NCME national council on measurement in education

Volume 16, Number 3 NEWSLETTER September 2008

FROM THE PRESIDENT: HOW EARLY EVENTS INFLUENCED HOW I THINK

Mark D. Reckase, Michigan State University

When I think back over my professional training and development, a number of minor events keep surfacing in my memory as influencing my future directions. Of course, there are many of these, but to keep this article reasonably short, I will only mention three. I am going to share these with you with the hope that they will stimulate you to reflect on how you came to be who you are and that you will then share those reflections with your students and colleagues.

Challenge a Test Item

As is the case for many students taking introductory psychology courses, I was required to serve as a participant in a research study. My memory is fuzzy about the details of the study, but I do recall that part of it was taking a short, paper-and-pencil intelligence test. I believe it was the *Otis Quick-scoring Test of Intelligence*, but I don't have a high level of confidence in that belief. Whatever test it was, it contained a test item something like:

The statement "rocks can think" is

(a) absurd.
(b) unlikely.
(c)
(d)

There were more than two answer choices, but I only remember two of them. The keyed correct answer was "absurd", but I picked "unlikely" and I was upset when it was scored as incorrect.

It is really not important about whether "unlikely" can be defended as a correct answer. The event awakened in me the idea that test items are not necessarily perfect and the way they are written is critical to the usefulness of the test results. To this day, I am concerned that not enough attention is paid to the quality of test items. Test items are generally accepted without critical evaluation. I am at a meeting right now where a change of a few words in a test question was stated to be of no consequence. I know from personal experience that changing one word can be of serious consequence, resulting in a test question with no correct answer or more than one.

Computer Programming Error

When I was in my fourth year of undergraduate study, I was hired as a computer programmer for a faculty member. This was back in the distant past when elaborate statistical analysis programs where not common. I was asked to write a computer program that would compute the *t*-test statistics for dependent data. That is, each person had a pre-measure and a post-measure and the goal was to determine if there was a change in mean performance. I had the luxury at the time of having a person to key-punch the program into cards that would be read into the computer to analyze the data. The program statements were coded as holes in cards that were read by a card reading machine. I am sure many of you have no conception of this type of programming and data processing. It was from the dark ages of computing.

Well, when the program was run, all of the results came out exactly as predicted by the faculty member and he was very pleased with my work and the result. But, he noticed that two mean values were exactly the same and that seemed unusual so

he asked me to check the results. At first, when going through the program line by line, everything looked correct. But later, while the faculty member was at a conference, I looked at the program again and discovered that the character "I" was punched in a program card as a "1". The "I" was the index of a loop that was accumulating data for a sum. By punching it as a "1", the same data entry was used over and over again. This was a difficult error to catch. When the error was fixed, all of significant results disappeared. I was embarrassed and devastated. I had to communicate that all effects were "washed out."

This lesson has stayed with me and now whenever I see anything that looks the least bit unusual in the report of analyses, I always ask that the initial data entry and programming be checked. It is very easy to miss little things that have big influences on the reported results. The measurement literature is full of examples from equating problems to scoring errors. We as a field need to guard against trying to do things too quickly without enough checks on the accuracy of the work.

Semantic Differential

The first two example events are negative examples that have led me to be cautious in what I do. A more positive event occurred when I took a course on attitude change. This was in social psychology course that was not official a measurement course, but a large component of the course had to do with the measurement of attitudes. One of the measurement methods discussed in the course was the semantic differential. This is a measurement instrument that has a descriptive statement followed by a series of bipolar scales such as the following:

Psychometricians

Hot _____:_____:_____:_____:_____:_____:_____ Cold

Sharp _____:_____:_____:_____:_____:_____:_____ Blunt

Fit _____:_____:_____:_____:_____:_____:_____ Flabby

A mark is made in the spaces on the scale to indicate how similar the object is judged to be to the end words (I think of psychometricians as hot, sharp, and fit) and scores are determined based on a factor analysis of the scales. Results are presented by locating the target statement in a semantic space.

I was intrigued by this scaling method and immediately bought the book *Measurement of Meaning* by Charles Osgood. I still have the book on my shelf. This was my first exposure to methodologies like factor analysis. Over the long term, this has influenced me to study the measurement of achievement from a multidimensional perspective and to work on extending item response theory to the multidimensional case.

I am sure that many of you have similar events in your life. I hope you will share them with your students and colleagues so they can get a sense of the nuances in the development of ideas.

August Board Meeting

The NCME Board met for a day and a half to deal with the details of the functioning of the organization. A few highlights are provided here, but watch for the minutes of the meeting. They are posted on the NCME web site.

- NCME continues to have the Rees Group as the company that provides management support to the organization. Since the beginning of NCME's collaboration with this organization, Bruce Wheeler has been the person at the Rees Group who has worked directly with the Board. He has been our Executive Director. I am sure that many of you have met Bruce at the Annual Meeting. At this Board Meeting, we met Bruce's replacement. Plumer Lovelace is the new Executive Director of NCME. I am sure Plumer will do a great job for us over the coming years.
- The Board has accepted a plan from the Standards Management Committee for the process for the revision of the "Standards for Educational and Psychological Tests." This plan has also been accepted by AERA and APA. That means that the Committee can start recruiting individuals to work on the revision of specific sections of the standards. Please be cooperative if you are contacted to work on the revision to the "Standards".
- A new Web-site Management Committee has been formed and is being organized as you read this. It will be charged with determining the content of the NCME web site and how it should be presented.
- NCME is on a financially sound footing, and the Board is considering how to best use the organizations funds.

MORGAN STATE UNIVERSITY'S GRADUATE PROGRAM IN PSYCHOMETRICS

By, Pamela Scott-Johnson[1]: Morgan State University and Kurt F. Geisinger[2]: University of Nebraska-Lincoln

Morgan State University recognized the need for increasing the number of well-trained professionals in the field of testing and therefore is in the process of establishing a graduate program in Psychometrics, one that is beginning in the fall of 2008. Morgan State University's graduate program in Psychometrics is the first program of its kind at an HBCU. For more than 130 years, Morgan State University has been an important part of the higher education system in the Baltimore area, the State of Maryland, and the nation. Designated as Maryland's *Public Urban University*, Morgan State's mission is to serve a multi-ethnic and multi-racial student body and to help ensure that the benefits of higher education are enjoyed by a broad segment of the population by offering a comprehensive program of studies at the undergraduate level and degrees in selected fields at the master's and doctoral level. Morgan State is also one of the nation's Historically Black Colleges and Universities and as such awards more bachelor's degrees to African-American students than any campus in Maryland. The program will be housed in the university's Department of Psychology. Students can earn a Masters of Science (M.S.) degree and/or a PhD. degree. This program is the 13th doctoral program and the 26th Master's program to be offered at Morgan State.

This program is being started for several reasons. First, educational and psychological testing is playing a larger and more important role in our lives than ever before, due in part to the testing of students required as part of the Federal No Child Left Behind law. There has also been an increased use of many other high stakes tests, including entry tests into college and graduate school, licensing and certification tests, and various psychological tests to assess ability, personality, and other characteristics.

Because of the increased use of and importance of testing, there is a heightened need for top-notch professionals in every aspect of testing – from development to administration to scoring to assuring proper test use. Yet, according to a May 2006 NY Times article, psychometrics-related doctoral programs are producing at most 50 graduates a year, not nearly enough to satisfy the demand for these professionals in school districts, colleges/universities, testing companies, government agencies, and corporations. As of 2005 fewer than 100 colleges and universities had a graduate program in a measurement/psychometrics-related area. More importantly, especially in this instance, there are relative few African Americans in undergraduate and graduate programs (less than 16% of all students enrolled in postsecondary education according to the U.S. Census Bureau in 2003), as is also true of most other minorities. At most a handful of African American and other minority students are pursuing doctoral programs in testing. This small number of minority students (and even fewer graduates) makes it difficult for individuals from minority groups to compete on an equal academic-credential footing for leadership roles.

The Program's Goal

The goal of the Psychometrics program is to develop scholars who possess sophisticated statistical and analytical capabilities and the quantitative and methodological skills (e.g., measurement theory, statistical analysis, research design, evaluation, and qualitative tools) needed to design, develop, interpret and use valid, reliable and fair measurements and assessments of what and how individuals learn. Graduates from the program will have both these as well as specialized analytical skills and cultural competence to provide effective and innovative interventions that address issues within the field itself as well as inform policies that influence minority or special populations (e.g. African Americans) and those within urban environments.

The Uniqueness of the Program

This program is unique for two reasons. The first, of course, is that it is the only such program at an HBCU. A second unique feature of the program is the University's collaboration with Educational Testing Service (ETS). The collaboration between Morgan State and ETS assures the development of a rigorous M.S. and PhD program in Psychometrics that integrates academic rigor with the acknowledged need for the testing industry to increase the number of minorities engaged in psychometrics.

The program's advisory board has representation from professional organizations, which provide insight regarding competitiveness and consistency of practices and principles within the field. In particular, participation in the program is consistent with NMCE's own strategy goal of increasing membership and interactions in the field, with emphasis on minorities and under-represented groups.

Contact Dr. Pamela E. Scott-Johnson at 443-885-3290 or pamela.scottjohnson@morgan.edu for information about the program. You may also access information through the University's website - http://www.morgan.edu/academics/Grad-Studies/gs/doc/psychometrics.html.

[1] Dr. Scott-Johnson is the chair of the psychology department at Morgan State University.
[2] Dr. Geisinger is the NCME representative on the Morgan State University Psychometrics Advisory Committee.

ASSESSMENT IN THE 1980S: RUBIK'S CUBE OR HELLBOY?

By, Anthony J. Nitko: Professor Emeritus, University of Pittsburgh

When I was asked to muse on testing and measurement issues in the 1980s and early 1990s, I was a bit taken aback. I wondered if I even remembered what happened in the 80s: Was it Rubik's Cube? (Yes, invented in the 70s but marketed in the USA in the 80s.) Madonna? (Well, she was born in 1958, but did gain prominence in the 80s) Hellboy? (No, he didn't appear at Comic-Con until 1993!) Breakdancing? (No, that was the 70s!) A good bit of my time in the mid-80s was spent teaching assessment in the University of Malawi: The 1980s' Malawi had no Internet, no TV, and a state controlled radio and newspaper -- breakdancing was an innovation that my 16 year old son introduced to his friends in the Zomba in 1984! He even won a prize for breakdancing at a local fair.

I do remember many of the assessment issues in the United States in the 1980s, however. I will muse about them in this brief article. Space doesn't permit a detailed discussion, but I will give some references in case someone wants to check out the details. But first, let's look at what current measurement experts project is important for measurement research for the next ten years.

What Measurement Research is Needed in the Future?

In the June 2008 *NCME Newsletter,* Ron Hambleton reported on the Graduate Student Issues Committee's invited session, *Future Directions for the Field of Educational Measurement* (Magda, 2008). The six speakers discussed the following as important needs for concentrated future research: (1) cognitive psychology as the basis for designing and linking assessments to instruction (Gorin, Lane); (2) defining and developing ways to increase the knowledge and competence of teachers and other educators in educational assessment and statistical methods (Wick); (3) designing and implementing accountability tests that are sensitive to instruction and which will be of value to teachers to improving instruction (Popham); (4) procedures and applications for Bayesian estimation, pictorial representation of data, causal inference from nonexperiments, and analytic ways of interpreting results from data sets that are missing data (Wainer); and (5) developing procedures for describing the reliability, generalizability, and validity of aggregated multilevel test data used for accountability (Zumbo).

Not surprisingly, interest in these same measurement issues was sown in the 1980s and early 1990s. This article summarizes the profession's activities in the 1980s in some of these five areas. (The fourth area listed above is not included in this review, however.)

Validity Reconceptualized

Perhaps the most sweeping work the 80s and early 90s was the development of a comprehensive reconceptualization of validity as a unitary concept, replacing the fragmented concepts of content, construct, and predictive validity. In addition, issues of values and consequences of using test results were introduced as belonging in a test validation argument. Values and consequences were considered innovative and controversial additions to our understanding how to validate assessment results. The seminal works of Messick (1989), Tittle (1989), Moss (1992, 1995), Kane (1982, 1992), and Linn (1994), moved validity theory and applications forward. Most of their work found its way into the revised *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Reliability and Generalizability

Similarly, Feldt and Brennan (1989) synthesized virtually all of classical reliability and gereralizability theory into an integrated and coherent treatise. Their work, still considered the definitive exposition on classical reliability, included reliability of group means as well as of individuals' scores.

Promulgation of Item Response Theories

The various IRT models, used almost ubiquitously now in off-the-shelf standardized tests, customized state assessments, and national and international assessments, built upon, extended, and modified the IRT work of the 1980s. Item response theories were born of the groundbreaking work of 1960s and 1970s and quickly became of age in the early 1980s. Seminars and training programs were held in every quarter to teach the basics, attracting students and researchers from around the world. Major didactic works were published, and speedy computers and computer programs became accessible. Wright and Stone's (1979) *Best Test Design*, Lord's (1980) *Applications of Item Response Theory to Practical Testing Problems*, Wright and Master's (1982) *Rating Scale Analysis*, and Hambleton and Swaminathan's (1985) *Item Response Theory: Principles and Applications*, provided the foundations for thousands of assessment specialists worldwide to learn about IRT models, to

conduct research, and to apply these models to practical tests, including paper-and-pencil, performance, and computer-assisted assessments. Item banking, test item scaling, rating scales, test score equating, and differential item functioning took on new directions using applications of IRT models.

The End of Behaviorism and Rise of Cognitive Psychology

As psychometric theory and measurement applications developed, the influence of behaviorism was waning. Cognitive psychologists developed various models and explanations of what thinking was like, including thinking required to answer test questions. These researchers invaded the behaviorists' black box of the mind to seek ways of understanding and defining various cognitive processes. Early in the 80s, Robert Glaser (1981) laid out a research agenda for cognitive psychology and psychometrics. Susan Embretson (1985) and her colleagues led the way in combining findings from cognitive science and IRT models to show how to design tests that are more meaningful. By the late 1980s, there were numerous calls for using cognitive processes to develop tests to assess persons in ways that tests developed using behavioral analyses could not. Snow and Lowman (1989) summarized much of this research and showed its relationship to aptitude and educational achievement testing. And others were calling for using cognitive paradigms for meaningful student assessment, especially at the classroom level (Nitko, 1989; Lane, 1989, 1991; Shepard, 1989, 1991). Popham continued urging us throughout the 1980s for meaningful assessment that teachers could use to improve instruction.

Accountability and Crises Expands the Demand for and Criticism of Testing

Several authors (e.g., Berliner & Biddle, 1995; Glass, 2008) have written about the manufactured crisis in American education that arose in the 1970s and 1980s, and how results from national and international tests were politicized and misused in efforts to push forward certain educational and political agendas and testing programs, including state accountability programs, and, eventually, the No Child Left Behind Act of 2001. I will not repeat these analyses here. I should mention, however, that this expansion of educational testing in an accountability context spawned numerous standard-setting methods (summarized, for example, by Berk (1986) and Jaeger (1989)) and concerns about identifying differential item functioning (summarized, for example, by Cole and Moss (1989)). Standard-setting methods came under severe criticism in the 1990s and still remain controversial today. Accountability and the cries of crisis also spawned a huge increase in student testing that seems to be peaking only now. The increases in state testing programs, and the resulting demands for customized testing with new items each year, have created the current shortage of technically trained educational measurement specialists.

Lake Wobegon Stimulates Changes in State Testing Programs

As accountability through student assessment grew from the 70s into the 80s, it was discovered that all 50 states were above the national average (Cannell, 1987), a phenomenon that was dubbed the *Lake Wobegon effect*. At first, measurement specialists snickered at this seeming impossibility, but closer examination of states' data revealed it to be true (Phillips and Finn, 1988). Several investigations soon concluded that this effect resulted from pressures on schools and states to be accountable by using standardized test results, from using the same test form and same norms several years in a row, and from narrowing teaching to improve students' results (Phillips, 1990; Linn, Graue, and Sanders, 1990; Shepard, 1990; Burstein, 1990; Smith & Rottenberg, 1991). Eventually, the technical specifications of state testing programs were changed resulting in today's programs using tests tailored to each state's content standards, item banking for fresh samples of items each year, and IRT technology to scale items and equate test scores. This is not to say that all state programs have avoided curriculum narrowing.

Alternative Assessment Woos WYTIWYG

As concerns about the negative impact of accountability primarily through standardized testing grew, there was an increased criticism of using multiple-choice items for testing and an increased call to use assessment modes that were alternative to MC items. The feeling was that in a high stakes environment "what you test is what you get" (WYTIWYG). If the assessments did not represent genuinely desired learning outcomes, students would never learn the desired outcomes. The strong arguments put forth by the Resnicks (1985, 1989), Archibald and Newmann (1988, 1998), and Wiggins (1989a, 1989b, 1991), among others, combined with curriculum reformers' calls for applying constructivist learning theories and more student-centered teaching activities, swayed educators and state assessment authorities to include performance assessments in state programs and in the classroom. In some instances, the performance assessments were erroneously called "authentic". Measurement specialists were put to work to not only write such assessment tasks, but also to find ways to properly field test, validate, determine their reliability, and scale them. Initial efforts and applications of doing these were summarized by Berk (1986). Several states adopted performance assessments and alternative assessments such as portfolios, but eventually their use diminished because of problems with scoring, cheating, costs, and the lack of evidence that state tests substantially improved students learning. Remnants of these assessment formats remain, but there has been a return to the dominance of multiple-choice formats in states' assessment programs, the exception being, for the most part, assessing students with disabilities who cannot be assessed validly with multiple-choice tests.

Reforming Assessment Training Teachers

With changes in curriculum, teaching methods, and assessment there came calls for reforming the way we teach teachers how to assess and evaluate their students. An NCME team worked with the American Federation of Teachers and the National Education Association to prepare the *Standards for Teacher Competence in Educational Assessment of Students*. Another NCME team later developed a complementary set of standards, *Competency Standards in Student Assessment for Educational Administrators*, by working with the American Association of School Administrators, National Association of Elementary School Principals, and National Association of Secondary School Principals. Throughout the 80s, Stiggins conducted research to demonstrate the frequency, importance, and need for sound classroom assessment by teachers (Stiggins and Conklin, 1992). Colleagues soon followed suit and began a movement to reform assessment training of preservice teachers. Shafer (1991), Stiggins (1991), Airasian (1991), and O'Sullivan and Chalnick (1991) presented summaries of the research on teachers' assessment training in the 1980s and made important suggestions for what should be included in preservice teacher preparation courses. Today, more colleges and universities include formal courses in classroom assessment for preservice teachers than was the case in the 1980s. AERA has a special interest group in classroom assessment research that grew out of the 1980s research in this area. Improving classroom assessment continues to be a challenge for teachers. However, we are still trying to determine how teachers can best use assessment results to improve their teaching and students' learning.

Conclusion

The 1980s saw significant developments in most core areas of educational measurement. Technical developments and subsequent applications set the stage for most of the assessment practices we see today, both good and bad. Most of these areas still need improvement and further research as describe by the presenters of the 2008 Graduate Students' Committee's seminar. On the one hand, assessment in the 80s might seem like working with a Rubik's Cube: Making order and organization out of a seemingly intractable mess. On the other hand, assessment in the 1980s had more akin to Hellboy than we realized – it had awesome technical powers that were put to the service primarily of educational authorities and governments. Wikipedia (2008) describes Hellboy as interacting "regularly with humans, primarily law enforcement officials, the military, and various "scholars of the weird," most of whom are not presented [in the story] as overtly reacting to his strange appearance." Hummmm.

References

Airasian, P.W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice*, *10(1)*, 13-16, 26.

American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, & National Council on Measurement in Education (1997). *Competency standards in student assessment for educational administrators.* Downloaded July 17, 2008 from http://www.unl.edu/buros/bimm/html/article4.html.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education. Downloaded on July 17, 2008 from http://www.unl.edu/buros/bimm/html/article3.html.

Archibald, D. & Newmann, F. (1988). *Beyond standardized testing: Assessing authentic achievement in secondary schools*. Washington, DC: National Association of Secondary School Principals.

Berk, R. A. (1986). *Performance assessment: Methods and applications.* Baltimore: Johns Hopkins University Press.

Berliner, D.C., & Biddle. B.J. (1995). *The manufactured crisis: Myth. Fraud and attack on America's public schools*. New York: Addison-Wesley.

Embretson, S. E. (Editor) (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. A*merican Psychologist*, *36*, 923-936.

Glass, G. V (2008). *Fertilizers, pills, and magnetic stripes: The fate of public education in America*. Charlotte, NC: Information Age Publishing.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125-160.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Lane, S. (1989). Implications of cognitive psychology for measurement and testing: Diagnosis of procedural errors. *Educational Measurement: Issues and Practice*, *8(4)*, 17-21.

Lane, S. (1991). Implications of cognitive psychology for measurement and testing: Assessing students' knowledge structures. *Educational Measurement: Issues and Practice*, *10(1)*, 31-33.

Linn, R. L. (1994). Performance assessment: Policy, promises, and technical measurement standards. *Educational Researcher,* 23(4), 4–14.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Magda, T. *(2008) Future directions for the field of educational measurement.* Invited sessions NCME Annual Meeting, San Diego, CA, Match 25, 2008.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14(2),* 5–13.

Newmann, F, Brandt, R., & Wiggins, G. (1998). An exchange of views on "Semantics, Psychometrics, and Assessment Reform: A closer Look at 'Authentic' Assessments". *Educational Researcher, 27(6),* 19-22.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447–474). New York: Macmillan.

O'Sullivan, R. G. & Chalnick, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice*, *10(1)*, 17-19, 23.

Resnick, D. P. and Resnick, L. B. (1985). Standards, curriculum, & performance: A historical and comparative perspective. *Educational Researcher, 14(4)*, 5-20.
Resnick, L.B. (1989). Tests as standards of achievement in schools. A paper presented at the Invitational Conference of the Educational Testing Service, New York, October 28, 1989. [ED335421]
Schafer, W.D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, *10(1)*, 3-6.
Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, *20(6)*, 2-16.
Shepard, L. E. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
Shepard, L.A. (1989). Why we need better assessments. *Educational Leadership*, 46, 4-9.
Smith, M.L. & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, *10(4)*, 7-11.
Snow, R.E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.
Stiggins, R.J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, *10(1)*, 7-12.
Stiggins, R.J. & Conklin, N.F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany, NY: SUNNY Press.
Tittle, C. K. (1989). Validity: Whose construct is it in the teaching and learning context? *Educational Measurement: Issues and Practice,* 8(1), 5–13, 34.
Wikipedia., The Free Encyclopedia (2008). *Helboy*. Downloaded July 18, 2008 from http://en.wikipedia.org/wiki/Hellboy#Story.
Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703-713.
Wiggins, G. (1989b). The futility of trying to teach everything of importance. *Educational Leadership, 47(3),* 44-48, 57-59.
Wiggins, G. (1991). Standards, not standardization: Evoking quality student work. *Educational Leadership, 48(3),* 18-25.
Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis.* Chicago: Mesa Press.
Wright, B.N. & Stone, M.H. (1979). *Best test design*. Chicago: Mesa Press.

PSYCHOMETRICS AT HOME: TELECOMMUTING IN OUR INDUSTRY

Chad W. Buckendahl & Susan L. Davis, Alpine Testing Solutions

The popularity of telecommuting is growing across many industries. Factors including technology capabilities, rising energy costs, commute times, and workforce demands are contributing to the increase in alternative work environments. Within the psychometric community, several companies are considering or have adopted policies that facilitate telecommuting for at least some job roles. In this article we evaluate some of the potential benefits and drawbacks to telecommuting from employee and employer perspectives.

Telecommuting is often viewed as perhaps the highest level of flexibility an employer can offer to an employee because of the level of trust that productivity will be at or above the level anticipated in a traditional work environment. For particular job roles that involve regular, national and/or international client interaction, travel may necessitate a communications infrastructure that includes email, cell phones, and virtual meeting software. Telecommuting is often an extension of this existing infrastructure.

We are regularly asked about the telecommuting experience and how our work environment is different from a traditional setting. In constructing this article, we wanted to combine our perspective with colleagues who work at least part-time as telecommuters. We used the results from an informal survey of these colleagues to inform our analysis of potential benefits and drawbacks to telecommuting in a range of psychometric job roles. To protect the anonymity of respondents in this small, non-random sample, we will not share demographic information. Some common themes emerged. Several of the psychometricians in our sample began telecommuting for personal reasons that required them to live in a specific location that was away from their primary office location. Others began working from home to reduce commute time and benefit from the type of work conditions offered by a home office. The portability of the work and the availability of positions within the psychometric community have also likely facilitated this trend.

Benefits of Telecommuting

From an employee's perspective the most frequently mentioned benefit of telecommuting was flexibility. For some respondents this flexibility meant work location (i.e., living where they wanted). Others defined flexibility in the context of work hours that could be scheduled around employer requirements, client demands, and their preferred schedule. In turn, some telecommuters mentioned that because they could help define their own schedule (e.g., start working earlier, start working later, taking breaks to coincide with family responsibilities) they found themselves being more productive as they saved time that would otherwise be spent preparing for work, commuting, and chatting with colleagues. Successful telecommuters also perceived that they worked in a more efficient manner when they designed a specific home office to reduce distractions and interruptions.

Employers may also benefit from allowing qualified staff members to telecommute. By offering positions that are eligible for telecommuting, employers can broaden the net of their recruitment efforts and attract staff members who may not have previously considered the physical location of the company's offices. Employers may realize increased productivity from these staff members as a result of the greater efficiency of the work environment. Telecommuters in our survey also generally

reported being more motivated to produce as justification of the management decision after being given the flexibility to work from home. For employers that offer full-time telecommuting positions, there are potential cost savings on office space, furniture, office equipment, utilities, and related overhead costs that may be otherwise provided by the employer. There is also the potential to reduce staff turnover if competitors in the market do not offer comparable flexibility in the work environment.

However, there are also a number of drawbacks to the increasing trend in telecommuting within the psychometric community.

Drawbacks to Telecommuting

Each organization has its own culture that is modeled by its leaders. Transferring that culture to new staff members and then maintaining it among existing staff members becomes an increasing challenge in organizations that allow telecommuting. For an employee, the reduction of face-to-face contact with colleagues means that it may be more difficult to appreciate and monitor effort. Some of our respondents reported feeling like their work and workloads were not recognized by colleagues who worked in a traditional office setting. It is also more difficult to establish collegial relationships with others who are in the office if there is not a professional reason to interact with them. Although instant messages, emails, and phone calls with colleagues can serve as effective communication tools, they do not replace the experience you get from sitting across a table from a colleague to collaborate on a project. This, in turn, can sometimes lead to feelings of social isolation as some respondents reported feeling out of the loop with respect to decision-making processes or general activities in the office. As these drawbacks suggest, telecommuting may not be the ideal working arrangement for everyone, but rather most appropriate for particular personality types. In addition to the personal demands, telecommuters also have to consider the practicality of providing sufficient office space, equipment, and "self" technology support that are often required for at-home offices.

Similarly, for employers there are many drawbacks to telecommuting that have contributed to its slow evolution. Without all staff members having at least some physical presence, employers may be challenged to promote a "team" approach to projects where employees share day-to-day accomplishments or frustrations when striving toward a common goal. In addition, having employees working remotely in a home office environment requires adjusting traditional management styles. Balancing the morale of both the in-house and telecommuter staff members becomes an added management challenge. Employers also need to consider the increased risk associated with physical and electronic security when employees and office equipment (typically computers) are housed outside the control of a common physical environment. Although there are additional steps organizations can take to mitigate these risks, these strategies may begin to outweigh the cost-benefit of telecommuting.

Advice for Telecommuters

Clearly telecommuting is not appropriate for all employees or employers. For NCME members who are considering telecommuting or would like to present telecommuting as an option to their employer there are several things to consider.

There appears to be some common personality characteristics of successful telecommuters. Research suggests that these individuals are very organized and have superb time management skills. In addition, they must be able to work independently without the physical presence of colleagues or a supervisor as motivation. It is important to understand and analyze your personal work style prior to considering telecommuting to avoid disappointing yourself or your employer.

For existing and future telecommuters, we conclude this article with three recommendations drawn from our respondents and organizational management research in this emerging area. First, the most common recommendation is to establish sufficient space within your home that can be solely dedicated to your office. This should not be the place where you pay your home bills, have your family computer, or perform any other type of daily function. This should be a space you enjoy working in that also has an office feel to help you stay on task.

Second, telecommuters need to be disciplined to focus for an appropriate amount of time during the day to complete work-related tasks. This means not being distracted by things around your house or other people who might be in your house. However, there is a related caution in that telecommuters need to establish a routine that forces you to stop working at some point. With your work environment in such close proximity to home, it is very easy to "stop by the office" during the evening or never really leave the office at the end of the day. This can be particularly challenging if your co-workers span across multiple time zones expanding the work day. In developing your work schedule, include opportunities to leave the office for coffee, lunch, exercise, or some other type of non-work activity.

Finally, it is important to establish guidelines and expectations with your employer about how this arrangement will work and what is expected of you. Part of these expectations includes building a strong communication infrastructure with your employer and your colleagues. Let your colleagues know how you like to be contacted (e.g., email, cell phone, landline) and then be responsive to that mode to develop the trust that they can count on you.

Technology advancements and the nature of many psychometricians' work will continue to open opportunities for telecommuting. These opportunities are in partial response to market demands for these specialized skills and changing attitudes in the workforce. Telecommuting can be beneficial for employees and employers when developed in a context that recognizes the factors that contribute to its success and mitigates the risks associated with this growing trend.

LARGE-SCALE ASSESSMENTS IN CANADA[3]

By, Don A. Klinger: Queen's University, W. Todd Rogers: University of Alberta, Tess Miller & Christopher DeLuca: Queen's University

Recently, there has been an increase in the number and purposes of large-scale educational assessment programs in Canada. However, due to the provincial/territorial control of education throughout Canada (Note 1), it is not known to what extent the format and purposes of these assessment programs vary. Therefore, the central purpose of the present paper is to document the format and explicit purposes of the current large-scale assessment programs in each of Canada's ten provinces and three territories. Using document analysis of publicly accessible policy documents, examination of Ministry websites, and telephone interviews with Ministry (Department) of Education officials, the explicit purposes for each assessment were identified. At the same time, the general characteristics of the assessments were obtained. The article begins with these general characteristics to provide a context in which to present and discuss the explicit purposes. This order was adopted given the controversy surrounding the purposes of large-scale assessments in the provinces and territories (Rogers & Klinger, 2007).

Current Assessment Programs in Canada

Table 1 provides a summary of the grades and subject areas each province/territory assesses. For example, students in Alberta are assessed in Language/Literacy and Mathematics at Grades 3, 6, and 9, in science and social science at Grades 6 and 9, and in specific academic subjects at Grade 12. In contrast, Grade 6 students in Québec are assessed in Language/Literacy and Mathematics at Grade 6 and in specific subjects in Grade 11.

Despite the provincial/territorial control of education, the provincial and territorial assessments are more similar than different. First, provincial assessments typically occur every three years starting in the primary division and continuing until early secondary school. The most common starting point for such an assessment program is either Grade 3 or 4, although New Brunswick first tests Anglophone students in Grade 2 and Francophone students in Kindergarten. The variations is starting times and the intervals between assessments are primarily due to the variance in academic divisions across the provinces/territories (e.g., Grade 6 marks the end of elementary school in Alberta, while Grade 7 marks the end of elementary school in British Columbia) rather than being based on the cognitive developmental stages of students.

Second, Language/Literacy (reading and/or writing) and Mathematics are most commonly assessed prior to Grade 11, with specific school exit examinations at Grades 10, 11 and/or 12. Literacy is typically defined by reading (e.g., demonstrate explicit and implicit understanding of reading passage and connecting their understanding of the passage with their own personal knowledge and experience) and writing skills (e.g., topic development, use of conventions). The range of mathematics topics and skills is somewhat more varied. For example, the topics typically include one or more of number sense and numeration; measurement; geometry and spatial sense; patterning and algebra; trigonometry; and data management, probability, and statistics; while the skills typically include computation, problem solving, and communication. The specific exams at the upper grade levels are in the core academic areas.

Third, teachers are typically involved in item development. They help develop and review items along with staff from the assessment division or agency in the province/territory. The assessment items are based on curricular expectations stated in provincial/territorial curriculum documents for each subject area, although literacy examinations may be considered to be cross-curricular. Further, assessments having a three-year cycle are based on the curriculum of the current and the preceding grades not having large-scale assessments. In contrast, assessments used to determine students' course marks only use the curriculum of the course they represent. Assessments typically have both multiple-choice and constructed-response items with few assessments containing numeric response items. The literacy and mathematics assessments in Québec are notable exceptions as they only contain constructed response items.

---

[3] The authors wish to thank the editor of *the Canadian Journal of Educational Administration and Policy*, for providing permission to reprint and adapt portions of the article entitled *The evolving culture of Large-scale assessments in Canadian Education* for use in this newsletter. This study has been supported by a standard research grant from the Social Sciences and Humanities Research Council (SSHRC).

Table 1
Provincial/territorial Assessments by Subject and Grade

	Subjects per grade				
Province/territory	**Language/ Literacy**	**Math/ Numeracy**	**Sciences**	**Social Science**	**Academic 11 /12**
Alberta	3, 6, 9	3, 6, 9	6, 9	6, 9	X*
British Columbia	4, 7, 10,12	4, 7, 10	10	11	X*
Manitoba	3, 8, 12	3, 7, 12			
New Brunswick-A	2, 4, 7, 9**	5, 8			
New Brunswick-F	K, 2, 5, 8, 10, 11	3,5, 8, 11	5, 8, 10	11	
Newfoundland & Labrador	3, 6, 9	3, 9			X*
Northwest Territories	3, 6, 9	3, 6, 9			X*
Nova Scotia	3, 6, 9	3			X
Nunavut		3			X
Ontario	3, 6, 10**	3, 6, 9			
Prince Edward Island	3, 6, 9*	3, 6, 9*			
Québec	6, 10*, 11*	6, 10*, 11*	10, 11*	10, 11*	X*
Saskatchewan	4, 7, 10 (reading) 5, 8, 11 (writing)	5, 8, 11			X*
Yukon	3, 6, 9, 10	3, 6, 9*, 10*	10		X*

* Assessment results are combined with teacher awarded mark to obtain a blended mark. The weights of the provincial assessments are: Alberta: 50%; BC: Gr. 10 & 11: 20%, Gr. 12: 40%; Manitoba: 30%; New Brunswick: Francophone: 40%; Newfoundland & Labrador: 50%; Northwest Territories: 50% [Don: is this correct?]; Nova Scotia: 30%; Prince Edward Island: Gr. 9: 10%; Québec: 50%; Saskatchewan: 40%; Yukon: Gr. 9: 10%, Grs. 10 and 11: 20%, Gr. 12: 40%).

** New Brunswick: High school graduation; Ontario: High school graduation.

Fourth, assessments are administered under standardized conditions to the students in the schools. All students are assessed except for the Grade 12 examinations in Saskatchewan (Note 2). Administrations typically occur in April or May for assessments having no direct consequences to students and at the end of the course for high school exit examinations and assessments that contribute to a final course grade. In the case of exit examinations, multiple administrations are held in those provinces with different school schedules (full year, semester, quarter). Three exceptions are the Grade 3 and 4 assessments in Manitoba and the elementary literacy assessments in Nova Scotia (administered in September and October) and the Ontario Secondary School Literacy Test (OSSLT), a graduation requirement that is administered in March. Generally, classroom teachers proctor the assessments having no direct consequences for students. For assessments with direct consequences, teachers other than those who teach the students in the subject area being assessed or school administrators proctor the assessments.

Fifth, teachers are involved in marking the constructed-response items. Not surprisingly, assessments having a direct impact on students' grades or graduation requirements are marked in central locations, and, usually, by more than one trained teacher marker. Assessments not directly impacting school grades are either marked centrally, in school district marking centers, or by classroom teachers using marking guidelines provided by the province's Ministry (Department) of Education. Multiple-choice and numeric response items are machined scored.

Sixth, there is a trend in the extent to which large-scale assessment results are included in course grades. Generally, there is no requirement to include the provincial/territory assessment results as part of an elementary student's grade. However, beginning with Grade 9, assessments may account for a portion (10 to 20%) of the final grade. This amount increases with increasing grade level, from as much as 20% (Grades 10 and 11, British Columbia and Yukon) to 50% (Grade 12, Alberta, Newfoundland

& Labrador and Québec). In three cases, assessments are used as a graduation requirement (Grade 9 New Brunswick-Anglophone; Grade 10 Ontario; and Grade 11 New Brunswick-Francophone).

Seventh, assessment reports are most commonly produced at the student, school, school board and provincial level, and less commonly at the classroom level. When centrally scored, reports tend to be released in the summer for exit examinations and in the fall of the following school year for other assessments. Student level results generally are reported prior to school and school board results. School and school board reports are distributed in a manner that protects the anonymity of individual students and teachers. Thus while the general public and policy makers have access to school, district, and provincial results, only the students, parents, teachers and school administrators have access to individual student results. When reports are published, caveats are provided noting the limitations of the assessment (e.g., restricted to learning outcomes that can be measured using a paper-and-pencil assessment) and errors of measurement. Technical reports detailing the assessment procedure as it was implemented are produced by most provinces/territories.

Lastly, for the most part, the assessment procedures briefly summarized above have remained essentially constant since the inception of each program. Local changes have occurred, including adding additional grades or assessment areas, making the assessment more efficient (e.g., Ontario), reducing the weight of exit examinations (e.g., from 50% to 40%) and, most recently, making examinations optional instead of compulsory (Grade 12 British Columbia; Note 3).

Purposes of Large-Scale Assessments in Canada

Based on the earlier work of McEwan (1995), Nagy (2000), and Taylor and Tubianosa (2001), the purposes/uses for the provincial/territorial assessments were classified as accountability, gatekeeping, and instructional diagnosis. Assessments with stated purposes that used the term "accountability" or associated terms (e.g., monitoring with respect to a provincial standard or across time, ongoing formal reviews) and for which the results were publicly reported at the school and board levels were placed within the accountability category. Assessments in which the results were used to support student promotion and high school graduation, certified competence in selected subject areas, or contributed to a final course mark were classified as gatekeeping. Assessment purposes/uses classified as instructional diagnosis included purposes that called for identifying student strengths and weaknesses. To be classified under instructional diagnosis, teachers and principals must receive the assessment results in a timely matter so that they could use these results to make instructional changes in either the current or following school year.

The assessment programs were not differentiated in terms of low- and high-stakes. This decision was based on two factors. First, high-stakes assessments are defined in terms of consequences and importance, for example, marks, grade promotion, graduation, or admission (Santrock, Woloshyn, Gallagher, Di Petta, & Marini, 2004). Second, although low-stakes assessments do not directly impact individual students' educational decisions, it is not clear students make these distinctions. This is further complicated by provincial policies encouraging or allowing teachers to include provincial assessment results in determining the students' school grades and the public ranking of schools by independent agencies (e.g., Fraser Institute rankings in newspapers).

Turning now to the purposes/uses, each assessment program had more than one purpose/use and that the purposes/uses for an assessment program fell into the three classes. For example, as shown in Table 2, some of the purposes/uses of the Alberta Provincial Achievement Testing Program (APATP) were classified as accountability while the remaining were classified either as gatekeeping (teachers are encouraged to use the results of the APATP and some do) or as instructional diagnosis, whereas the Alberta Diploma Examination Program (ADEP) serves the purposes of accountability and gatekeeping.

All of the provinces/territories explicitly state that at least one of their assessments is for accountability purposes. Accountability is largely accomplished through the public reporting of assessment results to the different educational stakeholders. The results reported are at the school, school board, and provincial levels for the assessment year and, in terms of previous year's assessment results, either in terms of status comparisons or through cohort analyses (employing regression procedures). Teachers and principals are to use the results to review their instructional practices and, based on the performance of their students on the assessment, make needed changes which are to be designed to maintain learning/teaching strengths and address weaknesses. By so doing, student performance in the next year will increase, thereby increasing the level of performance of the schools. The effect of the changes made is then determined by the performance of the students on a comparable form of the assessment in the next year. And so the accountability cycle continues.

Like accountability, gatekeeping purposes/uses were identified in all provinces/territories, most commonly associated with exit examinations contributing to students' final grades. At the extreme, Ontario and New Brunswick (Anglophone) have a graduation requirement attached solely to successful performance on a literacy test. The remaining contributions to final course marks vary between 10% (PEI, BC Grades 10 and 11, Yukon Grades 9, 10, and 11) to 50% (Grade 12 in Alberta, Newfoundland & Labrador, Northwest Territories). In some provinces (e.g., teachers are encouraged but not required to include the assessment results as part of a final course mark). Where available, the Grade 12 assessment results, typically

combined with school-based marks, are commonly used to support admission decisions to tertiary institutions (e.g., colleges, universities, technology institutes).

Table 2

Provincial/Territorial Assessment Programs: Explicitly Stated Purposes

	Purposes/Uses		
Province/territory	**Accountability**	**Gate Keeping**	**Instructional Diagnosis**
Alberta	APATP, ADEP	ADEP	APATP
British Columbia	FSA, BCGPPE	BCGPPE	FSA?
Manitoba	Senior 4 PEs	Senior 4 PEs	PAP
New Brunswick (Anglophone)	Grades 2, 3, 8, and 9 PEs	Grade 9 proficiency examination	Grades 2, 3, 8, and 9
New Brunswick (Francophone)	Évaluation de la petite enfance, PEs	Grade 11 French/ Mathematics	Évaluation de la petite enfance
Newfoundland & Labrador	CRT, PEs	PEs	CRT
Northwest Territories	APATP, ADEP	ADEP	APATP
Nova Scotia	PLANS, NSE	NSE	PLANS
Nunavut	ADEP	ADEP	Grade 3 assessment
Ontario	Grades 3, 6, and 9 assessments, OSSLT	Grade 9 Numeracy, OSSLT	Grades 3, 6, and 9 assessments
Prince Edward Island	Grades 3, 6, and 9 assessments	Grade 9	Grades 3, 6, and 9 assessments
Québec	Cycle 3 Examinations	CEs	Cycle 3 Examinations
Saskatchewan	AFL, DEs	DEs	AFL
Yukon	YAT, BCGPPE,	YAT (Grade 9), BCGPPE, LPI	YAT

Note: APATP-Alberta Provincial Achievement Testing Program; ADEP-Alberta Diploma Examination Program; FSA-Foundation Skills Assessment; BCGPPE-British Columbia Graduate Program Provincial Examinations; PE-Provincial Examination in New Brunswick and Public Examination in Newfoundland & Labrador; CRT- Criterion-referenced Test; PLANS-Program of Learning Assessment for Nova Scotia; NSE-Nova Scotia Examinations; OSSLT-Ontario Secondary School Literacy Test; Cycle 3 corresponds to Grade 6; CE-Certification Examination; AFL-Assessment for Learning; DE-Departmental Examination; YAT-Yukon Achievement Test; and LPI-Language Proficiency Index.

Again, like accountability and gatekeeping, instructional purposes/uses were found in all of the provinces/territories. Alberta, Manitoba, New Brunswick (Francophone), Northwest Territories, Nova Scotia, Ontario, and Saskatchewan have specific statements identifying the use of an assessment for instructional and diagnostic purposes. The other provinces and territories identify similar uses, albeit the language tends to be broader, identifying the use of the assessments to help make plans to support teaching and learning and improve student achievement. As an example, the AFL program's goal in Saskatchewan is to provide data to teachers and education leaders "to provoke debate and inform decision-making in order to improve student learning" (Saskatchewan Ministry of Education, 2007). In Ontario, schools are required to use the assessment results to develop "school improvement" plans and the Educational Quality and Accountability Office explicitly provides assistance for school improvement. Similarly, while not referring specifically to instruction and the improvement of learning, Québec and Prince Edward Island mention that their assessments could be used to help teachers examine the effectiveness of their classroom practices.

To allow the use of assessments for instructional diagnoses, school level assessment results are made available either prior to the beginning of the next school year or, in three cases, during the current school year. Given that the administration of the assessments generally occurs later in the school year, these instructional and diagnostic uses appear to be directed primarily to support future teacher practices rather than supporting the individual students who have completed the assessment. The three exceptions are the Kindergarten assessments in New Brunswick (Francophone), the Grade 3/4 assessment in Manitoba, and the Grades 3 and 6 literacy assessments in Nova Scotia. These three assessments are administered early in the school year and results at the student level are provided in time for teachers to work with individual students and make needed changes in their instructional practice.

While purposes/uses were found for each class for at least one assessment program in each province/territory, as indicated earlier, purposes/uses related to accountability, which in many cases were introduced after the introduction of the assessment program, have become the lightening rod for change and school improvement. This accountability purpose is typically comprised of both internal accountability mandate to be reviewed and acted on within the schools, and an external accountability to report to the general public. Internally, schools and districts are expected to use their assessment results to make data-supported curriculum and instructional decisions. Thus accountability and instructional diagnosis become inextricably interwoven in those provinces where school and school district results are publicly reported. External accountability tends to serve a larger role of educational monitoring, providing evidence that schools are developing foundational skills and knowledge in children, and reflecting the health of the provincial/territorial educational system.

None of the educational jurisdictions in Canada attach any negative consequences for teachers, schools, or districts based on assessment results. While teachers at Grade 12 with consistently poor results have been moved and some elementary teachers will not teach grades with provincial assessments, there are no formal sanctions for teachers or for schools.

Discussion

Large-scale assessments and provincial/territorial assessment programs are an active part of Canadian education. Inferences drawn from these assessments help shape and guide instruction, curriculum, and policy, and inform student-based decisions. Despite the fact that education in Canada is provincial/territory responsibility, the purposes, subject areas and grade levels tested, assessment procedures used, and timing of the assessments are more similar than different.

The *Principles for Fair Student Assessment Practices for Education in Canada* (1993) identifies the need for test developers and users to clearly indicate the intended purposes/uses of assessments (Joint Advisory Committee, 1993). Clearly, in contrast to the assessment procedures employed in the provinces and territories, the purposes/uses of these assessments have been expanded, principally to include differing levels of accountability, which are now, at least implicitly, the preeminent purposes in most provinces/territories.

Explicitly-stated purposes for large-scale assessments are essential and the assessment procedures employed must be sound and relevant to the purposes if valid inferences from assessment results are to be used to determine how to maintain strengths and positively address weaknesses. Further, in light the ongoing changes to the assessment purposes/uses, there is a real need to carefully examine the fit between the assessment procedures presently used and the currently identified purposes/uses for the assessment to ensure each inference and use of the assessment results are sound and valid. The caution remains; assessments constructed for one purpose may not yield sound and credible results for a newly added purpose/use, thereby resulting in educational decisions and reforms that may be misdirected.

Notes

1. The education of children in Canada falls under provincial/territorial jurisdiction. The Ministry or Department of Education within each province and territory is responsible for the development of curriculum and, with one exception, the monitoring of student achievement within its jurisdiction. In the case of Ontario, the monitoring of student achievement is conducted by an arms-length agency (Education Quality and Accountability Office).
2. The Grade 12 Departmental Examinations are administered in core academic subject areas (biology, chemistry physics math, and language arts) to Grade 12 students that are instructed by teachers who have not been accredited by the Ministry of Education. For these students, the examinations contribute 40% to their final course marks.
3. Prior to the 2006/2007 school year, Grade 12 students were required to write the Diploma Examinations in each of the examinable courses. These scores from these examinations were combined with teacher awarded marks, with the examination accounting for 40% of the blended mark for each course. The University of Victoria decided they would only use school awarded marks to make admissions decisions, citing the lateness of the examination scores and the fact that students applied from other jurisdictions that did not have exit examinations like the Diploma examinations. The University of British Columbia followed the University of Victoria, and no longer requires examination scores for admissions decisions. Grade 12 students who want scholarships must still take these examinations: other students may if they wish to.

References

Crundwell, R. M. (2005). Alternative strategies for large scale student assessment in Canada: Is value-added assessment one possible answer. *Canadian Journal of Educational Administration and Policy, 41* 1-21. Retrieved October 11, 2007 from http://www.umanitoba.ca/publications/cjeap/pdf%20files/crundwell.pdf

Kurial, R. (2005). Excellence in education: A challenge for Prince Edward Island: Final report of the task force on Student achievement. http://www.gov.pe.ca/photos/original/task_force_edu.pdf

Levin, B. (1998). An epidemic of education policy: (What) can we learn from each other? *Comparative Education, 34*(4), 131-141.

McEwen, N. (1995). Accountability in education in Canada. *Canadian Journal of Education, 20,* 1-17.

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education, 25*, 262-279.

National Assessment Agency (2008). Retrieved April 8, 2008 from: http://www.naa.org.uk/

No Child Left Behind Act. 2002: Pub. L. No. 107–10.

Organization for Economic Co-operation and Development. (2004). *Chile: Reviews of National Policies for Education*. Centre for Co-operation with non-members. Paris.

Qualifications and Curriculum Authority. (2008). Assessment and reporting arrangements. Retrieved April 8, 2008 from: http://www.qca.org.uk/eara/documents/KS2_v07aw-2.pdf
Principles for Fair Student Assessment Practices for Education in Canada. Edmonton, AL: Author.

Rogers, W. T., & Klinger, D. A. (2006, May).*Have the provincial achievement tests programs in Alberta and Ontario promised too much?* Paper presented at the Annual Conference of the Canadian Society of the Study of Education in Canada, Toronto, Ontario,
Santrock, J. W., Woloshyn, V. E., Gallagher, T. L., Di Petta, T., & Marini, Z. A. (2004). *Educational Psychology* (1st Canadian ed.). Toronto, ON: McGraw-Hill Ryerson.
Saskatchewan Ministry of Education. (2007c). Retrieved June 25, 2008, from http://www.sasked.gov.sk.ca/branches/aar/afl/aflreading.shtml
Taylor, A. R., & Tubianosa, T. (2001). *Student assessment in Canada: Improving the learning environment through effective evaluation*. Kelowna, BC: Society for the Advancement of Excellence in Education.

INVOLVING NCME IN POLICY DISCUSSIONS

In his "*From the President*" column in the June 2008 NCME newsletter, Mark Reckase called for the education measurement community to get more involved in policy:

> *Another major project for the NCME Board is determining how NCME can become involved in policy discussions related to educational measurement issues. It should be no major revelation that educational measurement plays a major part in educational policy. One of the strategic goals of NCME is to increase the organization's influence in the policy arena. It is important that all members of NCME help with this initiative. One way that you can do this is to help identify important policy issues that NCME should address as an organization. Another way is to become involved yourself. Be willing to make public statements about educational measurement policy.*

The present article represents reactions to that call from four measurement professionals:

- Steve Sireci: a university professor who contributes to the preparation of researchers and other professionals who work in the education measurement field;
- Wes Bruce: a state assessment director who is responsible for the administering statewide assessment and accountability systems;
- Jon Twing: a test vendor who contributes to the construction and operation of state-level assessments; and
- Ellen Forte: a consultant who advises state and federal agencies and other organizations on assessment and accountability policies.

While none of these individuals necessarily speaks on behalf of his or her colleagues, each provides a unique perspective on the notion of policy involvement and was asked to respond to three questions about the involvement of measurement professionals in policy conversations.

1. Should NCME become more involved in education policy?
2. How should other groups and individuals in the measurement community be involved in education policy?
3. What resources and supports would be necessary to engage measurement professionals in education policy conversations? In what ways should NCME be involved in providing these?

Panelists' responses to these three questions appear below.

Should NCME become more involved in education policy?

Sireci: As the preeminent professional organization for the educational measurement field, the National Council on Measurement in Education (NCME) needs to be involved in advising and setting educational assessment policy. Politicians and other policy makers view educational tests as tools that provide objective information. They are popular because there are no better alternatives for efficiently summarizing students' knowledge and skills. To the psychometrically uniformed (i.e., most people on the planet) students' test scores seem to be interpretable at face value. However, we know better. Even when tests are well developed, scored appropriately, and reported sensibly, the results are merely estimates of the underlying attributes we are trying to measure. To the well informed (i.e., active NCME members), the limitations of educational tests are considered alongside the benefits. Our knowledge of the benefits and limitations must not remain in our heads. It must be effectively communicated to those who set educational assessment policy so that educational measurement can have a positive impact on instruction and student achievement.

I would, therefore, recommend three activities for the organization: (1) publish policy briefs on key educational assessment policy issues, (2) respond to Federal and state assessment policy initiatives when needed, and (3) develop instructional materials for teachers and educational administrators to understand proper test use and the strengths and limitations of educational tests.

Bruce: NCME should absolutely be involved in measurement-related policy. Large numbers of education policy decisions that directly involve measurement issues are made with no regard to available data, best practice, sound understanding of principals involved, or how to evaluate the impact of the decisions. All levels of the "K-12" education system are understaffed, especially in areas of technical expertise. Policies are regularly written, review, approved and implemented almost without a single review by anyone in the measurement community. Graduation rates are a perfect example, after all the debate, a wide range of methodologies are still used and both critics and supporters routinely compare rates that cannot be compared (in a sound measurement sense) and use questionable methodologies. Do you know exactly how your state calculates "graduation rate?" Do you even know where to find the precise formula?

Twing: Mark Reckase's call for the membership of NCME to become more involved in educational policy is timely and relevant while perhaps at the same time, a little misleading. For example, some of my colleagues and I have been working with states, local schools and the USDOE regarding the implementation of policy decisions, if not the decisions themselves, for many years. Requests to testify at legislative hearings, presentations to Boards of Education, review of documents like the joint Standards as well as direct advice to 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 the areas of 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."

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. Such disconnects between what is arguably good purpose (like the high rigorous standards of NCLB) and desired outcomes (like all students meeting standard) can be trumped when measurement issues (like validity, student motivation, remediation, retesting and standard setting) are not carefully considered when the policy governing the implementation of the idea is generated.

Forte: As a general principle, NCME and its members should be more involved in policy because, like it or not, the uses to which test scores are put are defined by federal, state, and local policies. Policy provides the context for our work. Unfortunately, policy makers are generally very naïve about assessment. They focus on the policy goal and not the underlying measurement or evaluation model; thus, they assign far too much meaning and value to test scores and ignore the many assumptions and caveats that should appropriately accompany score use. To some extent, we should accept partial blame for this because our professional organization, one of the sponsors of the testing standards, has been absent from the policy conversation. In fact, if you'd like to see a bunch of blank looks, mention NCME on Capitol Hill.

While I think many of us might agree that NCME should be more involved the real questions are about the nature of that involvement and about how to get a seat at the table. A look at the organization's purposes listed on the NCME homepage (http://www.ncme.org) probably provides the best foundation for defining the nature of involvement as educational, rather than political, and Steve's three concrete activities are excellent ways to get moving on several effective fronts concurrently. Getting a seat at the table is the trickier venture, but if ever there were a time to step forward, this is it.

How should other groups and individuals in the measurement community be involved in education policy?

Sireci: Many of us are already involved in education policy. We serve on state TACs, Federal committees, or work at the local level. I believe good people are providing good advice, but I would like to see that expertise aggregated somehow to an organizational level so that NCME can promote coherent and sound advice on key assessment policy issues. The development of instructional materials for lay audiences and for NCME members is one way NCME can provide resources and support for members to engage and inform policy makers. But we must remember that NCME is essentially a volunteer organization. Therefore, we must all pitch in to make this happen. We need a system for synthesizing the excellent work by individual members and research centers so that we can create and disseminate our expertise to the policy makers who have a direct effect on education.

It's important to note, too, that the testing industry is essentially unregulated, measurement expertise within state departments of education is scarce, and TACs often have little power or meet too infrequently to have as much impact as they should. Many of us have witnessed excellent advice from a TAC that was ultimately rejected by policy makers for reasons that may or may not be clear. Moving from the state level to the national level, the need for measurement expertise is even greater. NCLB

policies like adequate yearly progress for schools and subgroups, and educational gain policies for adult education students under the National Reporting System are just two examples of Federal legislation that could benefit from more input from the psychometric community.

Bruce: Boards of Education at all levels have wide decision making authority and they need information, suggestions, and advice from measurement experts. That can come in a number of ways, speaking at a "Board" meeting, serving on committees, testifying to legislative committees or even advising districts and states. Often as a "cog" in the machinery of education we can be viewed as part of the "problem" or uninterested in change, an independent voice (or voices) can provide balance to discussions. I am not looking for a rubber stamp or a cheerleader, what I want is a reasoned and informed perspective on the policy at hand. As policies are formulated it would be great to have a group of independent "advisors" who would be willing to help me wrestle with correct methodologies and help weigh the pros and cons of a particular policy.

An example: Indiana's graduation rate has been the object of constant legislative revision each of the last five sessions, not once in all of the committee hearings or floor debates did any independent measurement professional speak to the strength or limitations of any of the iterations of the formula.

Twing: We as 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 writing papers and doing research regarding aspects of current or planned policy we do not like or endorse. We often seek out the simpler venue of conference presentations to let our opinions be known regarding what we think is wrong with education and how to fix it, which inevitably means, we make a policy recommendation.

Not only do I believe NCME and its members are involved in policy making, but I 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 methods used by the service provider. 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 causing the debate to happen it never would and resulting decisions would most likely be poorly informed.

Let me offer 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, even referring 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 state) ask me simple questions like: Why is NAEP the standard for comparison for our NCLB assessments? NAEP does not measure our content standards very well, why are our NAEP scores being scrutinized so much? What research exists that shows NAEP is a good vehicle to judge education policy both statewide as well as for NCLB? Now, why is it that my customers—the very people making educational policy at the state level—are 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, 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 education policy.

Forte: Every measurement professional can contribute to better measurement-related policies. Some do so by publishing articles about their basic research to ensure that our substantive foundation continues to strengthen. Others contribute by adhering to our professional standards when designing and implementing assessment systems and avoiding bad practices even when they're more lucrative than the "right" way. Some can get or stay involved with Technical Advisory Committees. We can all become more attuned to the challenges practitioners face, often as a result of an unwieldy policy mandate, and engage in research or evaluation projects related to real-world problems. A willing few can become more proactive in their state legislatures or with members of Congress and even get to know the people who extend the invitations to the policy table. Also, when attempting to communicate outside of our journals and technical contexts, we have to be willing to speak and write using language that doesn't require a dictionary to understand or take a policy-maker more than five minutes to read. The least help is provided by those in our field who (1) rail against the policies we all know to be technically unsound without offering politically-tenable alternatives or solutions and (2) describe or refer to measurement concepts as being too complex for policy-makers to understand. Protecting our expertise from the outside world is how we got AYP.

What resources and supports would be necessary to engage measurement professionals in education policy conversations? In what ways should NCME be involved in providing these?

Sireci: Given the need for psychometric expertise to inform educational assessment policy, what can NCME do? There are many possibilities. First, let me point out what NCME has already done. For over 50 years, we have participated in the development of the *Standards for Educational and Psychological Testing*[4]. These documents have stipulated and described professional standards for test development, evaluation, and interpretation, and have had a positive influence on educational testing. NCME was also a key player in development of the *ABCs of School Testing* (1993), a video produced to communicate important educational measurement concepts to lay people such as teachers, school administrators, and parents. NCME was also a driving force in the development of the *Standards for Teacher Competence in Educational Assessment of Students* (1990). This document came from a collaborative effort with the American Federation of Teachers (AFT) and the National Education Association (NEA) and described the key assessment knowledge and skills teachers need to function successfully. Although I am proud of these efforts, they are dated. The *Standards for Educational and Psychological Testing* are being revised with NCME participation, and a committee is being put together to consider revision of the *ABCs of Testing*. However, I do not think we have worked on anything with AFT or NEA in recent years, which is surprising given the increased use of tests in schools.

Bruce: Those of us who have to implement testing policies need help in answering question in the many areas where practice has outstripped research and evaluation. The education community and policy makers would all benefit if we had more information on questions like:

1. Is there a difference between mastery of "English language proficiency" and the mastery of "English standards?"
2. Do interim assessments work?
3. Can the use of "formative assessment" practices reduce later need for "remediation"?
4. Do accommodations work? Are they the "same" for students with disabilities and for English language learners?
5. Is the dropout rate the inverse of the graduation rate? What should it be? What are the policy decisions/implications of a "four year on time" graduation rate?

In addition it would be great to have policy papers, methodology guides and willing voices to discuss emerging issues where education and measurement meet.

Twing: When considering how NCME should support involvement in policy discussions, 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 of mine, 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 this 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.

Forte: Once upon a time, student assessment was a rather simple endeavor. A school district wrote a check to a testing company, required its students to spend a few hours answering some questions, and a few weeks later got some reports that it used as it – or its Board of Education – saw fit. Whether one considers those the good old days, the dark ages, or an era equally offensive to the present, times have changed. Control over who, what, when, and why to test have been wrested from the local level. Federal policy now defines the critical features of most large-scale testing and test score use and that policy is, for the most part, unencumbered by technical considerations. While it's silly to think that every one of us should get involved in policy, our professional organization has as part of its mission the dissemination of knowledge about measurement "theory, techniques, and instrumentation…and procedures appropriate to the interpretation and use of such techniques and instruments" (http://www.ncme.org/about/mission.cfm). In addition to the types of activities Steve and others have noted above, this should extend to the consideration of the programs through which we train measurement professionals. How many of our graduate programs require courses on education policy or opportunities for our state and local practitioners to become more technically savvy? The answer is "far too few" and I imagine the p-value for that item, if presented to the NCME community, would be near 1.0. The real question is what we're going to do about it.

[4] The first edition (*Technical Recommendations for Psychological Tests and Diagnostic Reports*) was developed by the American Psychological Association, the American Educational Research Association, and the original name of our organization, the National Council on Measurements *Used* in Education (1954). Subsequent revisions were published in 1966, 1974, 1985, and 1999.

Scott Bishop
Data Recognition Corporation
13490 Bass Lake Road
Maple Grove, MN 55311

Phone: 763.268.2029
Fax: 763.268.2529
e-mail: SBishop@DataRecognitionCorp.com