

PEARSON



# TEST, MEASUREMENT & RESEARCH SERVICES

*Quarterly Newsletter*

VOL 1 | NO 3 | 2008

## INSIDE

Editor's Note  
Announcements  
Editorial Board  
Awards  
Recent Publications  
Recent Research Reports  
Upcoming Conference  
Participation  
Seminars  
TrueScores

## EDITOR'S NOTE

*by David Shin*

### **WELCOME TO THE THIRD ISSUE OF THE PEARSON TEST, MEASUREMENT, & RESEARCH SERVICES (TMRS) NEWSLETTER.**

First, I would like to thank those of you who have sent me your feedback and contributions for this issue of the newsletter. In this issue, we announce lots of great news including Dr. Walter (Denny) Way's appointment to the committee to revise the standards for Educational and Psychological Testing; Dr. Edward Wolfe and Dr. Jennifer Beimers' joining Pearson; Pearson's contribution to the H. Paul and DeCourcy Kelly Endowed Fellowship in Psychometrics; the Pearson Research Grant Recipients; the new version of the Pearson Research Grant Application Guidelines; and the All-Star and Team Player Awards recipients.

In every issue, we routinely include a recent article from Dr. Twing's *TrueScore* blog. In this issue, Dr. Twing's article, "Policy Wonks We Are—Implications for NCME Members," provides insights into why and how measurement professionals should engage in education policy conversations.

In addition, this issue lists the publications, research reports, conference participations, and seminars conducted by the TMRS staff. You will find lists of eight publications, one book chapter, four research reports, 22 NCME papers, 19 AERA papers, 11 papers for other conferences, and one seminar in the newsletter. What a harvest of research activities!

For readers who are new to this newsletter, the TMRS Newsletter is issued quarterly. Past issues can be downloaded at [www.pearsonedmeasurement.com/research/newsletter.htm](http://www.pearsonedmeasurement.com/research/newsletter.htm). Our purpose is to provide news of Pearson TMRS research activities and efforts to the psychometric community. We encourage you to share this newsletter with your colleagues as our email distribution list is incomplete. Readers with questions, comments or who would like a printed version, please contact me by email at [david.shin@pearson.com](mailto:david.shin@pearson.com).

We wish you a great start to an even better year!

## ANNOUNCEMENTS

### *Dr. Walter (Denny) Way Has Been Appointed to the Committee to Revise the Standards for Educational and Psychological Testing*

Please join me in congratulating Dr. Walter (Denny) Way, Senior Vice President in Psychometric & Research Services Pearson, on his appointment to the committee of researchers and experts in education and psychological testing to revise the Standards for Educational and Psychological Testing. He was appointed by the Management Committee that oversees the revision process for AERA, APA, and NCME. The standards—designed to establish criteria for appropriate development, use, and interpretation of tests—have been widely cited by states, federal agencies, private organizations, legislative bodies, and even the U.S. Supreme Court. They are based on the premise that effective testing and assessment requires test developers and users to be knowledgeable about validity, reliability and other measurement issues. The standards are more important than ever given the current demand for educational accountability, the increase of testing in the workplace, and the popularity of computer-based testing. The Standards, first published in 1966, remains strong to this day with nearly one million copies sold.

*Continued on page 2*

**Dr. Chingwei (David) Shin**, Editor  
Assessment & Information, Pearson  
319.339.6480  
[david.shin@pearson.com](mailto:david.shin@pearson.com)

**NEWSLETTER  
ADVISORY BOARD**

**Paul Nichols**

Vice President of Psychometric & Research Services

**Bob Dolan**

Senior Research Scientist

**Edward Wolfe**

Senior Research Scientist

**Kelly Burling**

Research Scientist

**Jason Meyers**

Research Scientist

**Yuehmei (May) Chien**

Associate Research Scientist

**David Shin**

Editor, Senior Research Scientist

*The Pearson Test, Measurement, & Research Services 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.*

Send information for this newsletter to:

**David Shin**

Phone: 319.339.6480

Pearson

david.shin@pearson.com

2510 North Dodge Street

Iowa City, IA 52246

**ANNOUNCEMENTS (CONT.)**

*Continued from page 1*

***A Warm Welcome to Dr. Edward W. Wolfe and Dr. Jennifer Beimers!***

Psychometric and Research Services team is delighted to announce that Dr. Edward (Ed) Wolfe joined our group as a Senior Research Scientist on December 15, 2008. Ed's research focuses on psychometrics, particularly as it relates to applications of item response models to complex and innovative measurement contexts. His work involves extensive use of polytomous, multifaceted, and multidimensional Rasch models. He applies these models in projects involving instrument development and validation studies, particularly in the areas of achievement testing in education, attitudinal measurement in the behavioral sciences, and program evaluation. Ed's recent research projects have focused on topics such as the comparability of computer and paper instruments, development of statistical indicators and cognitive models of rater effects in performance assessments in education, evaluation of functional equivalence of a translated preschool ability test, and the development of guidelines for conducting instrument validation studies. Prior to joining Pearson, Ed was a faculty member at Virginia Tech, at Michigan State University, and at the University of Florida. He also received a post-doctoral fellowship at Educational Testing Service, and he worked at American College Testing prior to completing his Ph.D. at the University of California. We are so pleased to have him join our team!

We also want to welcome Dr. Jennifer Beimers to the Psychometric Service team as an Associate Research Scientist! Jen is working from Iowa City and providing assistance on the American Diploma Project (ADP). Jen is a recent graduate of the University of Iowa. Her research interests include NCLB policy and growth models. Jen will be presenting her dissertation entitled "Consistency of District Annual Yearly Progress (AYP) Determinations across Three Types of NCLB Growth Models" at the annual AERA conference in San Diego.

***Pearson Contributes to Psychometric Fellowship***

Pearson, in conjunction with the College Board, helped the University of Texas establish the "H. Paul and DeCourcy Kelly Endowed Fellowship in Psychometrics." Dr. Kelly was a Professor Emeritus of Educational Psychology and academic administrator at The University of Texas at Austin; his four decades of work to promote fairness in testing practices impacted the nation. This is an important fellowship because it not only provides a memorial tribute to Dr. Kelly's life of service in measurement and of championing the appropriate use of test scores, but also because it will continue to provide support to students, faculty, and research studies in measurement at the University of Texas at Austin.

***Pearson Research Grant Recipients Have Been Selected***

The Pearson Research Service Committee would like to thank all the individuals/groups that submitted an application to the Pearson Research Grant Program. The proposals were all exceptional and a variety of topics were covered. Applications were judged based on the following criteria:

- 1) the value of the proposed research to Pearson;
- 2) the value of the proposed research to the field;
- 3) the adequacy of the research schedule;
- 4) the adequacy of the research plan;
- 5) the adequacy of the dissemination plan; and
- 6) the involvement of other Test, Measurement, and Research Services staff.

We were looking to improve the implementation of the researcher-practitioner model in Pearson Psychometric and Research Services and we would like to congratulate the following individuals/groups that have been selected for funding:

**EVALUATING THE COMPARABILITY BETWEEN ONLINE AND PAPER ASSESSMENTS OF ESSAY WRITING IN THE TEXAS ASSESSMENT OF KNOWLEDGE AND SKILLS**

**Laurie Davis**, Director of Psychometric and Research Services

**Ellen Strain-Seymour**, Senior Test Development Manager

**Jadie Kong**, Associate Research Scientist

**Chow-Hong Lin**, Research Scientist

**APPLICATION OF MATCHED SAMPLES COMPARABILITY ANALYSIS TO ITEM-LEVEL ANALYSIS: AN EVALUATION OF THE METHODOLOGY**

Katie McClarty, Research Scientist

**DISTRACTOR RATIONALE TAXONOMY: A FORMATIVE EVALUATION UTILIZING MULTIPLE-CHOICE DISTRACTORS (MATHEMATICS)**

**VALIDATING MULTIPLE-CHOICE DISTRACTOR ANALYSIS FOR FORMATIVE ASSESSMENT (READING)**

(Serena) Jie Lin, Research Scientist

Kwang-lee Chu, Research Scientist

Ying Meng, Senior Statistic Analyst

**SENSITIVITY OF VERTICAL EQUATING DESIGNS TO ITEM PARAMETER DRIFT UNDER DIFFERENT ESTIMATION STRATEGIES, LINKING SET LENGTHS, AND RASCH MODEL-DATA MISFIT**

Tim O'Neil, Associate Research Scientist

**ASSESSING GROWTH IN ENGLISH LANGUAGE PROFICIENCY: FINDINGS FROM RECENT PSYCHOMETRIC RESEARCH STUDIES**

Jane-Zhen Wang, Research Scientist

**THE EFFECTS OF RESPONSE PROBABILITY CRITERIA ON THE SCALE LOCATION ESTIMATION AND IMPACT DATA IN STANDARD SETTING**

Kay Um, Manger, Psychometric Services Tulsa Office

Denny Way, Senior Vice President of Psychometric Services

Steven Fitzpatrick, Principal Research Scientist

**SCALE CONSTRUCTION AND CONDITIONAL STANDARD ERRORS OF MEASUREMENT**

Tony Thompson, Senior Research Scientist

**APPLICATION OF POWER PRIOR IN EDUCATIONAL MEASUREMENT**

Honglian Zhang, Senior Research Associate

*Pearson Research Grant Proposals for 1st and 2nd Quarter 2009*

Vice President of Psychometric and Research Services, Paul Nichols, announced that the applications for the second round of the Pearson Research Grant Program are due January 26, 2009. The Review Committee will make every effort to announce awards by February 6, 2009. Applications for the third round of the Pearson Research Grant Program will be due April 27, 2009.

Applications must follow the new version of the Pearson Research Grant Application Guidelines. The new version of the Guidelines includes three revisions.

- 1) Applications must include a separate cover page that includes the project title and the names and titles of all staff that will be involved in the project. This allows for blind review of proposals.
- 2) The application must include an email from the applicant(s) manager(s) certifying that the applicant(s) and the manager(s) have worked together to construct a project schedule so that both research activities and operational functions can be accomplished.
- 3) An example is given in the Guidelines illustrating the format for a month-by-month schedule of activities that covers the duration of the project.

## AWARDS

### *All-Star*

**MEICHU FAN, CINDI KREIMAN, LEI WAN, & BRIAN WROBEL**

All-Star award received on September 12, 2008. As a special assignment, Brian, Cindi, Lei, and Meichu participated in organizing and conducting challenging equating work. This was done with their usual display of talent and sophistication.

### *Team Player*

**AARON MCVAY**

Aaron McVay was honored with a Team Player award on November 14, 2008 for outstanding work on the ADP, Maryland, and Ohio projects. Clarifying the complexities of each project, Aaron helped the projects flow through the groups within Pearson, even when faced with a tight schedule.



## RECENT PUBLICATIONS

### Journal Articles

**Arce Ferrer, A.** (in press). Studying the Equivalence of Computer Delivered and Paper Based Administration of the Raven Standard Progressive Matrices Test. *Educational and Psychological Measurement*.

**Converse, Patrick D., Wolfe, Edward W., Huang, Xiaoting, & Oswald, Frederick L.** (2008). Response rates for mixed-mode surveys using mail and email/web. *American Evaluation Journal*, 29, 99-107.

**Greene, S., Wolfe, E.W., & Olson, B.** (2008). Assessing the validity of measures of an instrument designed to measure female employees' perceptions of workplace breastfeeding support, *Breastfeeding Medicine*, 3, 159-163.

**Meyers, J. L., Miller, G. E., & Way, W. D.** (2009). Item Position and Item Difficulty Change in an IRT-Based Common Equating Design. *Applied Measurement in Education*, 22(1) 38-60.

**Myers, Nicholas D., Feltz, Debra L., & Wolfe, Edward W.** (2008). A confirmatory study of rating scale category effectiveness for the coaching efficacy scale. *Research Quarterly for Exercise and Sport*, 79, 300-311.

**Nichols, P. D., Meyers, J. L., & Burling, K.** (in press). A Framework for Evaluating and Planning Assessments Intended to Improve Student Achievement. *Educational Measurement: Issues and Practice*.

**Wolfe, Edward W.** (2008). RBF.sas (Rasch Bootstrap Fit): A SAS macro for estimating critical values for Rasch model fit statistics. *Applied Psychological Measurement*, 32, 585-586.

**Wolfe, Edward W., Converse, Patrick D., & Oswald, Frederick L.** (2008). Item-level non-response rates in an attitudinal survey of teachers delivered via mail and web. *Journal of Computer-Mediated Communication*, 14, 35-66.

### Book Chapter

**Wolfe, Edward W., & Dobria, Lydia** (2008). Applications of the multi-faceted Rasch model. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71-85), Thousand Oaks, CA: Sage.

## RECENT RESEARCH REPORTS

**Meyers, J.L, Davis, L.L, Keng, L. & Nichols, P.D.** (in press). An Evaluation of the Feasibility of Using Automated Essay Scoring in the Texas Assessment Program. *TEA Technical Reports*.

**Thompson, T. D.** (2008). Growth, Precision, and CAT: An Examination of Gain Score Conditional SEM. *Pearson Research Report*. [www.pearsonedmeasurement.com/research/research.htm](http://www.pearsonedmeasurement.com/research/research.htm)

**Wang, Z.** (2008). NYSESLAT Linking Study Report for the New York State Testing Program. *Research Report for New York State Department of Education*.

**Way, D., McClarty, K., Davis L., & Keng, L.** (in press). A Review of Literature on the Comparability of Scores Obtained from Examinees on Computer-based and Paper-based Tests. *TEA Technical Reports*.

## UPCOMING CONFERENCE PARTICIPATION

### NCME

#### Alvaro Arce Ferrer

- » Comparing IPD Detection Approaches in Common Item Nonequivalent Groups Equating Design
- » An Investigation of Traditional and Alternative Approaches to Vertically Scale Modified Angoff Cut Scores

#### Jiao Hong, Shudong Wang, Lei Wan, & Lu Ru

- » Investigation of local item dependence in scenario-based science assessments

#### Kay Um, Denny Way, Steven Fitzpatrick, & Cindi Kreiman

- » The effects of response probability criteria on the scale location estimation and impact data in standard setting

#### Leslie Keng

- » A Comparison of the Performance of Testlet-Based Computer Adaptive Tests and Multistage Tests

**Nilufer Kahraman** (National Board of Medical Examiners)  
& **Tony Thompson**

- » Relating Unidimensional IRT Parameters to a Multidimensional Response Space: A Comparison of Two Alternative Dimensionality Reduction Approaches

**Jason L. Meyers, Ahmet Turhan, & Steven J. Fitzpatrick**

- » Interaction of Calibration Procedure and Ability Estimation Method for Writing Assessments under Conditions of Multidimensionality

**Katie McClarty, Jadie Kong, & Chow Lin**

- » How many students do you really need? The effect of sample size on the matched samples comparability analysis

**Xia Mao & Steven J. Fitzpatrick**

- » An Investigation of the Linking of Mathematics Tests with and without Linguistic Simplification

**David Shin & Yuehmei Chien**

- » The Conditional Randomesque Method for Item Selection in Computerized Adaptive Tests

**David Shin, Tsung-Han Ho, Yuehmei Chien, & Deng Hui**

- » A Comparison of Person-Fit Statistics in Computerized Adaptive Test using Empirical Data

**Ahmet Turhan**

- » The Effects of Anchor Item Position on a Vertical Scale Design

**Ahmet Turhan, Chow-Hong Lin, Kimberly O'Malley, & Michael Kolen**

- » Vertical Scaling for Paper and Online Assessments

**Tony Thompson**

- » Scale Construction and Conditional Standard Errors of Measurement

**Ye Tong & Mike Kolen**

- » A further look into maintenance of vertical scales, paper session
- » Vertical scaling methodologies, applications and research, training session

**Changjiang Wang**

- » Investigating the Effects of Speededness on Test Dimensionality

**Hua Wei**

- » The effect of test speededness on item and ability parameter estimates in multidimensional IRT models

**Lei Wan & George Henly**

- » Measurement properties of innovative item formats in a computer-based science test

**Ming Xu, Jane-Zhen Wang, & Sz-Shyan Wu**

- » A Predictive Validity Study of an English Language Proficiency Test

**Qing Yi**

- » The Impact of Ability Distribution Differences between Beneficiaries and Non-Beneficiaries on Test Security Control in CAT

## AERA

**Agnes Stephenson & Tian Song**

- » Using HLM to Investigate Longitudinal Growth of Students' English Language Proficiency

**Alvaro Arce Ferrer**

- » Linking Strategies and Item Screening Approaches: A Study with Augmented Nationally Standardized Tests Informing NCLB
- » Applying Rasch Modeling and Generalizability Theory to Study Modified Angoff Cut Scores for Reporting with Vertical Scales

**Bob Dolan & Paul Nichols**

- » Technical Quality of Formative Assessments with Online Instructional Tools

**C. Allen Lau, Liru Zhang** (Delaware DoE), & **Jenny Jiang**

- » Using Pass/Fail Pattern to Predict Students' Success for Standards: A Longitudinal Study with Large-Scale Assessment Data

*Continued on page 6*

## SEMINARS

### *Consequences as Validity Evidence: Differing Viewpoints*

*Organizer: Jason Meyers*

*Presenters: William A. Mehrens  
and Paul D. Nichols*

*Date: November 5, 2008*

*Location: Austin*

The Consequences as Validity Evidence seminar is one of the Pearson Seminar Series presented by Test Measurement and Research Services. This two hour session provided opposing viewpoints on using consequences as validity evidence and was led by Bill Mehrens, a national testing expert, and Paul Nichols, Vice President of Psychometric and Research Services. The seminar was held on Wednesday, November 5 in the AOCTR Congress Room.

## CONFERENCE PARTICIPATION (CONT.)

*Continued from page 5*

### David Shin & Yuehmei Chien

- » Conditional Randomesque Method for Item Exposure Control in CAT
- » Using Bayesian Sequential Analyses in Evaluating the Prior Effect for Two Subscale Score Estimation Methods

### Hyun Jung Sung

- » Developing a Short Form of the Enright Forgiveness Inventory Using Item Response Theory

### Jane-Zhen Wang, Husein Taherbhai, Ming Xu, & Sz-Shyan Wu

- » Modeling Growth in English Language Proficiency with Longitudinal Data Using the Latent Growth Curve Mode

### Jennifer Beimers

- » Consistency of District Annual Yearly Progress (AYP) Determinations Across Three Types of NCLB Growth Models

### Jessica Yue, Elizabeth G. Creamer & Edward W. Wolfe

- » Measurement of self-authorship: A validity study using multidimensional random coefficients multinomial logit model

### Leigh Harrell & Edward W. Wolfe

- » A comparison of three global fit indices as indicators of multidimensionality in multidimensional Rasch analyses

### Michael McGill & Edward W. Wolfe

- » Assessing unidimensionality in item response data via principal component analysis of residuals from the Rasch model Validation of measures of the quality of the mentoring experiences of new teachers.

### Paul Nichols

- » The Psychology of Writing Items: Improving Figural Response Item Writing
- » Leveraging Evidence-centered Design (ECD) within Scenario-based Statewide Science Assessment

### Tasha Beretvas (University of Texas at Austin) & Jason L Meyers

- » Modeling Rater Severity Using Multiple Membership Cross-classified Random Effects Models

### Tony Thompson & Denny Way

- » Using CAT to Achieve Comparability with a Paper Test

### Tsung-Hsun Tsai & David Shin

- » Generalizability Analyses of a Case-dependent Section in a Large-scale Licensing Examination

### Yuehmei Chien & David Shin

- » The Weighted Penalty Model and Conditional Randomesque Method for Item Selection in Computerized Adaptive Tests

## ATP

### Jason L. Meyers, Leslie Keng , Laurie L. Davis, & Paul D. Nichols

- » Operational Considerations for Implementing Automated Essay Scoring in K-12 Testing

### Jon S. Twing, Dave Bartram, Roy Swift, Wayne Camara, & John Framer (Moderator)

- » Development and Use of Testing Standards and Best Practices (ATP session)

### Jon S. Twing (Moderator), Lisa Ehrlich, Wes Bruce, Dirk Mattson, & Wayne

- » ATP Best Practice Working Group: Update on Operational Best Practices (ATP session)

## Southwest Educational Research Association Conference

### Raymond S. Brown

- » Examining the Effect of SES and Ethnicity on ELP Scores using HLM



## 35th International MEXTESOL Convention— New Ways for New Needs in ELT

**Luis Perea**

- » Teacher evaluation of item formats for an ESL assessment
- » Corpus linguistics use in developing better language proficiency assessments

## The Texas Assessment Conference

**Gloria Zyskowski & Kimberly O'Malley**

Going Vertical: Up, Up, and Away — In this session, TEA and Pearson shared plans for the 2008-2009 vertical scale in TAKS grades 3-8 reading and math. Presenters shared information about what a vertical scale is, how it relates to measuring student growth, and how the new scale will impact score reports this spring.

**Gloria Zyskowski & Kimberly O'Malley**

How are Cut Points Set on Texas Tests? — Attendees learned about the process used for setting the performance standards, or cut points, on the TAKS tests. TEA and Pearson staff summarized different standard-setting methods, explained which method was used to set TAKS/TAKS-M/TAKS-Alt standards, and used audience participation to practice part of the process with sample test questions.

## The International Conference on Outcomes Measurement

**Leigh M. Harrell & Edward W. Wolfe**

- » Effect of correlation between dimensions on model recovery using AIC

**Michael T. McGill & Edward W. Wolfe**

- » Assessing unidimensionality in item response data via principal component analysis of residuals from the Rasch model

## TRUESCORES

*Each issue of the Pearson Research Services Quarterly Newsletter will include a recent entry from the TrueScores blog written by Jon Twing. For more information on TrueScores, please visit [www.truescores.com](http://www.truescores.com).*

## Policy Wonks We Are—Implications for NCME Members

*by Jon S. Twing*

This is the “full text” of a contribution I made to the NCME newsletter. I thought you might like to get the full inside story!

Mark Reckase’s call for NCME members to become more involved in educational policy is timely and relevant, while perhaps also a little misleading. For example, some of my colleagues and I have been working with states, local schools, and the USDOE regarding implementing policy decisions for many years. Testifying at legislative hearings, making presentations to Boards of Education, reviewing documents like the Technical Standards, and advising policy makers are all examples of how psychometricians and measurement experts already help formulate and guide policy. Nonetheless, I still hear many members of technical advisory committees (experts in psychometrics and applied measurement) “cop out” when asked to apply their experience, wisdom, and expertise to issues related to education policy, often citing that they are technical experts and the question at hand is “a matter of policy.”

I have commented and I believe that we no longer live in a world where the policy and technical aspects of measurement can remain independent. In fact, some good arguments can be made that when such independence (perhaps bordering on isolation) between policy and good measurement practice exists, poor decisions can result. When researchers generate policy governing the implementation of ideas, they must carefully consider a variety of measurement issues (e.g. validity, student motivation, remediation, retesting, and standard setting) to avoid disconnects between what is arguably good purpose (e.g. the rigorous standards of NCLB) and desired outcomes (e.g. all students meeting standards).

In this brief text, I will entertain the three primary questions asked by Dr. Reckase: (1) Should NCME become more involved in education policy? Why or why not? (2) How should other groups and individuals in the measurement community be involved in education policy? (3) What resources and supports are necessary to engage measurement professionals in education policy conversations? In what ways should NCME be involved in providing these?

I think I have already answered the first question, but let me elaborate. I maintain that we measurement professionals are already involved in policy making. Some of us influence policy directly (as in testifying before legislatures developing new laws governing education). Some of us influence policy in more subtle ways, by researching aspects of current or planned policy we do not like or endorse. We often seek out the venue of conference presentations to voice our opinions regarding what we think is wrong with education and how to fix it, which inevitably means we make a policy recommendation.

*Continued on page 8*

## TRUESCORES (CONT.)

*Continued from page 7*

Not only do I believe that NCME and its members are involved in policy making, but I also believe it is critically important for all researchers and practitioners in the measurement community to seek out opportunities to influence relevant policy. I recall recently being involved in some litigation regarding the fulfillment of education policy and the defensibility of the service provider's methods. After countless hours of preparation, debate, deposition, and essentially legal confrontation, I asked my colleague (also a measurement practitioner) why we bother defending best practice when there are so many agendas, so many different ways to interpret policy, so many points of view regarding the "correct way" to implement a measure. Her response was surprising—she said we do it because it is the "right thing to do" and that if we stop defending the right way to do things, policy makers will make policy that is convenient but not necessarily correct. Her argument was not about defining right from wrong; her argument was that if we were not there instigating debate there would be none, and resulting decisions would most likely be poorly informed.

So, my simple answer to the second question is to get involved. If you don't like NCLB, what did you do to inform the policy debate before it became the law of the land? If you think current ESL, ELL, or bilingual education is insufficient to meet the demands of our ever-increasing population in these areas, what are you doing to help shape policy affecting them? Across the country debate rages regarding the need for "national standards" or state-by-state comparability. Why aren't NCME, AERA, and all other organizations seemingly affected by such issues banding together to drive the national debate? Do we not all claim to be researchers? If so, is not an open debate what we want and need? When was the debate where it was decided that the purpose of a high school diploma was college readiness? When did we agree to switch the rhetoric from getting everyone "proficient" by 2014 to getting everyone "on grade level" by 2014? The input of measurement experts was sorely missing in state legislation regarding these issues. It is still desperately needed.

For the purpose of this presentation, let's assume that all measurement and research practitioners are in agreement that we need to take part in policy discussions directly. What resources, tools, and/or procedures can we use to implement these discussions and how can NCME help? I stipulate that there is a feeling of uneasiness surrounding the engagement of researchers and measurement practitioners in policy debates or decisions.

Perhaps this is an unfounded concern, but there seems to be an air, forgive me, of such debates being below our standards of scientific research. Policy research is very difficult (to generate and to read), so why leave the comforts of a safe "counter-balanced academic research design" to mingle with such "squishy" issues as the efficacy of policy implementation? Perhaps NCME could strive for a division or subgroup on Federal and State Policy that would focus on measurement research as it applies to education policy (policy, law making, and rule implementation) to lend more credibility to such a scientific endeavor. Maybe NCME could work with other groups with

similar interests (like AERA, ATP, CCSSO) and maybe even get a spot in the cabinet of the next Secretary of Education for the purpose of promoting the credibility of measurement research and application for informing policy. Perhaps less ambitious things like including more policy research in measurement publications, sponsoring more policy discussions and national conventions, and encouraging more policy-related coursework in measurement-related Ph.D. programs would be a good place for NCME (and other organizations) to start.

Let me close with a simple example of why this interaction between applied measurement and education policy is so important. Many of you are firm believers in the quality of the NAEP assessments. Some of you have even referred to NAEP as the "gold standard" for assessment. NAEP is arguably the most researched and highest quality assessment system around. Yet, to this day many of my customers (typically the educational policy makers and policy implementers in their states) ask me simple questions: Why is NAEP the standard of comparison for our NCLB assessments? NAEP does not measure our content standards very well; why are our NAEP scores being scrutinized? What research exists demonstrating that NAEP is a good vehicle to judge education policy—both statewide and for NCLB?

Don't get me wrong. My argument here is not against NAEP or the concept of using NAEP as a standard for statewide comparability. My question is why my customers—the very people making educational policy at the state level—were not at the table when such issues were being debated and adopted? Did such a debate even take place? As measurement experts, when our customers come to us for advice or guidance, or with a request for research regarding the implementation of some new policy, I believe it is our obligation to know and understand the implications of such a request from a policy point of view, not just a measurement point of view. Otherwise, we will be acting in isolation and increasing the divide between sound measurement practice and viable educational policy.

