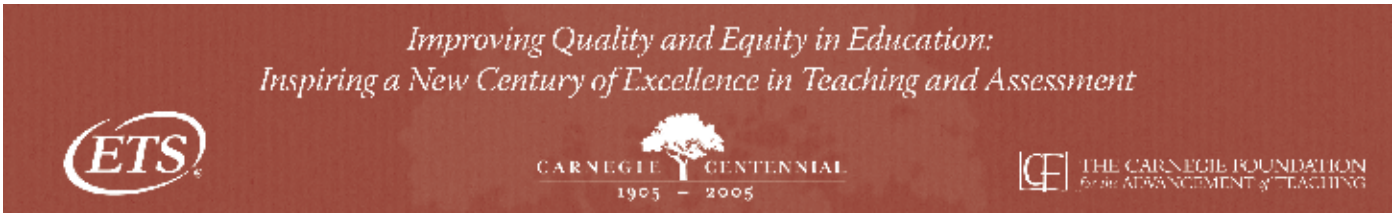
POLICY NOTES

News from the ETS Policy Information Center

Volume 15, Number 1 | Policy Evaluation & Research Center | Winter 2007



Session I. Uses of Assessment in Influencing the Outcomes of the Nation's Broad and Diverse Population

Jerome Karabel, Professor of Sociology at the University of California, Berkeley, and author of the recently published book, *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton*, proceeded from the social consequences of university admissions policies to the issue of the inequality of opportunity as it relates to the 100 to 150 most selective U.S. higher education institutions. Using test scores and biographical data from the College Board, he examined the role of standardized testing in promoting social mobility and "social reproduction" (the transmission of privilege from generation to generation). His four-part presentation focused on (1) the socioeconomic composition of the nation's elite colleges (which educate only 4 percent of the college-age cohort);

ETS President Kurt Landgraf (far right) opened the conference by thanking ETS's founding organizations, represented by (l. to r.) Gaston Caperton, President of the College Board; Lee S. Shulman, President of The Carnegie Foundation for the Advancement of Teaching; and Ellen R. Babby, Vice President for Advancement, Membership, and Planning for the American Council on Education.

This Issue — *Highlights From the Carnegie Centennial Conference — Improving Quality and Equity in Education: Inspiring a New Century of Excellence in Teaching and Assessment.*

ETS honored one of its parent organizations, The Carnegie Foundation for the Advancement of Teaching, with an invitational conference in Princeton, NJ, on June 1–2, 2006. The gathering commemorated both the centennial of The Carnegie Foundation (1905–2005) and its role, along with that of the American Council on Education and the College Board, in creating Educational Testing Service in 1947. Leaders of all four organizations welcomed invitees to the two-day conference, which included presentations by U.S. and international educational leaders.

Speakers and audience members examined assessment trends and effects from a number of perspectives, and discussed policies and practices that can lead to

(continued on page 4)

(2) differential performance on the SAT® by socioeconomic status (SES) and race (with special attention to the class and race of high-scoring students); (3) the small but present pool of high-scoring students with low SES; and (4) current attempts to increase the socioeconomic diversity of the nation's elite institutions.

Karabel demonstrated that the lowest SES quartile contributes only 3 percent of the students admitted to selective colleges, while the top SES quartile contributes 74 percent. The SAT score distribution by parental income is similar. Students whose parents have college degrees are four to six times more likely to score above 650 in Math and Verbal on the SAT.

Karabel noted that while institutions have developed admissions guidelines that take into account race, alumni legacies, and athletic prowess, most have not yet attempted to diversify their student body with respect to their students' SES. He proposed a new direction in higher education admissions that would be both class- and race-sensitive. To implement a policy that would change the distribution of opportunity in higher education, he suggested that we measure social class more accurately by looking closely at three dimensions: (1) family (not only parents' income and education, but also occupation, net worth, and grandparents' education); (2) neighborhood (socioeconomic composition of applicants' zip codes and census tracts); and (3) school (public and private).

Based on these data Karabel advocated the development of an algorithm to predict a median score on the SAT and SAT II. Students who are intensely or moderately disadvantaged along the dimensions listed above, and who perform considerably better than expected on the standardized tests, would be identified by this algorithm. This information would be stored in a database that colleges seeking to diversify their admissions cohort along SES dimensions could use in their recruitment process. Identifying these students and publicizing the database would encourage elite institutions to enroll more students from disadvantaged backgrounds. And, ideally, this would help students believe that equal opportunity is a reality in elite college admissions. It could enhance social mobility and, in the process, give new life to the American dream.

Alden Dunham, former Admissions Director for Princeton University, and **Neil Grabois**, Vice President and Director for Strategic Planning and Program Coordination at the Carnegie Corporation of New York, served as respondents for Karabel's presentation. Dunham offered comments on the triumph of vocationalism in higher education, the inconsistency of Division I athletics with university values, the promise of distance learning as a democratizing force, and the major advances neuroscience is bringing to assessment. He also registered his disagreement with what he called the "sinister cabals" described in Karabel's *The Chosen*, and opined that elite college admissions practices are of concern to relatively few people, chiefly those residing in the Northeastern United States. Dunham advocated gender-based affirmative action for boys, whose enrollment in higher education is in steady decline.

Grabois, on the other hand, was inclined to agree with Karabel's conclusions about elite colleges' past admissions policies that discriminated against certain groups of students. He also lauded the capacity of admissions officers to recognize

talent, stating that imaginative institutions can help students move through social class. Since the 1980s, he said, some admissions officers at selective colleges have recognized a need for "affirmative action for poor people." Such a program is costly, but institutions such as Amherst College are pursuing funding for such an approach. With respect to the potential of higher education to improve students' social class, Grabois' main concerns are the distribution and redistribution of wealth that has taken place over the past 15 years, and the deterioration of the K–12 school system.

Session II. Uses of Assessment in Institutional Accountability and Action

Within the context of current education policy discussions about the merits of standardized testing in higher education, **Richard Shavelson**, Professor of Education and Psychology at Stanford University, offered "A Brief History of Leadership in Assessing Undergraduates' Learning." His historical overview began with the genesis and development of college learning assessment and its early proponents at The Carnegie Foundation. Shavelson's discussion covered the origins of objective testing (1900 to 1933); assessment of learning in graduate education (GRE®: 1933 to 1947), the increasing numbers of test providers, including the founding of ETS in 1947 (1948 to 1978); and the era of external accountability (1979 to present).

Michael T. Nettles, ETS Senior Vice President for Policy Evaluation and Research, welcomed participants to the conference to celebrate The Carnegie Foundation centennial.

Shavelson noted that one of the ironies of this history is that early 20th-century educators saw objective testing as a way to distance the examinee from the examiner, to probe for content knowledge, and to get away from the tradition of essay-writing for college entrance examinations. Yet, today, written examinations are once again at the forefront of contemporary learning assessment reform. In an attempt to assess reasoning ability and the value added by college learning, test creators are now using technology to design, develop, and score examinations that require performance tasks such as analyzing complex material and providing written responses. Shavelson suggested that The Carnegie Foundation might conduct research on how assessment information from external testing could be integrated into institutions' efforts to use assessments to change and improve teaching and learning.

(continued from page 1)

equity and excellence in education. The conference sessions featured prominent educators and academic leaders speaking on the following topics:

- Uses of Assessment in Influencing the Outcomes of the Nation's Broad and Diverse Population
- Uses of Assessment in Institutional Accountability and Action
- Reliance on Assessment for Judging the Quality of Educational Systems
- Case Study of Inventive Uses of Testing: National Board for Professional Teaching Standards®
- Leveraging Powerful Teaching: The Importance of Performance Assessment
- A Union of Insufficiencies: Measurement, Assessment, and Judgment in Supporting the Future of Educational Quality

This issue of *ETS Policy Notes* offers an overview of the sessions.

Linda Tyler, Group Executive Director for New Product Development in the ETS Higher Education Division, picked up the theme of improvement in teaching and learning in her presentation, "Collecting Evidence for Action: Help From a Test Design Methodology." She suggested that institutions might consider using ETS's evidence-centered design (ECD) methodology as they work on student-learning outcomes. ECD can help higher education institutions collect evidence and use it to improve learning. Tyler outlined the steps in the ECD methodology:

- Step 1: Claim. Determine what a successful student should know or be able to do upon completing a given course. Further measurement will always be against this claim.
- Step 2: Develop the evidence model. Define the evidence needed for supporting the claim. Determine what evidence, and how much of it, would be needed to make the claim (e.g., What am I preparing my students to do? How should I design my assignments and tests?). What would be the best evidence possible to support the claim? Faculty must deconstruct the claim statement into various competencies and then decide what evidence is necessary to satisfy the

Commemorating the centennial, ETS President and CEO Kurt Landgraf presents a grandfather clock to Lee S. Shulman, President of The Carnegie Foundation for the Advancement of Teaching.

claim that the student has mastered the competency.

- Step 3: Design the task. Create assignments and assessments that serve as evidence-collecting devices. To do this, faculty need to ask themselves questions about, for example, whether to have weekly writing assignments, short-answer or essay tests, an in-class test, or a take-home exam. There is usually a summative assessment, and this must be as close as possible to the integrated claim statement.

Other, interim assessments are used to build confidence and gather evidence of student achievement. Tyler stressed that ECD can be useful not only for making choices in the classroom, but also at the program or institutional level, to collect evidence that will improve instruction and curriculum.

The role of testing in accountability was at the center of remarks by **Henry Braun**, Distinguished Presidential Appointee at ETS.[1] His presentation was titled "On Test Quality… And Beyond." Braun set the context by examining the key goals (access, equity, quality, and efficiency) and the means (infrastructure, funding, human capital, regulations, and oversight) of education policy. Within this framework, he noted that society sets the policy goals through legislative actions and bureaucratic interpretations. Testing serves as an instrument of education policy through regulations and oversight; but its main function is to address issues of quality. Test results can be used to define the standards by which quality is determined (e.g., proficiency can be defined as obtaining a score of 80 percent or better on a test). They can also pinpoint where quality is lacking and identify possible problems (e.g., students in a particular school are having trouble with algebra).

Recently, however, test results are being used for less benign purposes. They are used for institutional and/or teacher accountability, based on a logic that views evidence of student learning as an indication of the quality of schools and teachers. A current example of this is a relatively new methodology called value-added modeling (VAM). VAM attempts to isolate the contributions schools and/or teachers make to student learning, as measured by patterns in test-score trajectories. Braun cited methodological concerns related to interpreting the output of a value-added analysis as an accurate indicator of school or teacher effectiveness. The concerns relate to the problem of making causal inferences from observational studies. He noted that teachers and schools do not come together in random ways: All sorts of selection bias effects can confound any estimate of effectiveness.

Moving to a discussion of test quality, Braun advocated a broad view that combines measurement, politics, and a value system. In this context, he quoted Campbell's Law[2], which says: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

With respect to testing, this paradox can be rephrased as: The better tests we build, the more we want to use them; and the more we

[1] Braun is now the Boisi Professor of Education and Public Policy at Boston College.
[2] Donald T. Campbell, *Assessing the Impact of Planned Social Change*, 1976

use them, the more likely they are to result in unintended consequences. The less we use them, the less anyone is interested in their results. Braun stated that poor test quality increases the chances that consequences of Campbell's Law will occur; but poor assessment system design and poor implementation also invite resistance and corruption. He urged attendees to think not in terms of test quality or test validity, but rather in terms of systemic validity. He said that education policies are systematically valid if they result in decisions and actions that lead to progress toward one or more intended goals without causing regression from other goals. Braun argued that we must recognize the need for an interactive policy-design process in which original design decisions are modified in light of analysis of alternative scenarios and anticipated costs.

Braun concluded that school and teacher accountability systems can play a constructive role in improving student learning, provided reasonable steps are taken to enhance systemic validity. He admitted that the technical and political challenges of establishing such systems are made more daunting by ideological polarization in both educational research and policymaking.

Ellen R. Babby, Vice President for Advancement, Membership, and Planning at the American Council on Education, helped to kick-off the conference.

Freeman Hrabowski III, President of the University of Maryland, Baltimore County (UMBC), was the featured dinner speaker. Asked to address the topic of assessment and access, he posed the question: How can we encourage the academy not to be content with the status quo?

Higher education, which has not traditionally held itself accountable, needs to be constructively self-critical, he said. But that requires data — and, so far, we haven't been able to collect data on student learning, nor have we been able to document what students know. While positing that higher education enrollment will continue to grow (by approximately 15 percent over the next 15 years), Hrabowski drew attention to low and declining retention rates. He predicted that retention rates will continue to decline if we don't change the culture of higher education. At the heart of the matter is the tension between the public's understandable demand for accountability and the faculty's historic mistrust of external measures. To rectify that — to improve education, assess what students are learning, and insure that they stay enrolled and make progress toward their degrees — we must try to find a way to bring these two sides together.

Two issues are paramount:

- How can we most effectively assess the quality of teaching and learning in our institutions?

- What can we do to insure that more students succeed in college, even using the most basic of measures, such as retention and graduation?

Hrabowski cited the Collegiate Learning Assessment (CLA) project and the National Survey of Student Engagement (NSSE) as worthy examples of attempts to gather data on what students are learning and thinking. It is important for campuses to have faculty who are always thinking about continuous improvement in instruction. The challenge is to persuade faculty to become involved in this discussion.

Assessment and access in the STEM fields (science, technology, engineering, and mathematics) are particular areas of concern, Hrabowski said. Too few students enter college with interests in the STEM fields and even fewer complete STEM majors — and among those who do graduate, many express the desire to leave the field. Forty-three percent of STEM doctoral degree recipients are international students. We cannot continue to rely on international doctorates, said Hrabowski; we need to increase the number of native-born students in the STEM fields, and as the U.S. minority population grows, it is essential to train more people of color for careers in the STEM fields.

Changing how STEM subjects are taught is crucial for increasing the number of students majoring in these fields. Years ago, the National Academy outlined some steps that need to be taken to prepare people in the life sciences. Institutions should re-examine their courses and teaching approaches and consider providing more courses in mathematics, physical and computational sciences; and more interdisciplinary laboratory work. Unfortunately, few higher education institutions have done anything about these recommendations, Hrabowski said. Resources and attitudes account for why not much has changed in science education, even though it's been clear for years what changes need to be made to curriculum and instruction.

To encourage change, Hrabowski said, people both inside and outside of the academy need to work together to find valid methods for measuring student learning. To this end, he said, it is critical to both know the data and understand the issues. We have to understand why, for example, at UMBC, half of our students are failing chemistry, and we need to consider whether changing how we teach chemistry could improve this outcome. We looked at what works with teaching STEM subjects to minority students and then used this approach for majority students; now we are moving from the sciences to other areas. The university gathered evidence through focus groups with faculty who teach students from diverse backgrounds, and through focus groups with students who were asked to characterize what they consider "good teaching."

Hrabowski stressed that this approach will promote the use of assessments to bring rigor to the discussion about how to help students complete their education and remain in the STEM fields.

Session III. Reliance on Assessment for Judging the Quality of Educational Systems

Juergen Baumert, Director of the Max Planck Institute for Human Development at Humboldt University in Berlin, began his remarks by noting that results of large-scale multinational assessments have triggered significant developments in national educational policy in various countries. In his presentation,

"International Comparisons at the Transition to Adulthood," Baumert contended that the worldwide standardization of schooling in terms of time structure, social organization, and content has intensified notions of competition and accountability. The latest generation of international educational assessments boasts a strong theoretical framework, alignment to internationally shared standards, high-quality test items, and proficiency scaling with a broad array of items to illustrate competence levels. Baumert noted that ETS has played a fundamental role in the international standardized testing movement.

But Baumert pointed out that international assessments such as the Trends in Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) possess both strengths and weaknesses. They can serve as benchmarks of what can realistically be achieved in compulsory schooling, monitor school system development, identify at-risk populations in terms of what students are expected to be able to do but cannot do, and compare inequality structures in educational outcomes. They are limited, however, in how much they can tell us about education in any given country. They may offer misleading conclusions on the basis of limited analyses and, due to small sample sizes on the country level, they do not include theoretically-based multivariate modeling.

To compensate for these weaknesses, international cross-sectional surveys are sometimes embedded into a broader analytic research program on teaching and learning. Examples include the TIMSS video component in a longitudinal framework, which examines the surface structure of classroom interaction and the logic of learning mathematics, and an expansion of the PISA 2003 Study to a longitudinal design focusing on teacher expertise. Using these two examples, Baumert examined teachers' content knowledge in mathematics, as well as their pedagogical content knowledge in mathematics, asking the question — Does pedagogical content knowledge contribute to student learning? He concluded that while insufficient content knowledge limits pedagogical content knowledge, content knowledge alone is not sufficient to make mathematics accessible to students. In ways such as these, international large-scale assessments can promote knowledge about how to improve teaching.

Continuing the discussion of international assessments, **David Baker**, Professor of Education and Sociology at Pennsylvania State University, asserted that over the past 15 years these assessments have assumed an important role in educational policy debates. His presentation, "The Good, the Bad, and the Future," provided insights into the dimensions of the worldwide education revolution. This movement toward a schooled society, encompassing a global expansion of all levels of education, and especially higher education, has made these multinational assessments routine. PISA, for example, is conducted every three to four years.

There are also multinational studies of teachers and teaching that reveal a very small percentage variance in approaches and teaching methods across countries. While differences between individual teachers may be great, no generalized picture emerges of the teachers of any given country. For example, all math teachers across a variety of countries use, more or less, the same

Governor Gaston Caperton, President of the College Board, encouraged conference participants to embrace change in education.

approaches and hold similar beliefs about their students. Country differences emerge in terms of outside tutoring, however. Baker cited the situation in Korea, where tutoring franchises are growing and families voluntarily pay $.85 for tutoring for every dollar the government spends on K–12 education.

Another assessment, the 1999 International Civic Study of Political Knowledge, Skill, and Attitudes, surveyed 14-year-olds in 28 countries (including Eastern Europe and the United States, but not including Asia or the Middle East) on their civic education knowledge. The study found that the nation in which the students were raised had no influence on the extent of their civic knowledge. While the nature of the particular political regime did not influence the production of an informed citizenry in a given country, schools and schooling qualities did. These qualities related to a democratic classroom, educational expectations, teachers with civic experience and civic training, and availability of a civics curriculum. In contrast to the findings for mathematics and science mentioned earlier, this study revealed differences in teacher effects.

Unfortunately, said Baker, a great deal of political pressure surrounds these international assessments, sometimes with deleterious effects. The results of the first TIMSS assessment caused education experts to charge that there was a major crisis in mathematics in the United States. Although the study seemed to indicate that the nation's math curriculum was broken, the truth was that this debate pushed aside any discussion of the inequality of educational resources in the country.

International assessments are here to stay, noted Baker. For better or for worse, we will continue to use these assessments to compare ourselves with other countries.

Matthias von Davier, Senior Research Scientist in the ETS Center for Psychometric Infrastructure, carried the international theme forward by turning his attention to the large-scale educational assessments themselves, their models, and their targets of inference. He noted that the National Assessment of Educational Progress (NAEP) served as a model for the TIMSS, PISA, and Progress in International Reading Literacy Study (PIRLS) assessments. These instruments offer group-level reporting only and do not report individual student scores, and they have minimum reporting group sizes. From NAEP, they inherited some design principles that make it more difficult to do cross-country analysis or analysis for small-population countries (due to small per-country sample sizes). While multinational assessments can serve many purposes, they cannot substitute for national assessments. Some of the reasons for this are:

- The cross-sectional design of multinational assessments does not allow for measurement of growth or change.
- They are descriptive and non-experimental, and therefore no causal inferences can be made from them.
- They are designed for broad-construct coverage and group-level reporting at a single point in time, with item response theory linking across cycles for describing trends.
- They are not effective for comparing achievement across grade levels.

A number of important issues were identified by von Davier that must be addressed when designing the international tests. They include coverage of curriculum and choice of measurement model; guaranteeing a diverse sample of countries (developing and industrialized countries); comparability of measures, national adaptations, and differential item functioning; and differences in the meaning or interpretation of background data. For the latter, he cited as examples two possible interpretations of the presence of animals in a household: signifying either a wealthy family or a poor rural family.

Asserting that international comparisons such as TIMSS and PISA actually tell us rather little about teachers and teaching, **Suzanne Wilson**, Professor of Teacher Education and Director of the Center for the Scholarship of Teaching at Michigan State University, concluded that we still lack a model for capturing the dynamic of teaching. She examined the questions asked of students and teachers in these surveys and found that the questions did not represent what it takes to be an effective teacher. Wilson disagreed with Baker's contention that there are minimal differences in teachers and teaching across countries. She commented that the surveys do not have the capacity to capture teacher knowledge, beliefs, and practices that can have an impact on learning.

Wilson spoke of methodological and conceptual challenges associated with these large-scale multinational assessments. For methodological challenges, she noted that this assessment approach involves self-reporting, which can be suspect, and that the surveys do not get close to identifying individual expertise or the value added by a particular teacher. She faulted the questionnaires for their flat dimensionality, in contrast to the multidimensional TIMSS video study.

Wilson called for a conceptual map of teaching. Noting the attention paid in the United States to what is taught and how it is taught, she found that these assessments co-mingle these two issues. In fact, she found that these assessments entangle content knowledge with teacher beliefs and practices. In Wilson's view, TIMSS does not allow an understanding of what teaching is.

The surveys contain technical and structural features of teaching, but do not capture what the teacher does to activate engagement in learning. For this, she said, we need smaller-scale studies that get closer to the student and the teacher. Assessments drive instruction and help us to reconceptualize what's going on in the teaching-learning process; they also represent

what we think teaching is. But large-scale assessments focus on educational practices, which do not represent the ethics of teaching or its interpersonal and moral aspects. Wilson suggested that creating smarter large-scale assessments and using them sensibly is an important goal for the educational community.

Session IV. Case Study of Inventive Uses of Testing: National Board for Professional Teaching Standards®

Lloyd Bond, Senior Research Scholar at The Carnegie Foundation, traced the origins of the National Board for Professional Teaching Standards® (NBPTS®) and described how this assessment of teaching practice was developed. In his presentation, "Assessing Accomplished Practice in Teaching," Bond noted that the influential 1983 report, *A Nation at Risk*, put education on the national agenda, but teaching and teachers were not part of the ensuing discussion. The Carnegie Foundation responded in 1986 with the publication of *Teachers for the 21st Century*, which made a number of recommendations, one of which was to establish the NBPTS, which occurred in 1987. The following six years were spent in defining professional standards, with the first pilot assessments conducted in 1993.

Concerns were raised by a number of constituencies (such as higher education and unions), but in the end, the educational community supported NBPTS. Owing to the work of a group of visionaries, psychometricians, and assessment development and scoring experts, an assessment was created that met what Bond called the essential validity challenge: Being an accomplished teacher versus demonstrating it in a formal assessment. The assessment's developers decided that the assessment would be voluntary and confidential, and that applicants had to be K–12 teachers with at least three years of experience in the classroom. The group identified five core propositions (above) as well as a set of underlying assessment development principles (below).

Five Core Propositions

1. Teachers are committed to students and their learning.
2. Teachers know the subjects they teach and how to teach those subjects to students.
3. Teachers are responsible for managing and monitoring student learning.
4. Teachers think systematically about their practice and learn from experience.
5. Teachers are members of learning communities.

Two important philosophical underpinnings guided the team's work: (1) both preparing for and undergoing the assessment should be "deeply educative" experiences; and (2) teachers themselves should have primary control over the definitions of quality and "accomplished practice" and the determination of who meets the desired

Underlying Assessment Development Principles

- Tasks should be authentic and therefore complex.
- Tasks should be open-ended, allowing teachers to show their own practice.
- Tasks should provide ample opportunity for analysis and reflection.
- Knowledge of subject matter and knowledge of students should underlie all performances.

standards of quality. The team identified two settings for the assessment: one in the classroom itself (the teaching portfolio) and one in an assessment center. The teaching portfolio was to reflect the richness and complexity of teaching in real classrooms. Considerable latitude was allowed in choosing assignments to feature, and actual student work and student feedback were to be included. Time was set aside for reflection and analysis. At the assessment center, candidates were presented with on-demand tasks that gauged content knowledge and pedagogical content knowledge.

The psychometric challenges for this new assessment were considerable. They included construct underrepresentation, construct irrelevant variance, scorer training and calibration, and adverse impact and bias.

Bond also noted some significant challenges for the future of the assessment. Among these are the demands of measuring teacher performance in the technology-equipped classroom of the future, in the distance learning environment, and in the context of ever-increasing student diversity. The ultimate issue, however, will be to relate teacher performance on the assessment to student learning.

Building on the historical overview offered by Bond, **Joseph Aguerrebere Jr**., President and CEO of the NBPTS brought the participants up to date on National Board Certification. He asserted that the current assessment truly does cover the complexities of teaching. By developing standards and then developing an assessment system to measure teachers' performance against those standards, the National Board has been able to capitalize on the assessment results to improve American education. Aguerrebere reviewed the requirements for the teaching portfolio, the constructed-response exercises in the assessment center, and the 200 to 400 hours necessary for preparing for National Board candidacy.

The NBPTS has contributed to the "wisdom of practice" by examining teaching practice in order to improve it through capturing and documenting evidence in a teaching portfolio, and by facilitating learning communities among teachers. The assessment forces teachers to evaluate their success and understand the components that have had an impact on student learning. It also has given rise to a "language of practice" — a common vocabulary for describing what teachers do.

As of June 2006, there were 47,500 Board-certified teachers, half of whom had achieved this status in the past three years. (By January 2007, that number had climbed past 55,000.) The certificates cover 90 percent of teaching areas and are reviewed and revamped in a set cycle to ensure they are current and aspirational. The framework and standards have also influenced higher education. Some 500 higher education institutions are using aspects of NBPTS in their teacher education programs.

Session V. Leveraging Powerful Teaching: The Importance of Performance Assessment

Quality teaching is recognizable, and it matters in terms of student achievement, stated **Linda Darling-Hammond**, the Charles E. Ducommun Professor of Education at Stanford University. Changes in the field of teaching and in student

demographics in the 21st century mean that powerful teaching is needed more than ever, and the right kind of teaching assessment can act as a lever to produce better teaching. Darling-Hammond drew attention to the huge inequality of resources devoted to education in the United States, where K–12 graduation rates are low compared to those of other developed nations. She noted the close correlation between failing in school and the school-to-prison pipeline.

It's easy to teach children who already know what you want them to learn, already know how to learn, and have educated parents and the financial resources to hire tutors if necessary, she said. It's far more challenging to teach those without these advantages. Given the reward structure in the teaching profession, Darling-Hammond noted that schools tend to allocate the easiest students to the most experienced teachers. But these are not the students who most need experienced teachers — those teachers who understand content in all the ways students can understand it and who can plan around both the demands of the content and the needs of the students. There is a tendency to associate the success of easily taught students with what it takes to be a successful teacher; but powerful teachers have a repertoire of teaching strategies, assessment strategies, and the ability and disposition to reflect on learning and practice and adapt to what the students are learning, she said. The evolution of standards of practice has helped to articulate what this kind of teaching looks like.

The National Board has created an authentic representation of teaching, built on a base of evidence. Darling-Hammond reflected on the "courageous moment" when those involved in developing the NBPTS assessment rejected an early prototype that did not take student learning into account. The Board's standards have had a wide-ranging effect on the profession, influencing the characterizations of teaching in the Interstate New Teacher Assessment and Support Consortium (INTASC) standards[3], in the standards put forth by the National Council for the Accreditation of Teacher Education (NCATE), and in teacher education programs. She noted that assessments, too, can leverage change in the profession, but they must be true assessments of teacher performance, not multiple-choice tests or subject-matter tests. Such assessments must be embedded in teacher education programs.

Thanks to NBPTS, she said, we have the language and a forum where exchanges on teaching practice can occur. Shared norms of practice across the profession can serve as a vehicle to help transform the preparation of teachers. These shared norms and practice will also help the public perceptions of teaching and the debate on what makes quality teaching. Darling-Hammond concluded her talk with the hope that in the future we will be able to assert, in a revised reprise of an old adage: Those who can — do. Those who understand — teach. And those who can't — go into a less significant line of work.

Session VI. A Union of Insufficiencies: Measurement, Assessment, and Judgment in Supporting the Future of Educational Quality

As a preamble to Lee S. Shulman's concluding presentation, ETS Vice President **Ida Lawrence** called the audience's attention to a number of research efforts under way at ETS to support the use of assessment in the service of teaching. She described five areas of focus for the Research & Development Division:

[3] These standards reflect the requisite knowledge, skills, and attitudes necessary for teachers starting their career.

1. New Constructs and How to Measure Them. These are both the familiar cognitive measures, such as critical-thinking skills and communication skills, as well as the so-called non-cognitive skills, such as dependability, persistence, and teamwork.

2. How to Improve the Quality of Teaching. ETS is developing new teacher professional development materials and products, including formative assessment techniques. These should help teachers along the teach-assess-teach-assess progression.

3. English Language Learners. Given the U.S. demographic projections and the growth in English as the international language of business and communication, ETS is developing new assessments and learning tools for non-native speakers of English.

4. Reading – Math – Writing Assessments. Researchers are working on assessments for the K–12 market that are more construct-rich and cognitively based; these formative assessments in reading, writing, and math are intended for use by teachers to help students make progress.

5. Technology Tools for Scoring. Devising automated ways to score writing and speech samples is the focus of another group of ETS researchers. Automated scoring technologies applied to our own assessments will lower their costs.

Lee S. Shulman began his concluding discussion by noting that he had deliberately recycled its title, "A Union of Insufficiencies," from an article he had written when he and colleagues were developing the National Board assessment. What they discovered as they thought about how to measure teaching is that it is too complex a process to be judged by a single metric, too nuanced and too rich to be evaluated by a single method. Multiple methods and multiple indicators are required. The powerful tools that are most appropriate for measuring a particular thing are, by design, incomplete or insufficient for measuring other things. With this in mind, those working on the Board certification set themselves to developing a collection of assessment tools and indicators ("a union of insufficiencies") that could be used to judge the quality of teaching.

Shulman reflected on the 1948 meeting of college examiners at the American Psychological Association conference, where they discussed the need for a shared lexicon and conceptual framework for evaluating undergraduate general education. They were looking for diagnostic tools to assess what was being learned in general education, but they discovered that there was a serious mismatch of what was being taught and what was being learned and assessed on their campuses. Led by Benjamin Bloom, these discussions eventually gave rise to Bloom et al. Taxonomies, with their revolutionary cognitive and affective domains:

Bloom et al. Taxonomies

Cognitive	Affective
Knowledge	Receiving
Comprehension	Responding
Application	Valuing
Analysis	Organizing
Synthesis	Internalizing
Evaluation	

This became the shared lexicon and broad understanding the examiners were seeking. Shulman emphasized that a broad, comprehensive view of the goals of education must precede the design of measures. He noted that decisions of consequence, such as choosing a mate, buying a car, or judging the health of the economy, are done on the basis of multiple indicators. If there is a fatal flaw in how assessment is translated into policy in our society, he said, it is in the sanctification of the single indicator.

Currently at The Carnegie Foundation, work is progressing on how to assess professional learning. Three categories of learning are central to professionals: habits of mind, habits of practice/skill, and habits of the heart. These are generally assessed by visible, public performances of understanding, skill, and disposition either embedded in the course of instruction or residencies. They constitute a collection of multiple indicators.

Asserting that the past century's assessments stand outside of the learning process, Shulman described the five principles that guide the Foundation in the 21st century:

- Error of the single instrument: designed insufficiency
- Multiple indicators: union of insufficiencies
- Aggregating for validity: transparent policies of aggregation
- High stakes/low yield must give way to low stakes/high yield: timing, embedding, coaching, and repeating
- Resistance to corruption: educative assessments

The Foundation has devised a theory of action that uses assessments as:

- Mirrors (seeing your teaching and student learning)
- Lenses (seeing teaching and learning in new ways, e.g., through the National Survey of Student Engagement and the Collegiate Learning Assessment)
- Windows (seeing how you compare with your peers via windows and/or lenses)
- Reflection for Action (convening to consider data and explore options; e.g., faculty inquiry groups, NBPTS support groups)

For the future, we need to reverse the usual validity argument and collapse our comfortable distinctions between formative and summative evaluations; between high stakes and low stakes tests; between lower and higher level understanding; between cognitive, affective, and formational assessment. Shulman charged that the educational measurement field, by focusing on single indicators, has colluded in the truncation of merit and method, and he called on researchers to invent a new psychometric that represents embedded, systemically valid assessments that provide evidence of the enduring consequences of learning and not only the evanescent, passing ones.

Edmund W. Gordon, Richard March Hoe Emeritus Professor of Psychology and Education at Teachers College, Columbia University, and John M. Musser Professor of Psychology, Emeritus at Yale University made the invited commentary on Shulman's remarks. Gordon paid tribute to Shulman for his major contributions to the fields of assessment and education. Gordon expressed sympathy for Shulman's attempt at

reconciling the contradictions in the history of assessment in education, but warned against our settling for Shulman's "union of insufficiencies." Rather, Gordon offered more radical ideas:

- eliminating standardized approaches to on-demand academic performance as a legitimate pedagogical function, since such assessments may serve accountability functions but only modestly inform pedagogical intervention;
- distilling the information we need for accountability purposes from the records of assessments that are embedded in teaching and learning transactions;
- revising criteria for what it means to be an educated person to include the capacity to use complex systems of representation, the capacity to examine phenomena from more than a single perspective, the capacity to impose order onto chaotic or complex data, the ability to make sense of one's environment and solve familiar as well as novel problems, and integrating these assessment probes into the teaching and learning process.

Gordon concluded with the suggestion that the use of standardized tests to determine merit may be problematic since supporting the notion of a meritocracy could be anti-democratic in a society where opportunities to achieve merit are unequally distributed. Under such circumstances, we face the risk of reinforcing inequality.

ETS Policy Notes is published by the ETS Policy Information Center
Educational Testing Service
Rosedale Road, MS 19-R
Princeton, NJ 08541-0001
(609) 734-5949

e-mail: pic@ets.org

www.ets.org/research/pic

Director: Richard J. Coley

Editors: Linda H. Scatton, Richard J. Coley, and Amanda McBride

NON-PROFIT
ORGANIZATION
US POSTAGE
PAID
EDUCATIONAL
TESTING SERVICE

Listening. Learning. Leading.

www.ets.org