

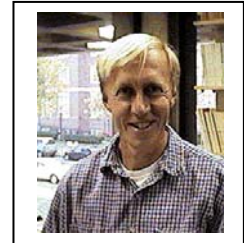


FROM THE PRESIDENT: UPDATE ON NCME ACTIVITIES!

Terry A. Ackerman, University of North Carolina - Greensboro

Dear NCME members,

The last couple of months have been pretty busy and I just wanted to update you on NCME activities since the last Newsletter.



A few weeks prior to our October Board meeting I met with Judith Rizzo, Executive Director of the Governor Hunt Institute, here in North Carolina. Former NC Governor Hunt established this institute to serve as a resource to governors throughout the country especially on issues related to assessment and NCLB. Several NCME members have served as resource for this organization. We had a good discussion about their mission and how NCME could collaborate with the Institute to further their mission.

The day prior to our Board meeting in D.C., Plumer Lovelace, NCME's Executive Director, and I met for the third face-to-face meeting with AERA's Executive Director, Felice Levine and AERA's Senior Adviser, Jerry Sroufe to discuss NCME/AERA relationships and changes we foresee in the upcoming contract renewal. Our contract with AERA expires in June 2010. We will be sharing a draft of changes we envision making, in accordance with our strategic plan, with AERA in January. I will keep you updated as this process unfolds.

Our Board meeting which was held on October 28-29, in the AERA Boardroom, was very productive. My goal was to take advantage of being in the D.C. area and talk with policy makers about logistically how NCME can serve as a resource. I was able to make arrangements for several people to meet with us. We heard from John Easton, Director of the Institute of Education Sciences, John Tanner, Strategic Initiative Director, Center for Innovative Measures at CCSSO, Natalia Pane, American Institute of Research, and Judy Wurtzel, a consultant in the U.S. DOE. Much of our first day was spent talking with these people. I think the message was very clear. First, there are quite a few members of NCME, who already provide expert advice on a regular basis to these people and organizations. Secondly, there is an important need for NCME to step up and serve as an unbiased resource when it comes to issues of assessment and accountability. Wayne Camara and I have formed a task force that will try to flesh out this process. We actually held our first phone conference and talked about the charge and timelines for providing feedback to the Board. As details of proposed logistics begin to take shape, I will keep you informed.

Besides the AERA contract, we also have two other contracts that will have to be renewed relatively soon. One is with our publisher, Wiley Blackwell. Wayne, Plumer and I just spent a day and a half up in Boston meeting with the Blackwell staff and talking about future collaboration. I think this was a very good meeting and extremely informative. We talked about the different services that Wiley Blackwell can provide that not only would help our journals but also help market NCME, increase our membership,

and develop different means to serve as a resource to practitioners and policy makers. We talked about the Wiley-Blackwell staff working with the Publications Committee to developing a multiple-year strategic plan.

Finally, our contract with our management company, The Rees Group (TRG) also expires in 2010. I think we have developed a great relationship with TRG and one that we need to continue and nurture. We need to better understand all that TRG can do for NCME. I think we have just begun to tap some of their many services, especially with our annual meeting. Since becoming president I have worked very closely with our Executive Director, Plumer Lovelace. We are extremely fortunate to have someone of his caliber overseeing the day to day running of NCME. Wayne and I will be meeting in Madison (where TRG is located) for an orientation meeting with 2011 program chairs, Sandip Sinharay and Cara Laitusis in January. During that visit we plan to meet with Susan Rees and talk about the contract renewal process.

The last piece of NCME business I want to mention is the NCME Fun Run/Walk. Plumer and I talked with Brian French who has coordinated this the past several years. This year we have decided to make it into a fun competition. There will be two categories (academia and private organizations) with impressive trophies for each. Teams can score points by finishing the run/walk, presenting papers/posters at the conference, having unique team attire, etc. We're even looking into theta-hat hats for participants. Stay tuned for more information and let the trash talking begin...

I know many of you have had a memorable year. I just wanted to share with you a brief glimpse of my life during the past couple of months since the end of October to welcome you to my world.

On Sunday, October 25th I ran the Marine Corps Marathon in D.C. I'm sure there are easier ways to see D.C. but running the twenty-six miles and 385 yard was a goal I set for myself this year. It's about my twentieth marathon.



On the way back from the marathon (we were driving my son's 2007 Chevy Cobalt which we had just gotten for him two weeks earlier), we hit a 5-point buck head-on, going down I-85 about 30 miles south of Petersburg in Virginia. The deer appeared instantly and tried to jump the car but hit the hood. The hood flew up and smashed the windshield. There are a lot of things that can be said about driving 70 mph blind because your hood is plastered against your windshield, but none are good. I was in the right lane and was able to slow down and make it to the side of the road. My wife, Debbie was doing a Sudoku and started questioning me because she thought I had run into a wall on the freeway. We called AAA and the state patrol. It took about two hours to get a tow truck from Petersburg to come because the first one they sent out also hit a deer. They took us back to Petersburg and we spent the night there. The next day we got a rental car (another Chevy Cobalt) and drove back to Greensboro. Our car was actually totaled. (Not often is it the parents who have to tell their kids they totaled their car.) We were lucky though because if it hadn't leaped, we might have run over the deer and come to a screeching halt, setting off the airbags. If it were higher in the air it could have come through the windshield. So, we were very fortunate.



Then the very next day I had to get another rental car and drive back to D.C. for three days of NCME Board meetings. You guessed it...the rental car was a white Chevy Cobalt. When I got to the same spot on I-85 where we had hit the deer, there was another mangled deer in the left northbound lane that must have been killed several hours earlier. I switched lanes to avoid the deer. But when I was next to the deer a great big pick-up ran right over the deer and sprayed deer entrails all over the driver's side of my car. When I got up to DC I'm sure people thought I was a hit and run driver. Anyway, there were no problems

returning down I-85 going back home through "deer alley." But I did spend about a half an hour cleaning chunks of deer off my car before I returned it.

For me the best news of all is that on December 10th, my daughter Janelle gave birth to our first grandchild, William Paul. He was 8 lbs 15 oz and 22 inches. Both Will and Janelle are doing great. Debbie and I flew up to Chicago the next day. Those of you who are grandparents will clearly understand how proud Debbie and I are. I think being two weeks old is a little early for IRT (even the 1PL model) so I'm just starting him out on simple CTT books. I must confess though, he does cry a lot when I mention standard setting (go figure!)



Well, that's it for now. I wish you all a happy and safe Holiday Season!

Jerry

A NOTE FROM THE EDITOR

Thanos Patelis, The College Board

The active NCME Newsletter Advisory Board continued to offer some wonderful ideas for our newsletter. I'm sorry to say, but the editorial staff (me) couldn't keep up with all the great ideas for this issue. But, we do have a couple here. First, we have an informative presidential column by our president, Terry Ackerman, and we congratulate him on his 20th marathon, his beautiful grandchild, and thankful that he and Debbie are safe! Second, we are fortunate to have a second piece in the Legal Corner by S.E. Phillips. This time the column took on the issues associated with using test scores to evaluate teachers. S.E. Phillips provides an exceptional summary of the debate and presentation of the implications with commentary from Susan Brookhart and Helen Santiago. S.E. also offered some concluding comments. Next, we have our standing graduate student column by Carol Barry where she reports advice from our current NCME president, Terry Ackerman, and president-elect, Wayne Camara, on considering and navigating a career in educational measurement. This is Carol's last column; graduate student columnists have a one-year commitment. We thank Carol for the excellent articles this past year as they have all received positive feedback from our membership. Thank you, Carol! Next, David Frisbie offered an announcement by National Accessible Reading Assessment Projects indicating the publishing of the *Principles to Improve Reading Assessments for Individuals with Disabilities*. Finally, if you have not done so, please renew your membership and register for our annual meeting in Denver, CO scheduled for April 30 to May 3, 2010 by visiting: <http://www.ncme.org/meeting/index.cfm>. As always, please drop me an email with suggestions. Sincerely and at your service, Thanos.

LEGAL CORNER: USING STUDENT TEST SCORES TO EVALUATE TEACHERS

S.E. Phillips, Consultant

The federal regulations for federal *Race to the Top* state education grants place a high priority on state's plans to use student tests scores to evaluate teachers and principals. However, in response to comments from the national teachers' unions, the U.S. Department of Education (U.S. DOE) slightly revised the final grant-evaluation criteria to require states to include multiple measures, including student test score growth, in their teacher- and principal-evaluation systems. The use of student test scores to evaluate teachers and principals is controversial and has been extensively debated. Detractors argue that it is unfair to judge educator quality based on student test scores because student performance may be significantly affected by factors over which educators have no control and whose distributions may differ substantially across classrooms. Proponents argue that good teachers should be able to demonstrate achievement gains with all their students regardless of demographics, prerequisite skill levels or behavioral impediments. Other researchers argue that improved working conditions are more important than growth-model incentives for boosting teacher effectiveness in low-performing schools. Nevertheless, to remain eligible for the grant competition, California and Wisconsin recently removed legislative data firewalls that had prevented student test score data from being linked to individual teachers.

Judging teachers on the achievement test scores of their students is not a new idea. Contrasting legal cases challenging such practices occurred in Iowa and Missouri in 1973 and 1987, respectively. Although the teachers in both of these cases received unsatisfactory evaluations based on their students' low test scores, the results in the two cases were somewhat different. In the 1973 Iowa case, the federal Appeals Court held for the school system permitting students' test scores to be considered in the teacher's evaluation. However, in the 1987 Missouri case, the federal district court refused to dismiss the teachers' suit outright. Since the Appeals Court did not hear the Missouri case and the district court held only that its case could proceed to trial, the earlier Iowa opinion is apparently still in effect. While these cases considered the appropriateness of using student test scores in specific teacher-evaluation contexts, they did not settle the issue. Nonetheless, they may be instructive in illustrating some of the difficulties with such policies.

Iowa Case (*Scheelhaase v. Woodbury Central Com'ty Sch. Dist.*)

Norma Scheelhaase was a nontenured Iowa teacher with ten years of experience whose contract was not renewed because her students' Iowa Tests of Basic Skills (ITBS) standardized achievement test scores were too low. The academic curricula at her school had been criticized by an accrediting group and the school had been notified by the state that it would be removed from the approved list if the deficiencies were not corrected within one year. Among the remedial actions taken by the Superintendent was the nonrenewal of Scheelhaase's teaching contract. At the time, an Iowa statute provided that all teachers' contracts were renewable annually without tenure.

The Superintendent's decision was upheld at a hearing and Scheelhaase appealed. She alleged a substantive due process violation, describing the school district's decision as arbitrary and capricious and an abuse of administrative discretion. She claimed that the District had misinterpreted her students' test scores, and that her students had actually made normal progress. The District countered that the "use of the ITBS scores as a measure of teacher competence stood as a reasonable and valid exercise of administrative discretion."

The trial court held "that a teacher's professional competence could not be determined solely on the basis of her students' achievements on the [ITBS and Iowa Tests of Educational Development (ITED)], ... that [Scheelhaase] had a 'property interest' in her contract of employment" based on a legitimate expectation of renewal after her many years of service, and due process required fair reasons for termination supported by the evidence. The trial court ordered that Scheelhaase be reinstated and the District appealed. The Appeals Court found that the School Board had the authority to nonrenew without cause under the pertinent Iowa statute, the Board relied on the expert opinion of the Superintendent and good faith decisions by the Board could not be overruled because the court would have judged the evidence differently.

Missouri Case (*St. Louis Teachers Union v. St. Louis Bd. of Educ.*)

In 1987, the local teachers' union filed a class action in federal district court on behalf of tenured St. Louis teachers receiving unsatisfactory evaluations based on their students' standardized achievement test scores. The District had unilaterally adopted a new evaluation system for which a final unsatisfactory rating was given to a teacher with low student California Achievement Test (CAT) scores if the principal could document additional deficiencies.

Three indices devised by the District and applied to CAT reading, language and math scores were used to classify teachers in the District into three categories: *satisfactory* if student performance in at least two areas was positive, *in need of improvement* if student performance was positive in only one area, and *unsatisfactory* if student performance was positive in none of the areas absent mitigating circumstances. The Superintendent reviewed teachers who had received *unsatisfactory* ratings and assigned a final rating of *unsatisfactory* if the teacher's principal was able to document deficiencies other than low test scores, or *satisfactory* otherwise. The teacher class members claimed that their unsatisfactory ratings were unfair and unjustly affected their probationary status, possibility of termination, salary adjustments and reputations.

The teachers argued that the District's teacher-evaluation procedure was inappropriate because the CAT standardized test forms and norms being used were ten years old and because the tests had not been validated for the purpose of teacher evaluation. Moreover, the teachers argued that their unsatisfactory ratings were arbitrary and capricious in violation of their substantive due process rights. In addition, the teachers presented disparate impact evidence indicating that 53% of St. Louis teachers taught in nonintegrated schools but that 68% of teachers receiving final unsatisfactory ratings were employed in such schools. Finally, the teachers argued an equal protection violation because only English language, communications and math teachers were subject to evaluation based on CAT test scores. In response, the District moved to dismiss all of the teachers' claims.

In analyzing the equal protection claim, the court found no deprivation of a fundamental constitutional right or suspect class and, therefore, applied low level, rational basis scrutiny. The court concluded that there was no equal protection violation

because the District acted rationally in using CAT scores to evaluate teachers only in the subjects that they taught and in basing final unsatisfactory ratings on multiple criteria.

The court also held that the District had not violated the teachers' procedural due process rights because the unsatisfactory ratings which the teachers claimed stigmatized them as incompetent would be actionable only if they infringed on a liberty interest. However, the court found that "a statement that is basically one alleging conduct that fails to meet professional standards is a statement which does not impinge upon a liberty interest."

In sum, the court found no equal protection or procedural due process liberty interest violations in the District's use of CAT student test scores in its teacher evaluation procedure. However, the court decided that the teachers' claims of a procedural due process violation based on a property interest and their claims that the District's use of CAT scores for teacher evaluation was arbitrary and irrational in violation of substantive due process could proceed to trial and be decided on the evidence.

In particular, the court stated that if the teachers could establish a "legitimate claim of entitlement to salary advancement" under Missouri or common law, then they would have established a property right which could not be infringed without procedural due process. In addition, the court held that the teachers were entitled to present expert testimony related to the alleged inappropriateness of using CAT scores for teacher evaluation. Apparently, the case settled out of court because there were no subsequent published opinions by the court. In such cases, the parties are often prohibited from disclosing the specific terms of the settlement and sometimes are barred from making any public statements about the case or its settlement.

In contrast to the Iowa and Missouri cases challenging the use of student test scores in evaluating teachers, a 2006 Florida case challenged the failure of a school board to follow a state law requiring that teachers be evaluated *primarily* based on their students' test scores. In that case, the court invalidated a teacher termination that had been based on other subjective criteria.

Florida Case (*Sherrod v. Palm Beach County Sch. Bd.*)

Curtis Sherrod was a career teacher in the Palm Beach County schools who had been repeatedly transferred and placed on probation after complaints about his teaching. The complaints related to inappropriate curricula, lack of discipline, incomplete record keeping and excessive group work. He was eventually terminated, requested a hearing at which the termination was upheld and appealed.

A Florida statute required teachers' performance evaluations to be based primarily on their students' performance on annually administered state standardized tests. However, at the hearing, no evidence regarding the performance of Sherrod's students was presented. The School Board argued that "annual student assessments are not always the best means [of] evaluating the effectiveness or skills of a teacher." Nonetheless, the appeals court ruled:

When the meaning of a statute is plain, as here, our role is to enforce the law as written. ... [T]he term *primary* in the statute unmistakably makes student performance on annual tests the *first* consideration in any teacher evaluation. ... In circumstances such as those presented here, where the factors relied on by [the School Board] for termination are confined to pedagogical method rather than personal conduct unquestionably showing unfitness for teaching [e.g. mental diseases or criminal acts], the statute in question requires the school board to base a decision to terminate *primarily* on student performance on the annual tests. Because that was not done here, we have no alternative but to reverse the final order of [the School Board] discharging [the teacher].

Implications

The Iowa, Missouri and Florida cases suggest the following premises for evaluating the legal defensibility of using student test scores to evaluate teachers:

- The Eighth Circuit Court's ruling in support of the use of student test scores to evaluate an individual teacher in a low-performing school in Iowa might have been different if the teacher had been tenured or working under a collective bargaining agreement. For example, a collective bargaining agreement might specify whether and how student data might be used in determining merit pay for veteran teachers. Alternatively, the agreement might provide that participation in a merit pay system based on student test scores be voluntary.
- The use of multiple criteria for making the final decision is likely to be persuasive. For example, accumulating data for a teacher across classes or across multiple years, or providing a judgmental review of the decision that considers extenuating circumstances and/or corroborating evidence may provide sufficient multiple criteria to satisfy a court.
- When using student test scores to evaluate teachers, it is important to develop fair policies and implement uniform procedures to satisfy procedural and substantive due process requirements. This is particularly important for termination decisions utilizing subjective criteria and when personal conduct clearly indicating unfitness for teaching (e.g., mental diseases, criminal acts) is absent.
- There may be fairness issues if teachers who teach nontested subjects are treated differently than those teaching tested subjects or if student test scores for a content area are applied to a teacher with minimal or no responsibility for that content area. For example, which high school teacher(s) are responsible for the reading test scores of their students? Is it

English/language arts teachers only or also teachers of other subjects (e.g., math, science) for which success is at least in part related to reading skill? Alternatively, End of Course (EOC) tests at the high school level may provide a more direct link between instruction and testing and may be more fairly linked to a specific teacher.

- Expert opinions regarding appropriate use of test data and the quality of specific test instruments may play a role in court decisions involving the use of student test scores in teacher evaluations.
- Disparate impact data may be relevant if it suggests that a protected group has been disproportionately disadvantaged.
- Test security procedures and policies for investigating irregularities may need to be strengthened and more resources devoted to the detection and verification of unexpectedly large classroom score gains if student test scores become high stakes for teachers.
- If all parties have acted in good faith, the court is much less likely to second guess educational decisions.
- The specific wording of a relevant state statute may be decisive. Thus, the removal of prohibitions on linking student scores to individual teachers in California and Wisconsin resulting from the U.S. DOE threat to rule those states ineligible for *Race to the Top* federal grants may have significantly changed the options available to local school boards in those states.
- NCLB provisions regarding *highly qualified teachers* and escalating consequences for low-performing schools may have increased the viability of using student test scores to evaluate teachers irrespective of the actions of the U.S. DOE.
- U.S. DOE encouragement of state links between student and teacher data may increase states' experimentation with different methods for using student test score data in teacher evaluations that further refine defensibility principles.

LEGAL CORNER: COMMENTARY

Susan M. Brookhart, Consultant

I learned a lot reading S.E. Phillips's column, "Using Student Test Scores to Evaluate Teachers." The legal background was fascinating, and her conclusions about the legal implications from these cases seem like sound advice to me. I believe they would make good reference for anyone who evaluates teachers.

Part of the stickiness of the issue stems from the fact that both the detractor and proponent positions Phillips summarized in her first paragraph are true. True, student performance may be affected by factors over which educators have no control and whose distributions vary across classrooms. And also true, good teachers can help all students, regardless of demographics, achieve.

I will use this commentary to think aloud about some measurement-related issues inherent in the conundrum about whether and how to use student test scores to evaluate teachers. It is wonderful to have an opportunity to "think measurement" about this issue. Other perspectives – those of state legislatures, district administrators, teachers' unions, parents – are often expressed in the media, and I will not attempt to summarize those here.

Issue 1: What is the construct? Evaluating teacher quality with student test scores assumes that test scores can help answer the question "What makes a good teacher?" One problem is that the implied answer to the question "What makes a good (or at least acceptable) teacher?" is not simple.

My personal favorite answer is what Sara Lawrence-Lightfoot said in a Bill Moyers PBS interview, when asked "What is good teaching?" She replied:

Good teachers come in all forms and express themselves very differently. Teachers don't always connect successfully with all thirty kids in a classroom. But I think that one thing all good teachers have in common is that they regard themselves as thinkers, as existing in the world of ideas. This is true for a nursery teacher and a professor in the most distinguished university. The currency is ideas, but ideas as conveyed through relationships. (Moyers, 1989, p. 159)

This is a very different conception of good teaching from the one New York Mayor Michael Bloomberg implied in a recent comment about using test score for tenure decisions (Medina, 2009):

Referring to the law, Mr. Bloomberg said that banning the use of student achievement in tenure decisions is "like saying to hospitals, 'You can evaluate heart surgeons on any criteria you want – just not patient survival rates!'"

In this conception of successful teaching, the implied criterion is student achievement. Mayor Bloomberg didn't say achievement was the only criterion, but his analogy certainly expresses that for him it is the most important one.

At some level, these two ideas converge on a truth: conveying ideas to students implies they achieve something (at least the having of ideas, and maybe more than that). But for Lawrence-Lightfoot, the journey is important, too – the student's relationship with the teacher and with ideas. Teaching is about making students learners as well as achievers. The more "bottom line" view expressed by Bloomberg and others emphasizes the outcome over the journey.

Issue 1 would probably make the whole enterprise of teacher evaluation a non-starter if it were not for the administrative, economic, political need to evaluate teachers' work. Disagreement about what constitutes good teaching, arguably the construct implied in teacher evaluation, means measurement is going to be dicey.

Issue 2: What is the criterion? The Iowa, Missouri, and Florida cases Phillips described, while all different, all treat standardized achievement test scores as at least one criterion for good teaching. Known to the readers of this column, but not so obvious to the general public, is that standardized test scores are measures or indicators of student achievement. Even if there were such a thing as a standardized test with perfect reliability and validity for measuring student achievement – it still wouldn't *be* student achievement. What's inside a kid's head is a latent construct, with all its attendant issue of domain sampling.

This point does not make it into the metaphors and analogies used so well by the media and by many well-meaning citizens. In Bloomberg's analogy, patient survival is not an *indicator* of the desired end-state, it *is* the desired end-state for heart surgery. People confuse the measure with the essence in this way all the time. I once heard a fellow use the analogy of cleaning up the pollution in the Chesapeake Bay, which had been so bad the oysters had died. He likened oysters returning to the bay to the good test score outcome for which he was campaigning. But once again, oysters in the bay were not an *indicator* of the desired end-state, they *were* the desired end-state.

When I have made this distinction before, some people have said "Aha!" and others have looked at me as if I'm splitting philosophical hairs. What's the difference, and why does it matter? At the very least, understanding the good-test-score criterion as one imperfect indicator absolutely implies use of multiple measures of student achievement. In fact, I'd be happy if the general public could just get that far.

In a perfect world, I would wish to push thinking about the criterion as only one indicator by looking back at Bloomberg's analogy. We can talk to those heart surgery survivors. What was their experience like for them? No one would question that in selecting or recommending (teacher analog – employing) a physician (teacher), information like this would be helpful in making the decision. In a perfect world, it makes sense to me to talk with the students. I would see that not as another measure of achievement, but as an indicator of another aspect of the teacher's work.

Of course this analogy breaks down for the oysters; I'm sure no one cares what they think. So another lesson from the condemnation-by-analogy rhetoric is to understand the limits of analogies. But in all seriousness, the conclusion from this discussion of Issue 2 is that even if we can get political will and general social agreement on the fact that one of the important things teachers are supposed to do is help students achieve, we need a much broader set of indicators than standardized achievement tests.

Issue 3: What (who) is sampled? In teacher evaluation decisions, arguably the "unit of analysis" of interest is the teacher. Thus test scores aggregated over one group of students for one year is a one data point about the teacher's work for that occasion (school year) for those students. Better to have more students than fewer, of course, to make that performance estimate more reliable. But still one year, one teacher, zero degrees of freedom.

Samples of students change every year, and some years go much better than others. Even with two years of student achievement data, that's still a sample of 2 if the unit is the teacher. And by the time we accumulate enough data to make a statistically reliable interpretation, a teacher's career would be over.

So far I've been talking about more or less random sampling error (school groups are never completely random). There is also systematic sampling error, as expressed in the teacher quip, "If we're evaluated on our students' scores I want the honors classes." Statistical models can "correct" for systematic variation, but they do so by labeling and "controlling for" things like poverty, ethnicity, and student prior achievement. Some of those statistical controls translate to interpretations we don't want to make, not expecting (in the statistical sense) as much growth for some students as for others.

Again, the way around this is to think multiple measures. If at least two years of achievement data and other indicators of teaching all point to a similar conclusion about a teacher's performance, a reliable decision can probably be made. Issue 3 is just meant to remind us of the often-neglected issue of sampling error in all of this.

Issue 4: What are the consequences? I agree with the point of view that the consequences of using a measure should be part of a validity argument. Here I'll just mention two consequences that seem to scream out as issues in the use of test scores to evaluate teachers. One is the long-understood phenomenon that high stakes use of a measure corrupt the measure (Campbell, 1969; Lindquist, 1951). We already worry about the measurement and curricular implications as some teachers "teach to the test" (Koretz & Hamilton, 2006). As test scores become even higher-stakes for teachers it would be non-adaptive and ostrich-like for teachers *not* to teach to the test. What else would you expect? To ride Bloomberg's analogy a little, it would be like asking those physicians NOT to seek every possible avenue toward patient survival.

Another consequence-like issue I worry about is the chopping-block tone of all of this. A wise colleague once said, in another context, “We don’t need to fire him, we just need to help him do some things.” Formative and supportive strategies should be the first response when test scores flag a teacher as potentially not as effective as we would like. Professional development, coaching, even medical assistance – all sorts of things should come before the final tenure decision. Most of the school administrators I talk to would say that they do use formative strategies in their supervision, especially with teachers who need help. But in an evaluative culture, it’s hard for teachers to react formatively. The way schools are managed currently, a teacher can’t afford to have her supervisor see her wanting in any respect for very long.

Issue 4 is a sticky wicket. If we use test scores for teacher evaluation, we corrupt them as a measure. If we use test scores for teacher evaluation but tell teachers not to corrupt them, we are naïve and ignore over half a century of measurement wisdom. If we use test scores for summative teacher evaluation, we undermine their usefulness for formative purposes, where they really might make a difference.

Implications. Two conclusions jump out of these issues. One is the need for continued public conversation about just what exactly we want in good teachers and just what it means for students to achieve. Even as we have rushed forward because of the need to measure and report, we still need to revisit the main construct. True progress on this issue will not be made without that.

The second conclusion is to advocate for multiple measures of the outcomes of good teaching, most notably student achievement but arguably other outcomes as well. And a few measures of the process of good teaching wouldn’t hurt, either. There are good legal reasons for this (see Phillips’s column). There are also reasons that have to do with validity (all the issues above, especially 1, 2, and 4) and reliability (sampling error, issue 3).

It is worth pointing out that when one thinks about using test scores to evaluate teachers from the perspective of measurement theory, the courses of action that seem best may be different from those of legislators and community members, who view the problem from other perspectives. Political and social compromise may not lead to measurement-optimal solutions, and we may have to live with that.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409-429.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 531-578). Westport, CT: Praeger.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119-158). Washington, DC: American Council on Education.

Medina, J. (2009, November 26). Mayor says student scores will factor into teacher tenure. *New York Times*. Retrieved 1/25/09 from www.nytimes.com

Moyers, B. (1989). Sara Lawrence-Lightfoot: educator. In B. Moyers (Ed.), *A world of ideas* (pp. 156-166). New York, NY: Public Affairs Television.

LEGAL CORNER: COMMENTARY

Helen C. Santiago, *The College Board*¹

Having recently returned from facilitating a panel discussion on the hill in Washington with 5 teachers who participated in the College Board’s/Phi Kappa Delta International’s collaborative publication entitled *Teachers Voices*, and after having read Mr. Phillip’s article I write in response to his concluding implications.

I could not agree more nor advocate more strongly with the idea that there must be multiple criteria aside from standardized test scores for consideration as a part of any teacher’s evaluation process. I base this on my own personal experience, having evaluated teachers and primarily because of the very nature of schools policies and practices that impact the work of teachers everywhere.

¹ Helen Satiago is vice president of College Board Schools. Helen is retired from the position of senior instructional manager for teaching and learning in the New York City Department of Education. Helen's prior experiences include positions as community superintendent in Community School District 1 in New York City and deputy superintendent in CSDs 9 and 3; director of professional development; elementary school principal; and bilingual teacher not only in New York City but in Biloxi, Mississippi, and Dysart, Arizona, as well. Source: College Board, 2009.

What about Advanced Placement teachers who may only teach AP courses and honors classes while other teachers based on a schools scheduling or availability or student numbers are assigned to classes other than AP or honors? What is effective teaching in relation to that scenario? In schools where the student population can achieve 4's and 5's on AP exams such as Stuyvesant High School in NYC would those teachers be considered "more effective"? What about the teachers that have student class loads of 150 plus across a range of courses (requiring greater preparation), what about Physical Educators, Art Teachers and Music Teachers, content areas for which there are no standardized tests? All this to say that other criteria, such as student growth measured over a school year as determined by test scores, teacher grades and system requirements (local & state) must be taken into consideration. Other school variables including the quality of school administrative leadership, the impressions of parents and communities must be part of the equation.

We are left with the question; how do we determine teacher effectiveness in light of the many variables associated with a place called school that leads to equitable yet rigorous teacher evaluation?

LEGAL CORNER: CONCLUDING COMMENTS

S.E. Phillips, Consultant

Both commentators have highlighted the measurement and practical reasons why teacher evaluation is challenging and multiple criteria are preferred. Brookhart reminds us that the construct of effective teaching must be defined before it can be measured and illustrates why consensus will be difficult. Santiago further explains why the unique characteristics of each teaching assignment makes the task akin to deciding which fruit is better, apples or oranges. Using experience (seniority) as a proxy for effectiveness may satisfy some constituencies but may not retain those teachers most likely to turn around failing schools or satisfy federal criteria for additional funding. Using NCLB student test scores may be attractive because the data is already available but may have other consequences such as the potential for corruption described by Brookhart.

Brookhart also reminds us of the difficulty of using the same test score data for both formative improvement and summative decisionmaking. But in these tough economic times, school districts may have insufficient resources to design and implement dual evaluation systems. Some policymakers may have run out of patience for measurement arguments about appropriate criteria, sampling and consequences and may move forward with evaluation systems based on student test scores in an effort to achieve immediate results. Only time will tell whether sufficient safeguards can be built in to these proposed teacher evaluation systems to satisfy both federal accountability and legal notions of individual fairness. As they say, "the devil is in the details" and the president of the NYC teachers' union has not ruled out filing a lawsuit in response to the announcement of the details of the mayor's plan to use student test scores as a factor in teacher tenure decisions. Similar to the Florida case, existing state laws or collective bargaining contract language may play a major role in the final outcome of such a lawsuit.

CURRENT TRENDS AND FUTURE DIRECTIONS IN EDUCATIONAL MEASUREMENT: PERSPECTIVES OF TWO PRESIDENTS

Carol L. Barry, James Madison University

There is more emphasis on measurement in education now than ever and, because of this, the field of educational measurement is growing and changing at a rapid pace. Given the deluge of information that accompanies these changes, it can be challenging for graduate students to know what to focus on and where to concentrate their energies with regard to their studies. In order to provide guidance from two established professionals in the educational measurement field, we asked current NCME president Terry Ackerman of UNC-Greensboro and NCME president-elect Wayne Camara of The College Board to respond to a set of questions. Their responses are included below.



Given that the field of educational measurement is in a state of constant change and development, what trends do you see emerging in the field today?

Terry Ackerman (TA): I think this is a very exciting time to be in the field of measurement. No Child Left Behind (NCLB) legislation has certainly created mandates for accountability testing in every state that will have impact for at least several years to come. There are also continual hints from various sectors of the K-12 educational community regarding national curriculum standards and a corresponding national testing program (beyond NAEP). These types of initiatives create new opportunities as well as many challenges to produce high quality tests and interpret/use results in valid ways.

Currently, there are many individuals - teachers and administrators - directly involved in assessment at the local and state levels that lack formal training in measurement, statistics, or research design. I think that there will be a growing demand to educate these individuals; that is, to make them *measurement literate*. I also feel that we need to do a much better job of working with teachers to help them effectively and efficiently use test results to improve classroom instruction and to make better prescriptive and diagnostic decisions.

New areas that are beginning to emerge and have an impact on testing include: cognitive diagnostic modeling, assessment engineering, analysis of examinees' response times, more effective ways to display and report test results, and the development of more applied item types.

Wayne Camara (WC): I believe there will be even more migration of assessments to CBT, but new models will be required to address K-12 technology. Increasingly assessments will need to be delivered over extended testing windows which present significant security concerns for high stakes programs and also could require significant costs and resources. In addition, I believe we will abandon efforts to maintain comparability with paper because we cannot innovate with CBT if we are continued to meet paper demands. In addition, I believe that there will be more focus on how summative assessments might include student performance on interim assessments or projects completed during the school year. Many high performing nations have incorporated interim measures into their summative systems and used technology to allow teachers to score performances during the year in a reliable and efficient manner.

What are the skills and knowledge you see as being at the forefront of educational measurement within the next 5-10 years, and why?

TA: I would strongly encourage graduate students who are interested in measurement, whether it be in academia or in the private sector, to be well rounded in psychometrics, statistics, research design and evaluation. Equally important, students need to be capable of communicating technical information about measurement theory, testing results and implications to different audiences (e.g., teachers, administrators, parents, fellow researchers). Students need to be able to *bridge the gap between theory and practice*. This is my theme for the upcoming NCME annual meeting in Denver. Too often students learn theory but do not understand how to apply it in practice.

WC: I always believe that new graduates do need to have substantial technical skills in the traditional core areas such as measurement, statistics, evaluation, assessment and research design. In our field these are the basic skills that are required for entry level positions in applied settings. I would certainly advocate graduate students acquire strong technical skills, as well as experience in using a variety of statistical packages. Managing and manipulating large databases (whether through SAS, Access, or other programs), writing scripts and the ability to do some modest programming will be increasingly important in some areas.

If you were a graduate student today, what would you be learning, what skills would you be developing, and why?

TA: I would strongly encourage graduate students in measurement to develop a solid foundation in psychometrics, statistics, research design, and evaluation. In terms of coursework, I think it is necessary that students find a strong measurement program that will offer excellent breadth of assessment and statistical topics, allow students to collaborate with faculty on research, and identify internships that will help supplement their coursework with hands-on experience.

For example, here in the Department of Educational Research Methodology at UNCG we offer quite a wide range of courses in measurement (classical test theory and generalizability theory, three IRT courses including multidimensional IRT, data presentation, computerized testing, equating, standard setting, survey development and analysis, classroom assessment for teachers, differential item functioning, and cognitive diagnostic modeling) and statistics (multivariate analysis, structural equation modeling, hierarchical linear modeling, and factor analysis). Other courses that students can take include longitudinal data analysis, computer programming, qualitative analysis, and program evaluation.

Equally important to coursework is gaining hands-on experience by working with faculty on research projects as well as participating in internships. There are many internship possibilities offered by testing companies (ACT, ETS, College Board, Measured Progress, etc.) as well as certification and licensure programs (e.g., the American Institute of Certified Public Accountants (AICPA), the National Board of Medical Examiners (NBME), etc.) and publishing companies (e.g., Pearson, Riverside, etc.). After several semesters of coursework students seriously need to investigate doing a summer internship. Such experience will really help inform career decisions about going into academia or seeking a path in the testing industry.

Students should also take an active role presenting at national conferences (e.g., NCME, AERA, ATP, etc.) and international conferences (e.g., Psychometric Society). Based upon the feedback of discussants and fellow researchers, papers presented at

conferences could then be submitted for publication. The top graduates applying for positions in academia or in the private sector have a well established research agenda and are developing a strong publication record.

There are five pieces of advice I like to share with my students:

- 1) Chances are you'll only get one PhD in your life, so make sure you get the absolute most out of your program. Take as many courses and become as involved as you possibly can.
- 2) You want to leave your measurement program and be as competitive as possible for as many jobs as possible (e.g., academia, private testing companies, state accountability offices, government agencies, etc.) because your first, or maybe second choice may not be an option.
- 3) There is a good chance that your first job in the measurement field will not be your last job or may not even be your "ideal" job. That is why it is so important that you choose job(s) that provide you with opportunities (e.g., great colleagues, nurturing environment, etc.) to continue to grow, so that when the "ideal" job does become available you will be very competitive for it.
- 4) The top people in measurement did not become experts only by what they learned in the classroom. Most of their knowledge was acquired on the job, creatively trying to apply what they learned to everyday testing situations where theory breaks down.
- 5) Create bridges, never burn bridges. The measurement community is very well connected and a lot smaller than you think. It is amazing how people you befriend today will end up helping you tomorrow.

WC: In addition to the strong technical skills described in question two I believe that students should get some actual work experience prior to graduation. As students are completing their course work they should work with faculty to investigate potential internships or externships that can give them an opportunity to work in an applied setting with real world problems. Many firms, including the College Board, offer full-year on-site internships, summer internships, or fellowships that allow students to complete projects at their university. In addition, students also need to develop 'soft skills' if they hope to advance in our industry or work in many smaller organizations. At the College Board, we not only seek graduates with strong technical skills, but look for individuals with exceptional interpersonal skills, communications (oral and written), and team work and project management skills. We often need to have our psychometricians and researchers work on cross division teams to solve problems for clients or develop new assessments and products. We do need staff who can work directly with external parties independently and have the diplomatic skills and business acumen to function in meetings with state or district assessment leaders or business/project managers.

NATIONAL ACCESSIBLE READING ASSESSMENT PROJECTS (NARAP) PUBLISHES PRINCIPLES TO IMPROVE READING ASSESSMENTS FOR INDIVIDUALS WITH DISABILITIES

David Frisbie, University of Iowa

The National Accessible Reading Assessment Projects (NARAP) recently published a set of principles to inform state directors of assessment and special education and other groups about the best practices for creating accessible reading assessments for individuals with disabilities. The *Principles* are an outcome of an extensive research effort funded by the Office of Special Education Programs (OSEP) and the National Center for Special Education Research (NCSER) in the Institute of Education Sciences (IES). The *Principles* are based on the research conducted by NARAP, as well as other relevant research, existing standards, and expert consensus. The *Principles* address regular large scale assessments of reading and are focused specifically on reading assessments of content standards based on grade-level achievement standards used for accountability purposes (either school or student accountability). The document contains five principles that provide the frame for accessibility for reading assessments. Each principle is supported by guidelines that suggest specific ways to implement the *Principles*. The five principles listed in the document are:

1. Reading assessments are accessible to all students in the testing population, including students with disabilities.
2. Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.
3. Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.
4. Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.
5. Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

For a full discussion of each of these *Principles*, the guidelines for implementation of the *Principles*, and a summary of the supporting evidence for the *Principles*, please go to, <http://www.narap.info/publications/reports/NARAPprinciples.pdf>

**NCME ANNUAL MEETING
2010 ANNUAL MEETING AND TRAINING SESSIONS
APRIL 29-MAY 3, 2010
DENVER, CO, USA**

If you haven't already, please go to the following link to register: <http://www.ncme.org/meeting/index.cfm>

Program Highlights

Presidential Address: Bridging the Gaps, Terry Ackerman

Career Award Address: Defining and Controlling Errors of Measurement

Moderator: Michael Kolen

Presenter: Michael Kane

Discussant: Robert Brennan

Committee-Sponsored Symposia:

DIVERSITY ISSUES AND TESTING COMMITTEE

Ensuring Equitable Representation of English Language Learners in NAEP: Reactions to the Technical Advisory Panel Report to NAGB on Uniform National Rules for Including and Accommodating ELLs in NAEP

Organizer/Moderator: Charlene Rivera

Participants: Carlos Martinez, Jo O'Brien, Cornelia Orr, Sharif Shakrani, Deb Sigman, Katherine Viator

NATIONAL ASSOCIATION OF TEST DIRECTORS

Validity Issues for Interim Benchmark Assessment Systems

Organizer/Moderator: Jack Monpas-Huber

Participants: Judith Arter, Marty McCall, Lorrie Shepard

Discussants: Pamela Moss, Catherine Taylor

GRADUATE STUDENT ISSUES COMMITTEE

The Influence and Impact of Technology on Educational Measurement

Organizer: Mary Roberts

Moderator: Kimberly Swygert

Participants: Lisa Harris, Richard Luecht, Kathleen Scalise, Joe Willhoft

Invited Symposia:

Assessment of Learning in the Context of Educational Reform: Experiences from America Latina

Organizer/Moderator: Michael C. Rodriguez

Presenters: Michael Fast, Lorena Meckes, Mario Moreno, Fernando Rubio

Update on the Revisions to the Standards for Educational and Psychological Testing

Organizer: Barbara Plake

Moderator: Michael Kolen

Presenters: Laura Hamilton, Joan Herman, Barbara Plake, Denny Way, Laurie Wise

Discussant: Steve Ferrara

Measurement in Higher Education

Organizer/Moderator: Donna Sundre

Presenters: Peter Ewell, Gary Pike, Richard Shavelson, Donna Sundre, Tom Zane

Discussant: Lorrie Shepard

Common Core Standards and Coordinated State Assessment

Organizer/Moderator: Wayne Camara

Presenters: Wes Bruce, Pascal Forgione, Brian Gong, Suzanne Lane, Robert Linn, John Tanner

Are You Being Served? Operational Difficulties in Serving Real and Perceived Needs of State Assessment Clients

Organizer/Moderator: Luz Bay

Presenters: Luz Bay, Daniel Lewis, Diane Henderson-Montero, Paul Nichols

Discussant: Robert Brennan

View from the Top of the Mountain

Organizer/Moderator: Terry Ackerman

Presenters: Robert Brennan, Ron Hambleton, Robert Linn, William Mehrens, Barbara Plake, Lorrie Shepard, Wendy Yen

An Application of Assessment Engineering to Multidimensional Diagnostic Testing in an Educational Setting

Organizer/Moderator: Richard Luecht

Presenters: Mark Gierl, Jacqueline Leighton, Richard Luecht

Discussants: Steve Ferrara, Kristen Huff

Graduate Student Poster Session

This 13th annual poster session of NCME's Graduate Student Issues Committee provides an opportunity for graduate students to share their work and receive feedback from professionals and their peers.

NCME Fitness Run/Walk

Monday, May 3, 2010

5:40 a.m. - 7:30 a.m.

Organizers: Brian French and Jill van den Heuvel

See old friends and meet new ones while running a 5k or walking a 2.5k course on Denver trails. Commemorative t-shirts will be given to all participants (even if you don't wake up in time to make it!).

NEWSLETTER ADVISORY BOARD

CAROL L. BARRY, James Madison University (Grad Student Representative)

SCOTT BISHOP, Data Recognition Corporation

MARY LYN BOURQUE, Mid-Atlantic Psychometric Services

SUSAN M. BROOKHART, Consultant

SUSAN L. DAVIS, Alpine Testing Solutions

ELLEN FORTE, edCount LLC

EDWARD H. HAERTEL, Stanford University

SARA S. HENNINGS, Consultant

JOAN HERMAN, CRESST/UCLA

JOANNA GORIN, Arizona State University

THEL KOCHER, Edina Public Schools, Minnesota

GERALD MELICAN, The College Board

S.E. PHILLIPS, Consultant

CHRISTINA SCHNEIDER, CTB/McGraw-Hill

DONNA L. SUNDRE, James Madison University

XIANG (BO) WANG, The College Board

THANOS PATELIS, Editor, The College Board

Send articles or information for this newsletter to:

Thanos Patelis
The College Board
45 Columbus Avenue
New York, NY 10023

Phone: 212.649.8435
Fax: 212.649.8427
e-mail: tpatelis@collegeboard.org

The *NCME Newsletter* is published quarterly. The *Newsletter* is not copyrighted; readers are invited to copy any articles that have not been previously copyrighted. Credit should be given in accordance with accepted publishing standards.