, but only intermittently and with samples of modest size. Recent efforts to link NAEP results, which are available more frequently and with large national samples to mathematics and science results from TIMSS met with limited success. The measures are somewhat different, and test administration procedures were different, enough so that a sample of states who took the TIMSS had a different rank order on TIMSS results compared to their rank order on the corresponding state NAEP results.
At the same time that we want norms for larger populations of students, policy-makers are also asking for norms for disaggregated subsets of students, such as economically disadvantaged students, English learners, students with disabilities, or maybe even blondes over six feet tall.

[Slide 30] The final policy issue related to norms is that they change over time. The general idea of NCLB is that by 2014, all students would score above the median, or at least the proficiency standards. The issue of changing norms is confused by a mixture of good changes and bad changes. Good changes occur as instruct becomes more effective and students learn more. Bad changes occur when we teach too narrowly to the test. Sometimes, also, the demographic composition of our population shifts, such as when there is an increase in English learners, resulting in different norms for the newer population. The figure shown here illustrates how apparent improvements in test score distributions may only be illusory. When a new form of the same test was introduced, the gains nearly all disappeared.

[Slide 31] Measuring Higher-Order Skills

To meet the many technical and social issues that confront us, we need students to learn, not just simple facts and basic skills, but higher order reasoning and problem solving skills. [Slide 32] Several questions must be addressed, including (1) What exactly are these higher-order reasoning skills? (2) How can we teach them? and (3) How can we assess the degree to which students have learned these skills? This last question falls particularly to us.
In the early nineties several states, most notably Maryland, Kentucky, and California, introduced performance tasks as part of their assessment, primarily as a means to tap these higher order skills. The idea was that if we wanted to measure something like scientific problem-solving, then we should give students scientific problems to solve. The Fall 1993 issue of the Journal of Educational Measurement was devoted to discussions of these performance-based assessments.

Unfortunately, performance tasks were time-consuming, at least in comparison to simple multiple choice questions, they were costly to administer and score, and several studies indicated that they did not generalize well across different tasks. Thus to get a reliable estimate of the underlying skill, students would have to complete several of these time-consuming and costly tasks. Due largely to these findings, the use of performance tasks all but disappeared, particularly after the NCLB legislation requiring states to develop, administer, score, and report a large number of assessments for accountability purposes.

The one exception has been writing, where experts argued that; (a) writing was a critical, foundational skill and (2) writing skills could not be assessed just by asking multiple choice questions about writing. For writing, we have developed something like the ice skating model. We ask the test taker to write something and then have a panel of judges rate the results using a sometimes complex rubric to define and describe the different score levels. With careful attention to the development of rubrics and scorer
training and monitoring processes, we can generally achieve something like 70 percent exact agreement and 99 percent or better agreement to within one score point on a six point scale.

[Slide 34] Evaluation and reasoning are a more critical part of the Common Core State Standards for reading and mathematics and also the Next Generation Science Standards (NGSS). Both the Smarter Balanced and PARCC assessments include performance tasks. Initial analyses of data from last year’s field tests have been sobering, with relatively low student performance on these tasks. Generalizability and dimensionality analyses of these tasks have not yet been published, or in at least one case even completed. A great deal remains to be done to demonstrate that these are valid, reliable (generalizable), fair, and cost-effective measures of the higher-order skills we seek to assess.

One reason that I believe has slowed the development of measures of higher order skills is a tendency to value output over process. If they got the right answer, should we the process that led them to that answer. Conversely, how can we give more credit to a wrong answer with the right process than to a correct answer with an incorrect process. I worked for many years on developing measures of performance for military jobs. People thought this was exciting, as we used real performance tasks. But then we would explain that performance was scored as a sum of go’s or no go’s for each step in a tightly prescribed process. People did not think educational tasks could be so easily routinized and so stopped listening. However, process is something we can
teach. In writing an essay for example, there are a series of steps beginning with selecting the theme or message, creating an outline, writing a first draft, and then editing and revising. Each step has important features for students to attend to. In the past, with paper-and-pencil testing, we only had the final output to judge. With computer-based testing, however, we have the capability of recording and scoring a lot more process information, leading to scores that reflect the extent to which students learned and effectively executed processes that were taught. Also, attention to process may provide important diagnostic feedback of where students go astray and what they need to do to improve execution of particular parts of a process. So, can we teach scientific reasoning or critical evaluation as a process? I guess we’ll see.

[Slide 35] Providing Better Diagnostic Information

Students may spend a significant amount of time taking tests. To justify this effort, teachers and others believe that tests should provide useful diagnostic information. What did the student fail to understand? What processes were executed poorly? Where were the flaws in reasoning?

[Slide 36] Two general approaches to providing better diagnostic information are: (1) improved modeling of multiple skills and (2) better models for progression along a single learning dimension.

Most state assessments are built around blueprints that specify a fixed number or range of items for each content standard or group of content standards. So the tendency is use the items for each group of content standards to create subscores that
might assess individual student strengths and weaknesses. Technical advisors rail against this approach – the individual scores aren’t very reliable, the items are scaled as if there is a single underlying dimension and pretending that there are multiple dimensions to measure does not make it so. More importantly, teachers and others will interpret differences between different subscores as meaningful when such differences are almost always within the range of measurement error and so not meaningful for making instructional decisions.

One question about diagnostic information is whether we should focus on what the students do know and can do or what they don’t yet know or can’t yet do well. Most of the current cognitive diagnostic models are build around a paradigm that says it takes multiple skills to answer each item correctly. We create a skills by item Q matrix indicating which skills are required for each item and then work backwards from patterns of correct and incorrect responses to indicate the skills that a test taker does or does not have. In other contexts, diagnosis is more about explaining why things go wrong than why they go right. Some researchers are not focusing on the incorrect answers and, if a test is properly constructed, even multiple choice items can suggest specific misunderstandings.

[Slide 38] An alternative to multidimensional models of knowledge and skills is embodied in the idea of learning progressions. Learning progressions may be a way of charting student progress along a single underlying, possibly discrete dimension, where
the diagnostic information comes from knowing what students need to work on to reach the next stage.

[Slide 40] In Conclusion

In conclusion, I believe that policy makers need us to step up our gain with respect to attaching meaning to our score scales, developing more comprehensive normative information, assessing higher order skills, and providing better diagnostic information. Whether this assessment of policy demands is the exact truth or not, I cannot say. It is really only an estimate.

[Play sound on Slide 41, and show Slide 42]