THE STANDARD DEVIATION

January 2008

Washington Educational Research Association
University Place, WA

http://www.wera-web.org

Festschrift Honors Prof. Cathy Taylor

Four recent University of Washington College of Education doctoral graduates honored their teacher and advisor, Prof. Cathy Taylor, at the December WERA/OSPI State Assessment Conference. Fellow UW doctoral graduate (two decades earlier), Peter Hendrickson invited his colleagues last year to report on their current research which had been influenced by Taylor's teaching and ongoing counsel.

Presenters were:

- Nina Salcedo Potter of Shoreline, Integrated Math and WASL
- Feng-Yi Hung of Clover Park, Academic Growth in Math for Students in Poverty—WASL and MAP
- Yoonsun Lee of OSPI, Investigation of Test Structure of the WLPT-II
- Jack Monpas-Huber of Spokane, Validity Issues for Medium-Scale Assessments

In academic circles, this practice is called a Festschrift and often happens towards the end of a career. In Taylor's case, she has taken a leave of absence from UW to direct assessment alternatives and innovations at OSPI after many years consulting with OSPI on the development and analysis of the WASL. Associate Dean for Research Deborah McCutchen, a College of Education colleague, served as a discussant, as is done at AERA and other professional conferences. Hendrickson chaired the session and kept time, delighted with the fond tributes by these measurement professionals to their mentor who knew only that she was invited to the session.

Each of the presenters received feedback on their work and they have revised their papers for *The Standard Deviation*. Many of the PowerPoint presentations are available online at the WERA website. Several of the session evaluations suggested that future conferences provide similar paper sessions with a discussant to explore a theme from several directions.

The Festschrift Papers start on page 12

Festschrift participants from left P. Hendrickson, D. McCutchen, N. Potter, C. Taylor, J. Monpas-Huber, F. Hung, and Y. Lee

INSIDE THIS ISSUE

President's Column

Thank you to everyone who contributed to our learning and tested our assumptions at the WERA/OSPI Winter Assessment Conference by presenting.

I enjoyed each of my sessions and learned a great deal. I left Diane Browder's pre-conference session with concrete examples and strategies for aligning instruction and assessment to state content standards, including how to interpret the GLEs for students with significant cognitive disabilities. I have already incorporated the information into my work.

Jennifer Lloyd's session "On Analyzing Change and Growth When the Measures Change Over Time: Measurement and Methodological Issues and a Novel Solution" was fascinating and there was much discussion and professional dialogue during the session. While not immediately applicable to my work, it provided an opportunity to increase my knowledge and understanding of statistics and methodologies that might be used in educational research. I sincerely hope that the sessions you attended were as informative and challenging as mine.

The spring conference, "Leaders, Learners, and Change Agents," promises to provide leaders at all levels with research and strategies to further our work to create socially just schools. I encourage each member to invite someone who is not currently a member to attend the conference with you, or several someones!

Membership in WERA provides a professional connection to other educators who engage in qualitative and quantitative inquiry, read and analyze current research and use data and information to inform daily practice in educational settings from kindergarten through post-doctoral pursuits.

We all share the responsibility of preparing each child in our state for a productive and personally rewarding future in our democratic society. While the task often seems overwhelming in isolation, working in collaboration, educators armed with passion, data and current research about learning can make a difference in the future of the children for which we are responsible.

We have lots of room in WERA for more passionate educators who want to become members. I look forward to seeing you and your colleagues in March

-Lorna Spear, Ed.D, *is Executive Director For Teaching and Learning Services in the Spokane School District. She was a much decorated elementary principal and is WERA President.*

The mission of the Washington Education Association is to improve the professional practice of educators engaged in instruction, assessment, evaluation, and research.

WERA Services

- *WERA provides professional development through conferences, publications, and seminars.*
- *WERA provides forums to explore thoughtful approaches and a variety of views and issues in education.*
- *WERA provides consultation and advice to influence educational policy regarding instruction, assessment, evaluation, and research.*

Plan to Attend the Spring Conference!

March 26-28, 2008
Seattle Airport Hilton Conference Center

"Change" is the current political buzzword, and WERA is fitting right in. The Spring conference theme is *Leaders, Learners, and Change Agents*. It will explore a range of issues that will stimulate your thinking about what is changing, what needs to change, and how to lead and learn as change occurs in K-12 education in our state.

Two **keynote speakers** will challenge and inspire us with ideas about school and district change.

- Dean Fink, author and former teacher, principal and superintendent in Ontario, Canada, will give the keynote address Thursday morning on *Leadership for Mortals: Developing and Sustaining Leaders of Learning*. His address will dispel leadership myths and provide a model to develop and sustain individual leaders in schools.

- Carl Cohn, a former teacher, counselor, central office administrator, and superintendent in San Diego and Long Beach (winner of the Broad Prize) will give the Friday morning keynote address on *Lessons Learned about Leadership and Change in Urban School Districts.* The address will discuss how a collaborative leader uses non-confrontational methods to improve student achievement and close the achievement gap.

Optional **pre-conference workshops** will be held Wednesday, March 28. Dean Fink, Thursday's keynote speaker, will offer an all-day workshop on *Sustainable Leadership* based on the book he co-authored with Andy Hargreaves. Eight other workshops will be offered. Here are the topics:

All Day: Sustainable Leadership

Morning Only

- Managing the New Graduation Requirements: Lessons from the Field
- Leaders in Learning: Building Capacity for Teacher Leadership
- Making Sense of the Math Revisions
- Fighting Reform Fatigue: Frameworks and Formats for Continuous School Renewal

Afternoon Only

- Program Evaluation in a Nutshell
- Reframing Leadership and Cultural Competency
- School-Based Instructional Coaches
- Exploring the Impact of the New Math Standards

Breakout sessions on Thursday and Friday will provide useful information on a range of topics, including professional learning communities, school and district improvement practices, characteristics of effective educational leaders, updated information on the WASL and other assessments, and emerging state education policies.

With 2008 being an election year, the final event will be the **Pete Dodson Symposium panel** where candidates for state superintendent and representatives of the two main candidates for governor will discuss their views about education in our state, the problems that need to be addressed, and what should be done about them. For the first time, *we will open this event to the public for a small fee for those who have not paid the full conference registration fee.* Students from Tacoma will provide entertainment after lunch and before the panel begins.

Make your own **hotel reservations** at the Airport Hilton at $179/day for a single or double room. This rate is guaranteed until March 7. Call the hotel directly at (206) 244-4800 or use 1-800 HILTONS. Be sure to mention the WERA Conference to get this special rate. You can also register via credit card from a hotel link on the WERA Web site at www.wera-web.org.

The planning committee has done a great job putting together a great conference. Registration and pre-conference workshop information are available on the WERA Website. The complete program will be available in early March. We hope to see you there!

Dr. Dean Fink
Pre-conference Presenter
Thursday Keynote Speaker

Dr. Carl Cohn
Friday Keynote Speaker

Future Calendar

WERA Items

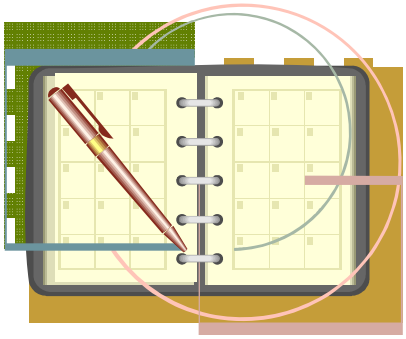
- Developing Computational Fluency in Mathematical Thinking, February 9, 2008
  Puget Sound ESD
- WERA Test Directors WASL Operations Meeting, February 14, 2008
  Hilton Seattle Airport Hotel
- 2008 Spring Conference, March 26-28, 2008
  Hilton Seattle Airport Hotel
- 2008 State Assessment Conference, December 3-5, 2008
  Hilton Seattle Airport Hotel
- 2009 Spring Assessment Conference, March 25-27, 2009
  Hilton Seattle Airport Hotel
- 2009 State Assessment Conference, December 9-11, 2009
  Hilton Seattle Airport Hotel

Contact: http://Wera-web.org

Other Calendar Items (Non-WERA)

- American Educational Research Association, National Council on Measurement in Education, National Association of Test Directors, Directors of Research and Evaluation Annual Meetings and Conferences, New York, NY. March 23-28, 2008
- American Evaluation Association Annual Conference, Denver, CO. November 5-8, 2008
- WSASCD Annual Conference, Spokane, November 6-8, 2008
  www.wsascd.org

OSPI Conferences Contact:
http://www.k12.wa.us/Conferences/default.aspx



WERA Assessment Director's Network coordinator Bob Silverman of Puyallup presents Robin Munson of OSPI an exceptional service award at the December pre-conference

Book Review: Leadership for Mortals by Dean Fink

Reviewed by Phil Dommes, Ph.D.

Dean Fink begins and ends *Leadership for Mortals* with the conclusion that great leadership is an attainable goal for ordinary mortals, but only with extraordinary commitment, effort and determination. The remainder of his book thoughtfully outlines a model for how such leadership might be developed and sustained in the contemporary educational milieu. Fink unfolds his leadership model with clarity and compassion, blending substantial theory with interesting and well-chosen stories from his rich career experiences.

Initially, he shows how today's leaders develop and work within organizational structures that have changed little in the last hundred years. Despite challenging conditions, Fink states that successful leaders remain "passionately, creatively, obsessively and steadfastly committed to enhancing deep learning for students—learning for understanding, learning for life, learning for a knowledge society." Good leaders express this commitment to learning by "communicating invitational messages…in order to build and act on a shared and evolving vision of a learning-centered school."

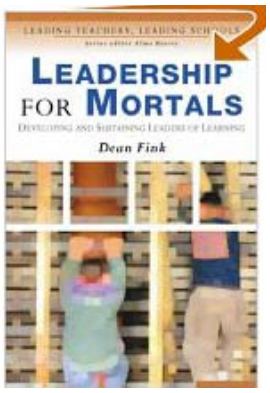
To communicate these messages to a receptive audience, leaders must reflect the values of trust, respect, optimism and intentionality. In order to maintain a proactive and positive stance in the face of inevitable obstacles to their vision, Fink encourages the appropriate development and balanced exercise of reason, ethics, common sense, imagination, intuition, and memory.

Furthermore, he notes that successful leaders must be armed with a variety of learnings – contextual knowledge, political acument, emotional understanding, understanding of learning, critical thinking, and an understanding of connections. Fink goes on to discuss the trajectories or career moves that help determine the extent to which leaders successfully move an organization towards a shared and realized vision. He concludes with a discussion of what it takes to sustain leadership across generations.

Leadership for Mortals offers a useful framework for reflecting upon one's personal leadership journey and upon leadership development more generally. It is a short text –only 164 pages – and many of the chapters are conceptually rich enough to merit a book of their own.

Publication Data: *Leadership for Mortals: Developing and Sustaining Leaders of Learning,* by Dean Fink, 2005. Corwin Press, Inc., Thousand Oaks, CA. Paperback, 170 pages, $31.95 (US) ISBN: 9781412900539

-Phil Dommes, Ph.D., *is Director of Assessment and Gifted Programs for the North Thurston School District in Lacey. He is currently a WERA Board member.*



Book Review: Sustainable Leadership by Hargreaves and Fink

Reviewed by Ali Williams

"Change in education is easy to purpose, hard to implement and extraordinarily difficult to sustain." (Hargreaves & Fink, 2006, p.1).

The authors of *Sustainable Leadership* provide research and clear evidence about the impact of sustainable leadership and the direct parallels between education, the environment and our lives. This is an excellent book for all educational leaders to read, share and discuss with their district peers and leaders. The authors provide strategies to help with change, not just in one school, but also as a system. The seven principles of sustainability in educational change and leadership are:

1. depth
2. length
3. breadth
4. justice
5. diversity
6. resourcefulness
7. conservation

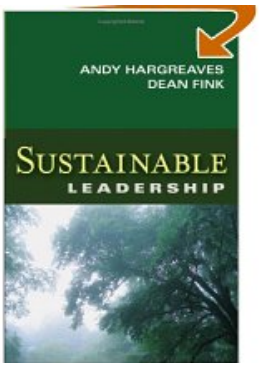
Each chapter goes into depth on the research that they have conducted around these principles. They conclude with five action principles which put theory into practice and are based on the authors' commitment to environmental sustainability which they believe, like leadership sustainability, is a moral imperative. The five action principles are:

1. activism
2. vigilance
3. patience
4. transparency
5. design

Fink and Hargreaves are clearly able to define leadership; they stress that successful educational leaders need to continue to work together for long-term positive impact on student learning. Education needs to be treated as a long lasting enterprise, not a temporary business looking for quick fixes. This is a great read for educational leaders who want to make a sustained difference and improve our educational system.

Publication Data: *Sustainable Leadership* by Andy Hargreaves and Dean Fink, 2006. Jossey Bass, San Francisco, 352 pages, $25.00 (US) ISBN: 10:-07879-6838-2

-Ali Williams *is a veteran teacher and recently became principal of Explorer Middle School in Mukilteo. She is active in WERA as a conference planner.*



Test Director Network Update –A WERA Affiliate

The WERA Test Directors Network will meet February 14, 2008 at the Seattle Airport Hilton to review WASL test operational issues with OSPI staff. The meeting will be in the Crystal B conference room from 9 A.M. – Noon (8:30 for visiting with colleagues). Attendees have been asked to provide any information regarding the locally determined assessment system for special education students. Test Directors are invited to bring an administrative assistant. Inquiries to convener Bob Silverman at Puyallup Schools. Contact him at SilverRJ@Puyallup.k12.wa.us.

Past WERA President (two terms) Duncan MacQuarrie has prepared a Test Director Survey to collect information about district assessment systems and duties of test directors. Members have been contacted by e-mail with a link to the web survey.

And the Bumper Sticker Winners from the WERA/OSPI State Assessment Conference (Praeger's Follies Contest) are…

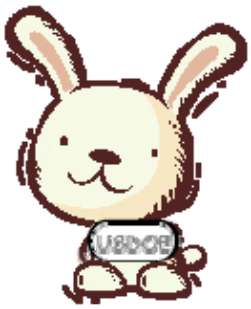
Class of 2007…Thank God!

–Brandon Lagerquist, Northshore School District
First Place

Bob Silverman (left) and Michael Power read entries to the Bumper Sticker Contest.

Brandon Lagerquist (left) receives Bumper Sticker Award from Bob

Continued from previous page...



It's cute how you think i'm listening.

–Diana Stray, Federal Way Public Schools
Second Place

So many CAA Options. So little time.

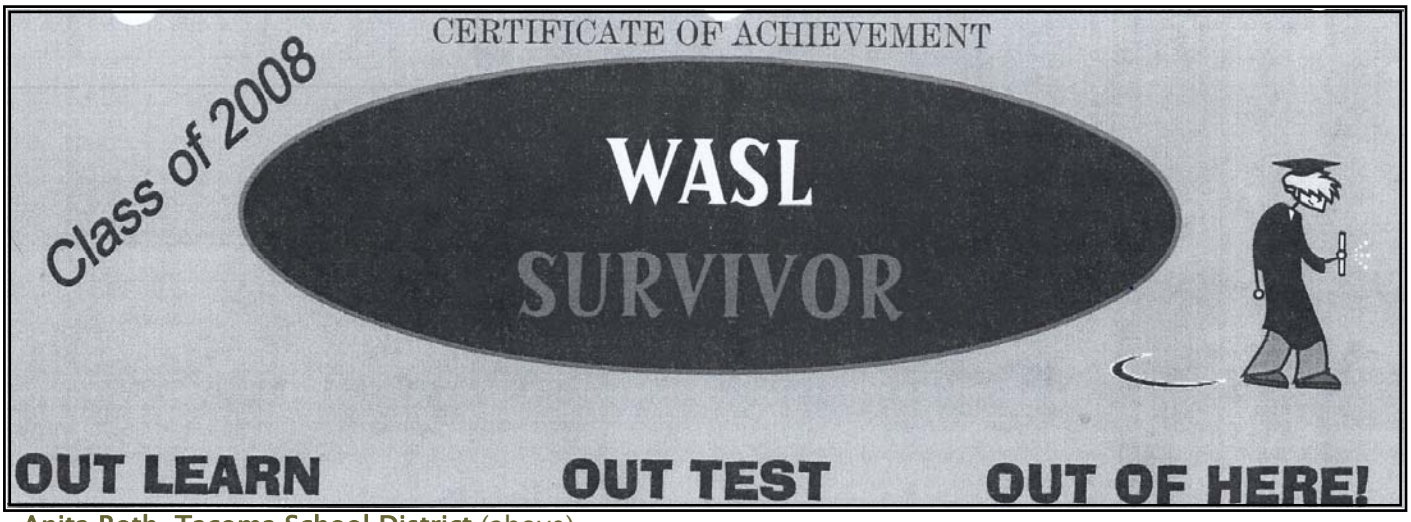
–Nancy Katims, Edmonds Public Schools
Third Place (tie)

My child is a CIA/CAA/CAA Option/GPA CoHort student at Washington High

–Bob Isenberg, Kent School District
Third Place (tie)

Continued from previous page...

Honorable Mentions



–Anita Roth, Tacoma School District (above)

"My other car is an armored assessment transport truck"

–Deidra McCollum, Pasco School District

The Lottery:
a tax on people who failed the probability & statistics strand

–Nancy Katims, Edmonds Public Schools

MATH: Don't Leave School Without it.

–Karrin Lewis, OSPI

Lessons from Bolivia: Multicultural Research

-By Peter Hendrickson, Ph.D.

Each of us conducting or supporting applied research in education from time to time finds ourself "just visiting" in another field, but curious. Over Thanksgiving, I accompanied a UW biostatistician (my wife) to a Bolivia planning meeting of University of Washington researchers and Latin American neurointensivists mounting an ambitious, randomized and observational study of traumatic brain injury (TBI) treatment.

Several similarities to our work with children unfolded over a long weekend in Santa Cruz, six flights from Seattle, south of the Equator, submerged again in Spanish after several years absence from my first teaching job in Ecuador.

Culture of Research

These physicians focus first on the needs of their patients as do teachers, principals and others with their students. Rarely is there resource or time to conduct high quality research within their profession. Like us, they've been trained and conditioned to practice, not to create knowledge. Most master's programs in education focus on instructional practice with scant attention to either reading studies in tier 1 journals or to conducting research.

Community of Researchers

We arrived from Quito (via Lima and La Paz) at 3:30 a.m. and by noon were meeting with the steering group, a three-hour lunch at an open-air restaurant. A cell phone call alerted an Argentine intensivist that a local hospital ICU had a new TBI patient. I had not been in scrubs since surgical orderly work in Los Angeles decades ago but the ER and ICU memories returned with the bedside briefing from a resident. The public hospital air-conditioning had failed and I nearly passed out in the steamy 40 c. heat. Santa Cruz is known for its proximity to dense jungles, although it is an agricultural center for the neighboring savannah.

An impromptu training by a UW resident neurosurgeon on the insertion of an Intracranial Pressure Monitor followed with a plastic skull as it was not in the patient's interest to undergo the procedure. We learned that Black and Decker drills with a surgical bit were the local instrument of choice to make a hole in the skull. Another three-hour meeting followed at a second restaurant—this before the Saturday and Sunday official meeting of physicians from Seattle, Columbia, Ecuador, Bolivia, Argentina, Uruguay and Brazil. Building a research team requires common time together, not all of it strictly focused on the study.

After graduate school, without any significant grants to fund research, and without close links to a research university, it is difficult to work out the details of a research project with colleagues. Where do we hear the cry, "Each and every person involved in collecting the data is 100% responsible for data accuracy, else error may obscure the power of the intervention"? Where do we hear, "Folks must be comfortable to say, 'We have a problem. These data are not accurate.'"? Where do we have the opportunity to turn clinicians (teachers) into researchers? I was heartened to hear several doctors reinforce a staffer's assertion that being a research group and developing partnerships is more important than the money. "Building capacity is a challenge we will all work on as partners," he said. This network began 10 years ago in Argentina resulting in the Gaucho Coma Data Base and this meeting marked the birth of a much broader coalition.

How wonderful to have colleagues of many years standing; how critical is the work to build those relationships so that important work can be done.

No Light Without Heat

A full day was spent on study design but it was preceded by introductions and brief PowerPoint presentations from the principal investigator at each of the seven sites. Questions about the study, the measures, and the applicability to local hospitals were long, detailed and frank. In every case, study leaders affirmed the validity, even the necessity of the questions. While the 5-year study was already approved and funded, refinement of the treatment and outcome protocols was yet underway. Commitment and fidelity to the final protocols by the physicians and their staffs relied on close initial scrutiny and influence over their final shape.

Continued from previous page…

Reliable Measures

Physicians, it turns out, are less experienced collecting reliable data in the service of meticulous, controlled science than they are at treating patients, how like our teaching and administrative colleagues. TBI research is well supplied with reliable scales of trauma severity, well established in the literature and practice. There are also scales measuring cognitive, vocational, emotional and other outcomes but these measures are more culturally and linguistically bound. Significant challenges exist to collect data from non-literate, subsistence farmers who live hours from the nearest trauma center or from homeless favella dwellers. I found myself envious of the TBI scales, wishing we had some as universally known in reading, writing or mathematics.

The center of this experience was the growing understanding that trauma doctors and school professionals have much in common around research issues. The urgency of meeting the needs of the patients or students leaves little room for thoughtful research. There is greater emphasis on staffing, instruction, curriculum and materials than on growing knowledge through research. And there is little reward for researchers outside of higher education or those businesses with a profit motive.

WERA's aim is to move us, when possible, towards becoming both practioners and researchers. *The Standard Deviation* and our conferences promote and invite that shift.

-Hendrickson *was a new teacher in Ecuador in the early 1970's, later a principal and curriculum director, and is now an assessment, research and program evaluation specialist for Everett Public Schools. He also edits The Standard Deviation.*

Traumatic Brain Injury researchers meet late into the night in Santa Cruz, Bolivia

Festschrift Papers (I)

Integrated Math Curriculum and the Math WASL

-By Nina Potter, Ph.D.

Introduction

There are currently a large number of curricula being used to teach math at the high school level across Washington State. While the curricula differ in a number of ways, I will be grouping them into three categories: Traditional curricula, Integrated/Traditional curricula and Integrated/Inquiry Based curricula.

- Traditional curricula typically teach topics such as algebra and geometry independently. Typically there is a three-year sequence: Algebra, Geometry, Algebra 2/Trigonometry. Students are introduced to topics and formulas and given examples of how and when to use them.
- Integrated/Traditional curricula typically have three year-long integrated courses, which cover the same material as traditional Algebra 1, Geometry and Algebra II courses. The textbooks are very similar as traditional textbooks, Algebra and Geometry topics are, for the most part, taught independently.
- Integrated/Inquiry Based curricula, such as IMP, integrate algebra, geometry and other topics, such as statistics, probability, curve fitting, and matrix algebra. Units are generally structured around a complex central problem. Although each unit has a specific mathematical focus, other topics are brought in as needed to solve the central problem, rather than narrowly restricting the mathematical content. Ideas that are developed in one unit are usually revisited and deepened in one or more later units. The textbooks are very different than traditional textbooks with more text and less formulas than traditional math text books.

Research Questions

The typical three-year course sequence for the Traditional Curricula is Algebra 1, Geometry, and then Algebra 2/Trigonometry. For Integrated curricula the courses are Integrated 1, Integrated 2, and then Integrated 3. Students who go on to a fourth year of math have choices such as calculus and statistics, using any of the types of curriculum. Most 10th grade student taking the WASL will be enrolled in either a Geometry or Integrated 2 course. Some students who struggle with math will be enrolled in lower level courses and some students who excel in math will be taking more advanced courses. The WASL is targeted at 10th grade skills so the assumption would be that students enrolled in (and succeeding in) Geometry or Integrated 2 should be successful on the WASL. Those students who are in lower level courses would not be expected to do very well on the WASL while students in more advanced courses should do very well on the WASL.

The main question for this study is whether there is a difference in performance on the WASL for students enrolled the different curricula. That is, is there a difference in performance for students enrolled in a traditional geometry course, the Integrated 2 course of an Integrated/Traditional curriculum or an Integrated/Inquiry Based curriculum? What about students who are taking a more advanced course, is there a difference in performance depending on the curriculum being used?

Data

Students who were in 10th grade at the time of the spring 2007 WASL were chosen for this study. Only students who took the regular WASL with or without accommodations were included.

The data collected included:

- WASL scores
- Math course enrolled in during spring semester
- Spring Semester grade in math course

Continued from previous page...

Results and Discussion

The first three graphs, Figures 1 to 3, show the percent of students enrolled in the typical second year course during their 10th grade scoring at each level on the 10th grade WASL.

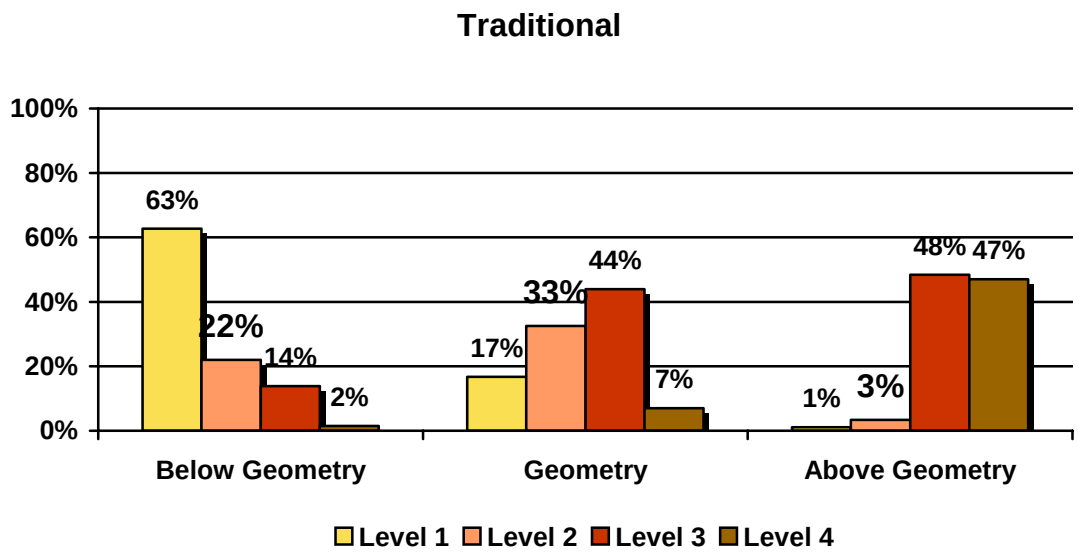


Figure 1. Math WASL pass rates, traditional math courses.

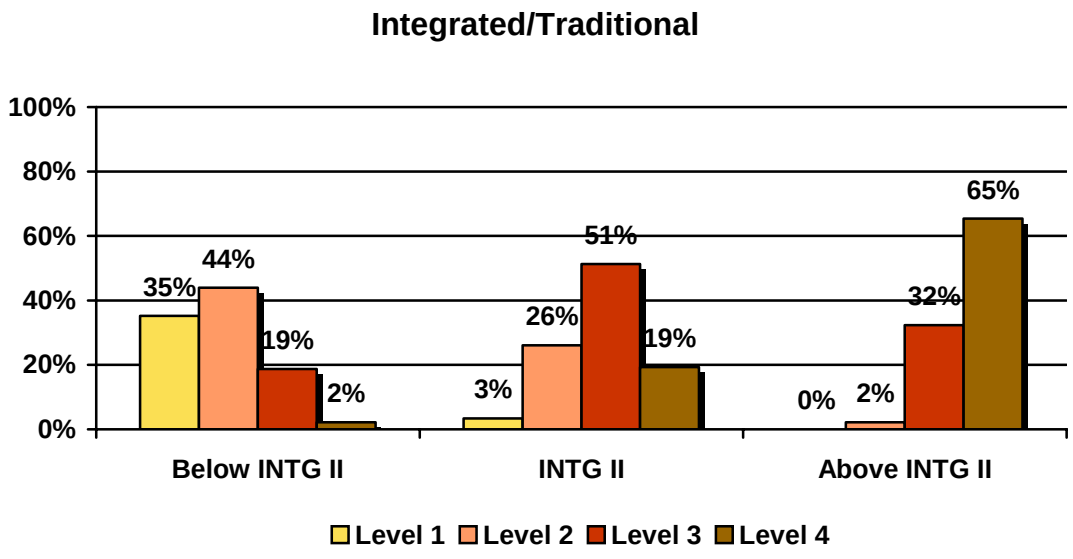


Figure 2. Math WASL pass rates, integrated/traditional math classes.

Continued from previous page...

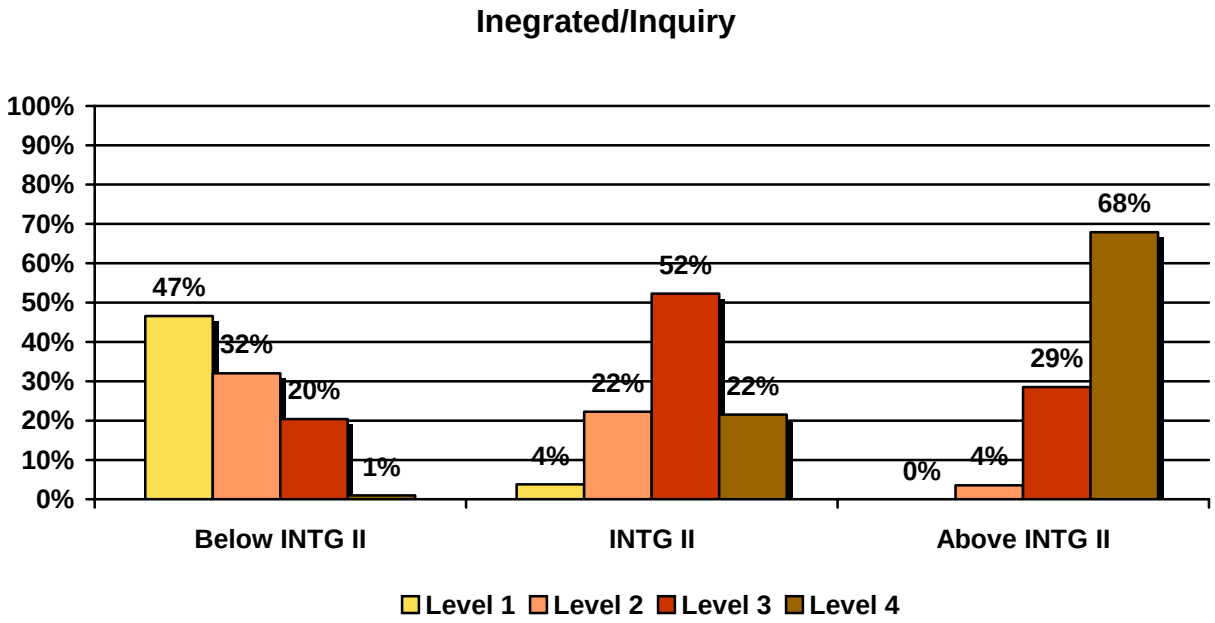


Figure 3. Math WASL pass rates, integrated/traditional math classes.

At first glance, it appears that the students in either type of integrated math curricula outperform students in a traditional geometry course. However, there are a few considerations when looking at this data. For example, since the data come from different districts and schools, it is possible that there are other confounding variables that would explain the difference in performance other than the curriculum. One simple way to explore this possibility is to limit the students in the analysis to those receiving an A or B in the class. Regardless of demographic or any other possible confounding variables, it can be assumed that students who receive an A or B in the class understand the material being covered. If we limit the comparison to only students receiving an A or B, we can more safely conclude that differences in performance on the WASL are due to how and what students are being taught. Table 1 shows the percent of students enrolled in geometry or integrated 2 receiving an A or B who met standard on the WASL.

Table 1 WASL math pass rates for A and B students

Traditional Geometry	73%
Integrated/Traditional	72%
Integrated/Inquiry Based	83%

These results suggest that students enrolled in the Integrated/Inquiry Based curriculum are outperforming their peers in the more traditionally taught courses.

Table 2 shows the results for students enrolled in more advanced courses. These results suggest that students enrolled in more advanced courses do well on the WASL regardless of the curriculum being used.

Table 2 WASL pass rates for advanced students

Above Traditional Geometry	95%
Above Integrated 2/Traditional	96%
Above Integrated 2/Inquiry Based	97%

Continued from previous page…

Other Questions/Future Research Studies

This is meant as a first look into differences in performance on the WASL based on math curriculum. I would like to do the same study with a larger sample size and see if the results look the same. Since very often there are other factors such as behavior included in grades, I would also like to do some kind of multiple regression analysis where I cannot only look at grades and curriculum, but at the same time explore whether there are other variables such as gender or socio-economic status that influence these results.

While there appear to be differences in performance on the WASL depending on the curriculum being used, in general students who are performing at grade level tend to do well on the WASL. On the other hand, those students who are taking courses below Integrated 2 or Geometry are not successful on the WASL. Students who get a grade below C in the regular 10th grade class are also less successful on the WASL. Currently we are looking more closely at these students and tracking some of their educational history, including grades and test scores. We want to know if and when we can identify these students early and find some kind of intervention or help early on.

-Nina Potter, Ph.D. *is Director of Assessment and Student Information in the Shoreline School District. Contact information:* nina.potter@shorelineschools.org

Festschrift Papers (II)

Investigation of Test Structures of the Washington Language Proficiency Test-II (WLPT-II)

-By Yoonsun Lee, Ph.D

The purpose of the study

This study is to investigate test construct on the Washington Language Proficiency Test-II (WLPT-II). A test construct (dimension) is defined as a theoretical representation of an underlying trait, concept, attribute, processes, or structure that the test is developed to measure (Messick, 1989; Ackerman et al., 2003). WLPT-II was developed for four grade bands: primary, elementary, middle, and high school; in four modalities: reading, writing (writing and writing conventions), listening, and speaking in accordance with the Standards for Educational and Psychological Testing (American Educational Research Association, 1999) and the Washington State English Language Development (ELD) standards. Table 1 summarizes the overview of the grade band. Because three to four grade levels take the same test, easy items for lower grades and more challenging items for higher grades needed to be included. To create better alignment with Washington State ELD standards, augmented items were added to the original test, which was based on the Stanford English Language Proficiency (SELP) test (2006 WLPT-II technical report, 2006). In this study, four models were tested to evaluate test structure of the revised WLPT test (WLPT-II). First, the unidimensionality was examined (Model 1). Second, the same unidimensionality in Model 1 was tested with errors correlated in each subtest. Third, multidimensionality with four independent factors representing each modality was investigated. Finally, the hierarchical model was studied.

Table 1. *Grade Levels in each Grade Band*

Grade Band	Grade Level
Primary	Kindergarten, Grade 1, Grade 2
Elementary	Grade3, Grade 4, Grade 5
Middle	Grade 6, Grade7, Grade8
High School	Grade 9, Grade10, Grade 11, Grade 12

Method

Data

The data for this study came from a statewide language proficiency test for Kindergarten through Grade 12. In 2006, the test was administered in four modalities: reading, writing (writing and writing conventions), listening, and speaking. Approximately 15,000 students were included in each grade band, totaling approximately 60,000 students for the study. The data contained nearly the same proportion of male and female students. The most populous ethnic group in Washington is the Latino group, followed by the Russian and Asian groups.

Instrument

The Washington Language Proficiency Test (WLPT) was developed for four grade spans (K-2, 3-5, 6-8, 9-12) in reading, writing (including writing conventions), listening, and speaking to assess the English language proficiency of ELLs. The test was developed based on the Stanford English Language Proficiency (SELP) Test. Because the SELP test did not fully cover the State's ELL standards and did not have three to four grade-level-appropriate items, augmented items were added to the test, creating the WLPT-II test. Table 2 shows the number of items and points in each grade band and each subject. Listening, reading, writing (writing and writing conventions), and speaking are assessed through multiple-choice (MC) and constructed-response (CR) items. The total number of items per grade band varies. As seen in Table 2, the number of items in reading and writing conventions increase as the grade band increases, whereas listening and speaking have the same number of items across the four grade bands.

Continued from previous page...

Table 2. Number of Items/Points on WLPT-II

Grade Span	Speaking	Listening	Reading	Writing			Total Number of Items (Points)
				Writing Conventions	Short Writing	Writing Prompt	
K-2	17 (38)	20 (20)	21 (21)	15 (15)	6 (10)	2 (8)	81 (112)
3-5	17 (38)	20 (20)	24 (24)	20 (20)	0	2 (8)	83 (110)
6-8	17 (38)	20 (20)	28 (28)	24 (24)	0	2 (8)	91 (118)
9-12	17 (38)	20 (20)	31 (31)	24 (24)	0	2 (8)	94 (121)

Note: Numbers in the parentheses indicate score points.

Data Analysis

To evaluate the internal factor structure, Confirmatory Factor Analysis (CFA) was conducted using EQS (Bentler, 1995) with Maximum Likelihood Estimation. The four models shown in Figure 1 were tested.

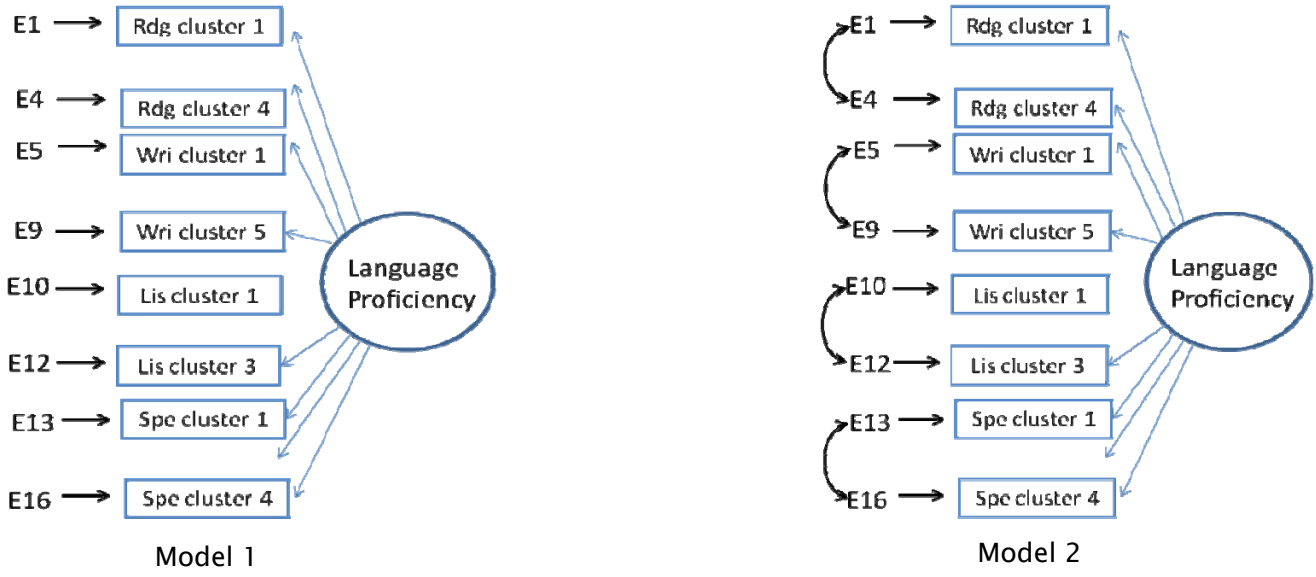


Figure 1 Four models tested continues on next page

Continued from previous page...

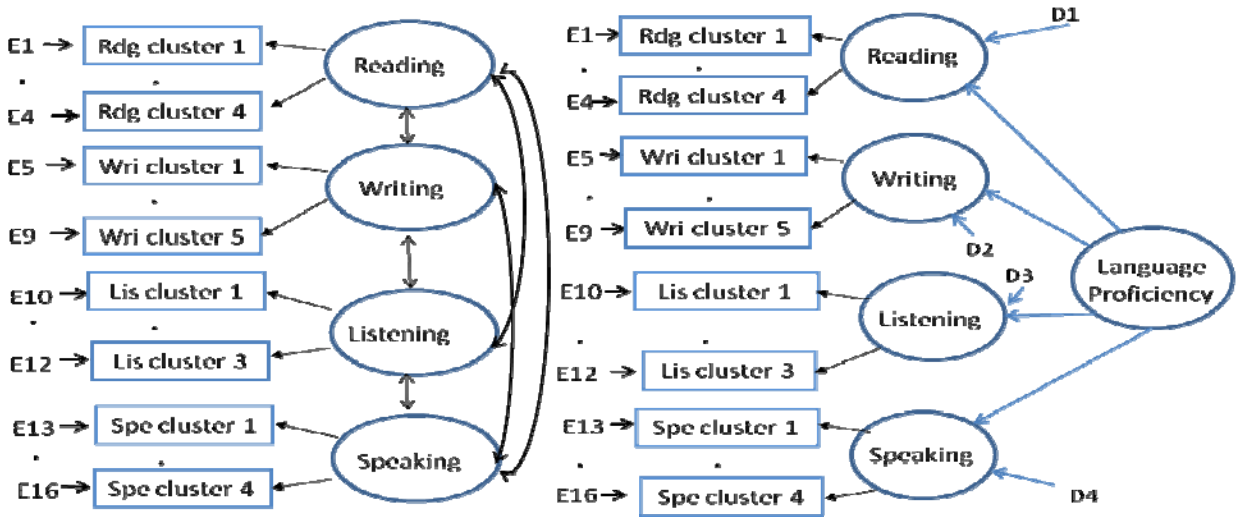


Figure 1. Four Models Tested

Sixteen variables were used in each analysis. Each variable (rectangular in each model) represents items clustered, based on test specifications. For instance, reading cluster 1 is reading comprehension and cluster 2 is reading analysis. Also, listening items measuring word/sentence comprehension were clustered as one observed variable, where items synthesizing information were combined as another variable. As a result, four reading variables, five writing variables, three listening variables, and four speaking variables were used in the analyses.

To assess how these models represented the data, absolute fit indices such as the Chi Square Statistic and the Goodness of Fit index (GFI), as well as incremental fit statistics, such as the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA), were used. Both GFI and CFI values greater than .95 constitute good fit, respectively. For the RMSEA, it has been suggested that values under .06 constitute good fit, values in the .05 and .08 range are acceptable fit, values between .08 and .10 range are marginal fit and values greater than .1 are poor fit (Hu & Bentler, 1999).

Result

Model 1 shows one factor called Language Proficiency, and all subtests are explained by the factor. In this model, all error variances are uncorrelated. The result showed poor fits. After using CFA results, Model 1 was revised. The big misfit in Model 1 was due to high correlation among error variances in each modality. Therefore, Model 2 was proposed, in which all error variances in each modality were correlated. Using the statistical information from Model 1 and content analysis, correlating errors in the subtest provided a reasonable explanation. For example, reading items are passage dependent. Each passage has items measuring main idea, summary, author's purpose, and so on. Therefore, different observed variables in reading include items from the same passage. Thus, unexplained variance by the Language Proficiency factor can be correlated because they are from the same passage. The results show acceptable fit with small RMSEA. Model 3 shows all four factors representing each modality. Although this model provided acceptable fit, Model 2 was significantly better fit than Model 3. Chi Square difference was examined to compare the four models. Model 2 produced a significantly better fit to the data than Models 3 and 4. The same result was found in elementary, middle, and high school. Table 2 shows the results in details.

Continued from previous page…

Table 3. *Model Fit Comparison in Four Models (Primary Level)*

Model	χ^2	df	GFI	CFI	RMSEA
1	16896.2	89	0.67	0.75	1.33
2	2445.4	142	0.95	0.97	0.05
3	5472.14	183	0.88	0.92	0.07
4	6192.9	185	0.87	0.91	0.08

No significant difference was found in test construct when augmented items were added. States developed augmented items for several reasons. One of the main reasons is because items are too easy for students in higher grades within each grade band using the SELP test items alone. Sometimes, items are too difficult for students in the lower grades. For example, in the primary level, writing two prompts is very difficult for kindergarteners and reading passages are too easy for second graders. Therefore, we developed easy writing items for kindergarteners and more challenging reading items for second graders. One of the main issues from adding augmented items was comparing test structure with the augmented items to the original SELP test structure. If there is a difference, we violate measurement assumptions, because we use item parameters calibrated on the original SELP test and adapted the vertical scale from the original test. The results showed that there is no significant difference between the models with and without augmented items, and therefore measurement assumptions are not violated.

References

Ackerman, T. A., Gierl, M, J., & Walker, C. M. (2003). An NCME instructional module on using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational measurement issues and practice*, 22, 37-52.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standard for Educational and Psychological Testing*. Washington, DC: American Education Research Association.

Bentler, P. M. (1995). EQS: Structural equations program manual. Encino, CA: Multivariate Software, Inc.

Harcourt Assessment (2006). *2006 WLPT-II Technical report*. San Antonio, TX: Author

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

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

Link to PowerPoint Presentation

-Yoonsun Lee,Ph.D., *is a Director of Assessment and Psychometrics at OSPI. She earned her Ph.D in measurement, statistics, and research design at University of Washington in 2004. Her primary research interests include Differential item Functioning (DIF), dimensionality, test validity, and equating. Contact information:* yoonsun.lee@k12.wa.us

Festschrift Papers (III)

Validity Issues for Common District Assessments

-By Jack B. Monpas-Huber, Ph.D

In recent years, many districts have implemented systems of common district assessments. By this I mean short assessments of students' skills that can be administered, scored, and reported quickly—hence the term "short-cycle" assessments. When administered under standardized conditions, these assessments can provide administrators with frequent estimates of students' status and/or growth toward the state proficiency standard. In these cases, such assessments serve a summative function much like a large-scale assessment such as the WASL. On the other hand, the short cycle of district assessments can provide teachers both timely feedback on instruction and identify students who may need additional help. These assessments can function as formative assessments in this way.

Both purposes of assessment are important. In most districts, accountability pressures create a need for summative data. Districts need to know what instructional programs and interventions are working most effectively. At the same time, districts value information that is timely and actionable to teachers. Further, assessments are expensive to develop and purchase for only one purpose. For all these reasons, it is probably common for districts to use their district assessments for both formative and summative purposes simultaneously.

In this paper, I critically examine these uses of district assessments from a measurement perspective. Using the example of district assessments in Spokane Public Schools, I argue that there are important validity issues behind these different uses of district assessments which districts should consider. In what follows, I first describe Spokane Public Schools' recent experience developing and using common district assessments. I then turn to the measurement literature to provide a brief review of various approaches to validation of districts' inferences and uses of assessments. I invoke Kane's (1992) conceptualization of an "argument-based" approach to validity to outline a strategy for validating district assessments. Using Kane's approach, I outline Spokane's argument for how it uses data from its common district assessments. I surface some of the implicit assumptions behind this argument and then attempt to sketch appropriate sources of validity evidence or validation strategies. In some cases, I can describe validity work that I have already done, while in other cases I describe challenges or barriers to validity work that should be done.

District Assessments in Spokane Public Schools

Several years ago, Spokane Public Schools embarked on a path of developing and implementing a centrally managed district curriculum and assessment system (English, 1988). In Spokane, this is called the "Written-Taught-Tested Curriculum," and its primary purpose is to bring curriculum, instruction, and assessment throughout the district into alignment with state standards. The managed curriculum provides a common district language and framework for instructional action for the district. To the extent that the curriculum contains the state standards, and teachers faithfully teach the district curriculum, the district can assume that all students at each grade level receive a common set of challenging educational experiences consistent with the state's expectations and will master the knowledge, skills, and abilities that will be manifest on the state assessment.

A cornerstone of this district curriculum policy and theory of action is common district assessment, which in theory provides several important pieces of information. They provide frequent reports of student achievement relative to the state standards and the curriculum—how well students are performing, how well *certain groups* of students are performing, *where* student performance is lower than desirable, and *on what state standards* performance seems to be lower than desirable. They also function as an "early warning system" for students who appear to need additional assistance.

Continued from previous page...

District coordinators in each content area have developed common district assessments in each content area for students at each grade level. The district assessments are relatively new and have evolved over the past few years to meet various district needs. These include both formative and summative purposes on which I elaborate in a forthcoming section.

Approaches to Validation

Arguably, all educators want valid and reliable data from assessments. Educators want to believe that they are drawing valid conclusions from assessment data, or that they are using assessment data in valid ways. But what does it mean to validate a district's use of assessments?

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 16). Evident from this definition, now shared widely among the measurement community, is that the object of validation is not an assessment instrument itself but the primary *inferences* and *uses* of the instrument (Messick, 1989; Test Standards, 1999). Arguably all uses of an assessment rest on inferences about what its scores mean, and not all inferences may enjoy the necessary empirical validity evidence or theoretical rationale. Some tests designed for formative purposes may not be able to support important summative purposes, and vice versa. Other tests intended to be measures of instructional effectiveness may actually not be very sensitive to instruction. Alternatively, claims that a series of local tests or other assessment results (such as grades) measure the same construct as a state assessment may turn out to be indefensible when confronted with disconfirming data.

There are multiple ways of conceptually organizing and embarking on the validation effort. One method is Messick's (1989) four-fold cross-classification of test inferences and uses by evidential and consequential bases. In the same work, Messick (1989) also outlined a variety of types of validity investigations:
We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to the items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test's external structure. We can investigate differences in these test processes and structures over time across groups and settings, and in response to experimental interventions—such as the instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and suing the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects. (p. 16)

Another method is to organize validation around the primary intended *uses* of a test (Shepard, 1993), foregrounding some validity questions and backgrounding others. One might also use the Test Standards (1999) as a guide.

In this paper, I use Kane's "argument-based" approach to organize thinking about validation of test inferences and uses. Kane (1992) suggests that validation might be framed most effectively in terms of a "practical argument" for particular inferences and uses which test users might advance for a test. According to this argument-based approach, test users first articulate as clearly as possible the major premises and claims of their argument for test meaning or use. They then surface the primary assumptions underlying these claims. Having articulated the argument and assumptions, users should then document the available validity evidence, or at least validation strategies, to support these inferences and uses.

The primary advantage of this approach is that it acknowledges the social context in which assessment operates. Educators make claims about student learning based on tests. They make arguments for using tests in some ways, but

Continued from previous page...

not others. They also question the validity of assessment data, based on different conceptions of evidence of student learning. An argument-based approach to validity has the potential to tailor validation to the particular frames of the public in question, such as the distinction between formative and summative uses of assessments which is meaningful to educators as well as somewhat analytically useful. I therefore use this distinction as the framework for validating Spokane's argument for its common district assessments.

Validating Formative Purposes of District Assessments

The purpose of formative assessment is to influence and inform instruction and action (Wiliam, 1998). Indeed, a major appeal of these systems to districts is their ability to provide content-rich assessment information to the district and classrooms on a regular basis in a way that informs next steps.

The district assessments in Spokane Public Schools have a strong formative orientation. By design, the assessments are relatively small (about 10-12 items) so that they can be administered quickly in classrooms, then quickly scored and reported district-wide so that teachers can make use of the information in a timely fashion. The district reports the results in terms of classical item difficulty statistics, disaggregated by school, so that teachers can see which skills have not been mastered by students and therefore where to focus additional instruction. By most accounts, this administration and reporting process is very effective with these assessments providing a valuable common language and metric of student achievement for school collaboration discussions. At the district level, content coordinators examine the data to identify areas of student need. Difficult items can point to areas for improved curriculum or professional development.

Wiliam (1998) suggests that formative and summative inferences of assessments should be validated in different ways. In his view, summative assessment should be validated on the basis of *score meaning* in the classical sense: What do the scores mean? How reliable are they? Do they behave as intended psychometrically? In contrast, the primary purpose of formative assessment is to stimulate and inform action: to inform next instructional steps, to guide instruction to areas of the domain where it is needed, to provide feedback to students, to motivate students to improve. For these reasons, formative assessments should be validated on the extent to which they produce the intended *consequences* (Wiliam, 1998).

Consequential validity evidence. What, then, counts as evidence of consequential validity? In Spokane, there is anecdotal evidence of the positive instructional consequences of the district assessments. The district assessments have helped teachers better understand what is expected of students at the state level, and as a result, teachers have become more consistent, aligned, and focused on the state standards in their instruction. Local data on what students know and are able to do have stimulated discussion among administrators, coaches, and teachers about instructional practice and pedagogy.

This is an important validity issue. Districts that invest in common district assessments for a strongly formative purpose should be clear about the intended instructional consequences of the assessments, and then gather evidence of the extent to which, and how, the assessments are then used to stimulate and inform ongoing midcourse corrections and differentiation in instruction.

Validating Summative Purposes of District Assessments

Arguably, consequential validity should not be only basis for validating common district assessments. Assessments lacking other important aspects of validity may produce fundamentally misleading information even if it is well-received and produces positive consequences. Evaluative pressures in districts may create needs for important summative inferences and uses of district assessments which may be designed for formative purposes. As schools become increasingly accountable for results, they will need to monitor progress of their students toward achievement goals. As far as these results are publicly reported, it will become important to ensure the technical quality of the testing process

Continued from previous page...

and data. Test results will need to be comparable despite year-to-year changes to items. Schools will need to show some measure of gain or growth for continuously enrolled students. It may also be important to have alternate forms of a test of comparable difficulty.

The district assessments in Spokane Public Schools have summative as well as formative purposes. Almost all of the district assessments are "end-of-unit" or "end-of-quarter" assessments administered at the end of instructional periods. Assessments are also administered under semi-standardized conditions (common instrument, predefined testing window, teacher scoring based on rubrics) in order to minimize variation due to administration factors and to facilitate comparison of the effectiveness of different instructional "conditions". More recently, the district has begun moving toward use of the summative district data as evidence of individual student achievement to populate a report card. This section articulates Spokane's argument for the functions of its district assessments as summative assessments of the written and taught curriculum and the validity issues and evidence surrounding them. In what follows, I use this argument as a framework for addressing inherent validity issues surrounding claims of this type and strategies for gathering validity evidence.

Inferences about Mastery of Content and Curriculum

All common district assessments are samples from a larger target domain of content knowledge, skills, and abilities. Most states make some effort to define this domain through frameworks and specification documents (Kolen & Brennan, 2004). Washington State provides a body of domain specification work at its Web site (www.k12.wa.us). This includes Grade Level Expectations (GLEs) documents which delineate exactly what students should know and be able to do at each grade level in each content area (OSPI, 2006). Teachers can internalize the GLEs to guide their own instructional planning, and districts can purchase or develop curriculum within which to embed these important learning objectives.

The "Written-Taught-Tested" Curriculum in Spokane Public Schools is a "theory of action" (Argyris & Schön, 1978) which makes strong claims about the content that its district assessments are measuring. The claim is that district assessments function as direct measures of the district's written curriculum which itself embodies the GLEs. The district's written curriculum takes the form of program guides that outline for teachers what content should be covered within a defined span of time, typically a unit lasting several weeks. These guides make clear the learning objectives—the GLEs embedded within the curriculum that will be taught if the curriculum is followed. The district assessments are designed to assess these learning objectives covered within curriculum units. A related claim is one of *alignment* between the district and state. Alignment to the state assessment system—for the district assessments to be "WASL-like"—is an important reason the district chose to develop its own curriculum and assessments rather than purchase these products from an outside vendor. The content argument is thus that the district assessments validly measure the content that students should know and be able to do within the framework of the curriculum.

These claims rest on various assumptions. One is that the district curriculum adequately captures the target domain of the state standards. Another is that the assessments adequately sample the domain of both the curriculum and the GLEs. The implication is that the district assessments are, to some degree, parallel measures of the state assessment, the WASL. Such claims should prompt a search for supporting evidence. Claims about the content of tests fall within the category of *content validity*. As Messick (1989) put it, "We can look at the content of the test in relation to the content of the domain of reference" (p. 16). Evidence of content validity can be documentation of test content and development procedures. All district assessments are developed according to the WASL item and test specifications (OSPI, 2007). All items in the district assessments are aligned (by expert judgment) to at least one GLE. All district assessments, like the WASL, have a mix of item formats: approximately half multiple-choice and half constructed-response items. The constructed-response items include short answer items (worth two points) and at least one extended response item (worth four points). All district assessments also include scoring guides to guide teachers in their scoring of student work. Arguably, these are good sources of evidence of content validity insofar as they make very clear the content of the assessments, the learning targets that the assessments are intended to measure, and the design and development of the assessments.

Continued from previous page...

Content validity evidence. Content validity is an important aspect of validity for medium-scale common district assessments. Districts that invest in common district assessments want to claim that the assessments are aligned to the state assessment system, that they are measuring the same content knowledge, skills and abilities that will be assessed on the WASL even if they are not strictly parallel forms of the WASL. Content validity may be an important issue especially in the marketplace of formative assessment systems. Outside vendors may claim that their products are aligned to state standards even if their items were not written according to state item specifications or their tests not developed according to state test specifications. This alignment may have been a *post hoc* process of aligning individual items to state standards through expert judgment. Spokane chose to develop its own common assessments specifically to build its own assessment capacity and to develop assessments specifically aligned to the Washington State standards using the WASL test and item specifications. To the extent possible, districts that go down the road of common assessments intended to prepare students for the state assessments should pay close attention to content validity.

Inferences about Student Proficiency, Constructs, and Traits

Claims from test results about what students know and are able to do in relation to a construct or trait is perhaps unavoidable. To ask any district assessment system to provide aggregate measurements of student *status* in relation to some predefined standard of proficiency is perhaps understandable. Such a standard can be the proficiency standard on the annual state assessment or a more proximal, locally-determined proficiency standard arrived at through some form of Angoff-based standard-setting procedure. In Spokane, inferences from test results about student knowledge, skills, and abilities are common. A *de facto* purpose of the district assessments is, as one coordinator put it, "to know where our kids are. WASL results should be no surprise." Another administrator said the purpose of the district assessments is to "fill the gaps between the WASLs." The construct argument is thus that the district assessments measure the same construct(s) as the WASL tests. Again, such claims about districtwide student abilities and achievements on the basis of district assessment data rest on assumptions which should be critically examined.

Correlational evidence. One form of construct validity evidence is correlations between scores from tests believed to be measuring the same construct(s)—what Messick (1989) refers to as the "external structure" of a test. Stronger correlations represent stronger evidence of parallelism and alignment between two tests purported to measure the same or at least very similar constructs. Correlations between scores from district assessments and WASL tests will likely be moderate, which gives rise to several interpretations. First, all correlations below 1.0 provide an opportunity to better understand the concept of measurement error and to temper hopes of perfect prediction. Second, moderate correlations suggest that although the tests share considerable variation, they are measuring somewhat different constructs (Kolen & Brennan, 2004). This makes sense when we consider the nature of the two constructs being measured. Both the WASL and the district assessments are samples from a very large domain. Being a larger test, the WASL represents a larger sample that uses more items to measure the domain. The district assessments measure only the GLEs embedded within curriculum units. Thus, they measure a smaller, more defined domain than the WASL, and they are smaller assessments in which a small number of items are used to measure as many GLEs as possible. As a result, the correlation between the WASL and the district assessments is attenuated by construct underrepresentation (because the construct measured by the district assessments is smaller and more constrained) and restricted range (because the district assessments cannot measure the full range of the construct measured by the WASL).

Convergent/divergent validity evidence. Another construct validation strategy for district assessments would be to explore the convergent and divergent validity through the use of the multitrait multimethod matrix (Campbell & Fiske, 1959; Crehan, 2001). Crehan (2001) used such an approach to examine the convergent and divergent validity of data from a district-developed performance assessment and found limited evidence of validity. Unsettling is the implication that the data provided by those assessments provided misleading information for decisionmaking in that district.

Continued from previous page…

Reliability evidence. Claims about performance on items or tests about what students know or are able to do also rest on the implicit assumption that scores from the sampled items and assessments can be generalized to the target domain without error or bias. The extent to which test scores possess this property is commonly known as *reliability*. The Test Standards (1999) are clear that inferences about student ability on the basis of scores of an educational assessment require some evidence of the reliability of scores from the assessment. Reliability is thus an important issue for district assessments.

In districts where students take a district assessment only once, it will not be possible to estimate reliability by means of test-retest or alternate forms analyses. Instead, one can use measures of internal consistency reliability such as Cronbach's coefficient alpha (Cronbach, 2004). The alpha coefficient is a measure of the extent to which a test is unidimensional based on covariation among items. Strong covariation between items and comparatively small amounts of individual item variation represent evidence that the items collectively are measuring one construct or trait (DeVellis, 2003). A good Cronbach's alpha value is .80 – .90, with .70 being a minimally acceptable value.

Estimates of reliability, like correlations, prompt a search for sources of unreliability, or measurement error. As described above, most district assessments are necessarily *short* (about 10 items) so that they can be administered and scored quickly. However, reliability is generally understood to increase with the number of items or tasks and the average inter-item correlation (DeVellis, 2004). In addition, items are typically written to measure different skills (GLEs). Thus, district tests may be multidimensional, rather than unidimensional, by design. Thus, test size and multidimensionality by design may place a ceiling on internal consistency reliability. Another potential source of error is inter-rater disagreement in the scoring of the open-ended items. Anecdotal evidence of variation in teachers' application of the scoring rubrics for their students' open-ended responses abounds. However, in my observations, open-ended items typically enjoy the strongest item-total correlations.

This finding suggests that open-ended items do a reasonably good job of discriminating examinees on the basis of achievement (contributing true variation) despite any inter-rater disagreement (error variation) that may exist.

Reliability estimates carry implications for summative inferences about student's level of achievement. Low reliability estimates may suggest that a large proportion of the observed variation in the scores is due to random error, or *noise*—or at least variation due to individual items that is unrelated to the primary trait being measured. In individual terms, this means that a student's observed total test score may lay at some variance from his or her true test score. In other words, low reliability produces unstable test scores. This becomes a problem when districts begin to make important decisions about students on the basis of these test scores, such as the assignment of a summative grade that will be reported publicly and become part of the student's permanent record. Low reliability will produce misclassifications. Students with test scores that overestimate their true achievement will receive higher grades, and students with test scores that underestimate their true achievement will receive lower grades. In addition, low reliability limits correlations with other measures (Carmines & Zeller, 1979). Some students who receive high marks in a content area based on district test scores will score below standard on the state assessment, and vice versa. Such results would be cause to temper strong claims about alignment with the state system.

New measurement research offers new ways of thinking about internal consistency indicated by Cronbach's alpha. Willett (1988) suggests that correlation-based reliability estimates might obscure important dynamics of student growth. Reliability describes the extent to which two measures produce the same rank ordering of examinees. However, that can be misleading when examinees are growing (both in negative and positive directions). By this thinking, low reliability estimates may be evidence of considerable intra-individual growth. Cronbach (2004) himself suggested that the alpha coefficient is less appropriate for the kind of mixed-format performance assessments currently in use today which are multidimensional by design. Marzano (2000) makes a similar argument in his application of measurement theory to

Continued from previous page...

formative classroom-level assessment. He suggests that most teachers rarely design classroom assessments to measure only one trait or construct. Thus, internal consistency may not be the most useful way to think about the reliability of these assessments.

These are important issues for district assessments. Consumers of common district assessments at all levels want to claim that their tests validly and reliably measure the psychological construct(s) they are intended to measure. Such claims may be tenuous without the kinds of evidence of reliability and construct validity outlined above. Minimal or weak evidence of validity and reliability in these areas may be cause to temper such claims.

Inferences about Growth in Student Achievement

Related to the issue of student performance status is *growth* in student achievement. Districts that invest in common district assessments, especially when those assessments measure the same students in a content area several times a year, might reasonably desire some form of information about growth in achievement. At the district or school level, pressure to improve performance may place a premium on data that could show students growing toward proficiency. Many administrators would like to use some measure of growth to more rigorously evaluate the effectiveness of various instructional programs and treatments (Lissitz, 2006), while many teachers would like to see growth over time in their students' learning—especially in regard to the state proficiency standards—as a result of instruction. Anecdotal evidence of these desires abounds. The appeal is understandable. Students in a grade level are assessed on their proficiency in a content area several times a year.

It is tempting to ask: Where are the students in relation to the state standards? How many are on track to pass the upcoming state assessment?

Fortunately, these desires for growth data come at a time when the supply of expertise and knowledge in this area is growing. An explosion of empirical work is currently happening in the area of growth research and longitudinal data analysis (Lissitz, 2006; Lloyd, 2007; Singer & Willett, 2003), and this work carries powerful implications for schools' efforts to measure and determine what can be done to cause all students to reach state standards. One immediate and important implication might be to provide educators with a more precise understanding of growth and the technical requirements for valid growth inferences. Measurement researchers restrict their use of the term "growth" to data that meet two important specifications: (a) the same examinees are observed on the same construct on repeated occasions; and (b) all measurements of examinees fall along a continuous score scale (Kolen & Brennan, 2004; Lissitz, 2006; Singer & Willett, 2003; Willett, 1988). In what follows, I discuss each of these requirements and efforts to provide valid growth inferences in the context of Spokane Public Schools' common district assessments.

The same examinees are observed on repeated occasions. A significant threat to observing the same examinees over time is mobility. Student mobility is a serious challenge in many schools. In Spokane, students move around considerably, both within and to beyond the district. In recent years, the district has not had a mechanism for collecting data for every student in the district, so it has used sampling methods. The district has collected district assessment data by means of independent random samples by which each of 100 teachers receives a new random list of students whose work is requested for central data collection. While this design had the advantage of a producing a large representative district-wide sample (stratified by classroom), it did not provide repeated measurements for the same examinees. One solution to this problem is to centrally select one district-wide random sample at the beginning of the year and follow it as a *panel* over time. This may not be a serious challenge in districts that have the resources or data collection mechanism to collect data from every student.

Continued from previous page...

All measurements of examinees fall along a continuous score scale (Kolen & Brennan, 2004; Lissitz, 2006; Singer & Willett, 2003; Willett, 1988). This is an important measurement issue which is not always well understood in education where growth has been defined broadly enough to include different kinds of analyses.

Consider a personal example: My five-year-old son stands up against a wall and I measure his height as 40 inches. Several months later, we repeat the procedure and find that his height now exceeds 40 inches. That is literally true growth in height along one continuous scale that spans the entire range of the construct of height. The implication for educational assessments would therefore be tests of varying difficulty, given at different ages, sharing one continuous scale. This is called a *vertical scale* which makes possible inferences about *absolute* growth.

Now consider a common alternative: My son is a student in a fifth grade class which receives a month of math instruction and then takes an end-of-unit assessment of the math content covered within the unit. My son scores at the mean of the distribution. His class then receives another month of instruction and then takes a second end-of-unit assessment of equal size and format to the first but which covers somewhat different material. My son scores at the 85th percentile. Can we reasonably infer that he has "grown" in math achievement? The two tests represent different rulers of achievement. We can much more validly infer that my son has grown *relative* to his classmates than we can infer that he has grown in math skill in any absolute way along any hypothetical continuous scale.

From the perspective of growth inferences, the district assessments in Spokane conform more to the latter example than to the former. Within a content area, the district assessments are discrete measurement points that capture what students are expected to have learned within the latest curriculum unit. Each unit, especially in mathematics, may focus on different content. Each assessment within a content area also has a slightly different number of total possible raw points. As a result, the assessments are like different rulers of different length and calibration.

However, it may be still be possible to "link" these assessments together on a continuous scale of proficiency in a content area. There are two issues. One is the construction of the continuous scale of achievement based on the district assessments. Once the same students in a grade level are observed on each of the district assessments, it may be possible to use IRT scaling techniques with single-group common person equating (Bond & Fox, 2007; Yu & Popp, 2005) to determine the extent to which the items from the different district assessments form a continuous scale of content area achievement. Such a common score scale for the district assessments could provide a foundation for educators to observe growth in the same students' content area achievement over the course of one year.

The other issue is to locate the WASL state proficiency standard on this locally developed continuous scale. What does a particular score on a district assessment, or some combination of performances on the district assessments, mean in relation to scores on the WASL? Is the WASL a more difficult assessment than the district assessments? An easier assessment? Or about the same? If WASL results included item parameters, it might be possible to use a single-group common person concurrent calibration strategy to construct a scale that includes the state proficiency standard. Then the district assessments would be on the same scale as the WASL and would provide better information throughout the year about where students are in relation to the state standard. However, the WASL results do not include item parameters, and so a weaker alternative may be to employ some kind of linking strategy (Kolen & Brennan, 2004) based on only total scores rather than items.

Growth research could have considerable practical implications for curriculum and instruction. If it were possible to give the actual WASL test on the first day of school, what would the distribution look like? How many students would already be at standard? How many students would be close, and farther away? What assumptions would be challenged by this result? How well does a linear model really represent the learning process? The ability to measure growth more precisely through the use of a continuous scale could stimulate considerable discussion about teaching and learning.

Continued from previous page...

Discussion

The purpose of this paper was to raise a series of validity issues for common district assessments whose use is increasing in public education. Worth repeating is that not all uses and inferences of common district assessments may be valid. However, complicating the matter, as this paper has tried to suggest, is the mix of conflicting pressures for both formative and summative purposes that converge at the district level. Clearly, small, frequent tests should be able to provide useful information to teachers. However, small tests developed for formative purposes may not be able to support important summative inferences and uses. To complicate matters, purposes of assessments may evolve over time in response to changing organizational priorities. Possibly district assessments themselves, like the state assessment, may evolve over time in response to changing needs.

Districts that choose to invest in common district assessments would do well to think through some of these validity issues and consider gathering appropriate validity evidence.

References

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

Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley Publishing Company.

Bond, T.G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Second Edition. Mahwah, NJ: Lawrence Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage Publications.

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational and Psychological Measurement*, 61(5), 841-848.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. (CST Technical Report 643). Los Angeles, CA: Center for the Study of Evaluation.

DeVellis, R. F. (2003). *Scale development: theory and applications*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

English, F. (1988). *Curriculum auditing*. Lancaster, PA: Technomic Publishing Co., Inc.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Second Edition. Springer.

Lissitz, R. (Ed.). (2006). *Longitudinal and value-added models for student performance*. Maple Grove, MN: JAM Press.

Lloyd, J. E. V. (2007). On the quantitative analysis of individual change: Unpacking the meaning of "change" and "commensurability". Manuscript submitted for publication.

Marzano, R. J. (2007). Applying the theory on measurement of change to formative classroom assessment. Retrieved November 10, 2007, from http://www.marzanoandassociates.com/html/resources.htm#papers

Marzano, R. J. (2000). Analyzing two assumptions underlying the scoring of classroom assessments. Retrieved November 10, 2007, from http://www.marzanoandassociates.com/html/resources.htm#papers

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Continued from previous page...

Office of the Superintendent of Public Instruction. (2007). *Test and item specifications for the Washington Assessment of Student Learning (WASL)*. Retrieved on November 17, 2007 from http://www.k12.wa.us/assessment/WASL/testspec.aspx.

Office of the Superintendent of Public Instruction. (2006, September). *Mathematics K-10 grade level expectations: A new level of specificity*.

Popham, J. (1987). Measurement-driven instruction. *Phi Delta Kappan*.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1993). Sampling variability of performance assessments. (CSE Technical Report 361). Los Angeles: Center for the Study of Evaluation.

Singer, J. S., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Wiliam, D. (1998). The validity of teachers' assessments. Paper presented Working Group 6 (Research on the Psychology of Mathematics Teacher Development) of the 22nd annual conference of the International Group for the Psychology of Mathematics Education, Stellenbosch, South Africa.

Willett, J. B. (1988). Questions and answers about the measurement of change. In E. Rothkopf (Ed.), *Review of research in education* (1988-89) (pp. 345-422). Washington, DC: American Educational Research Association.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yu, C. H., & Popp, S. E. O. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research & Evaluation* 10(4). Retrieved on November 18, 2007 from http://pareonline.net/getvn.asp?v=10&n=4

-Jack B. Monpas-Huber, *Assessment & Program Evaluation, Spokane Public Schools.*
Please direct correspondence to: Jack B. Monpas-Huber, Ph.D., Director of Assessment and Program Evaluation, Spokane Public Schools, 200 North Bernard Street, Spokane, Washington, 99201. E-mail: JackM@spokaneschools.org

Festschrift Papers (IV)

Academic Growth in Math Students in Poverty –WASL and MAP

–By Feng-Yi Hung, Ph.D

The need to improve the educational outcomes of students in poverty is urgent. Concentrated poverty, family instability, homeless situations, and military deployment are but a few hardships typical of growing up in Clover Park School District – Lakewood, Washington.

The 2006-07 school year marked the first opportunity for Clover Park School District students (middle schools) to participate in the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessments. MAP is a computerized adaptive assessment which enables teachers to assess and compare individual student growth in achievement with the growth of over 3 million students throughout the United States. This type of growth data is critical when we respond to the needs of students in poverty and evaluate their progress or lack of progress over time.

Research Questions:

- How will WASL and MAP assessment tell us the story of "progress" and closing the achievement gap for students in poverty?
- What are the significant factors impacting the achievement of students in poverty as measured by WASL and MAP?

Methodology

We selected a group of this year's 8th graders who met the following criteria –

1. Eligible for Free/ Reduced Price Lunch (FRL)
2. Participated in 2006 WASL (6th grade) and 2007 WASL (7th grade)
3. Participated in MAP Fall and Spring Assessments in the 2006-2007 school year

Two hundred and ten students were selected for this study. WASL scale score and MAP RIT scores are used to analyze student performance and growth during the 2006-07 school year.

WASL and MAP Results

Our results show that 71 (34%) students in poverty met standard in 2007 WASL Grade 7 Math (Table 1). The percentage of students meeting standard in this study is similar to the district-wide average (37%). MAP RIT score range for students not meeting standard in WASL almost double the score range of students who met standard (72 vs. 38).

MAP RIT mean score for students not meeting standard is 214 and for students meeting standard the RIT mean score is 238. Based on the NWEA math achievement and normative data, 214 is grades 4-5 instructional level and 238 is about grade 10 level. In other words, the FRL students who did not meet standard in WASL are at least two grade levels behind in math; in addition, if the FRL students who met the state math standard, they are approximately three grade levels ahead of their counterparts as measured by MAP. This highlights the spectrum of learning needs of students in poverty and, at the same time, the challenge for them to meet the state standards in math

Continued on next page…

Continued from previous page...

Table 1. 7th-Grade WASL and MAP

Descriptive Statistics

	WASL - Met Standard	WASL - Not Met Standard
Number of Students	71 (34%)	139 (66%)
RIT Range	38 (218-256)	72 (172-244)
RIT Mean	**238**	**214**
National Norm	**Grade 10**	**Grades 4-5**

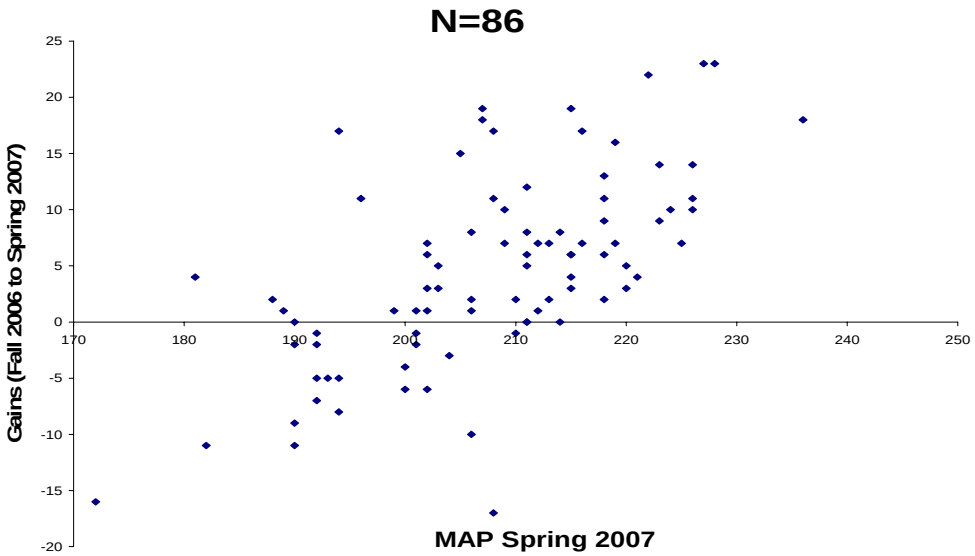
The NWEA 2005 Norms Study includes results from more than 2.3 million students in 794 school districts, representing 32 states.

Past research has indicated that students from low-income schools are concentrated at the low end of continuum of student achievement. Research also indicates that Hispanic students were the only group that evidenced consistently higher rates of change ("growth") than non-minority students (McCall, Houser, Cronin, Kingsbury, and Houser, 2006). We examined MAP growth distribution for FRL students who scored in WASL level 1 in both 2006 and 2007 testing administrations. These students were substantially below the state standard in two consecutive years of WASL administrations. Out of 210, 86 students were identified (41%).

Our results showed –

1. These "struggling" students had the average of 4.6 RIT point gains as measured by MAP assessments (Figure 1). Their RIT score growth ranges from -17 RIT points to +23 RIT points. This gain is slightly less than NWEA growth average for students with the same initial RIT point (203, Growth Mean: 5.8) at the 7th grade level.

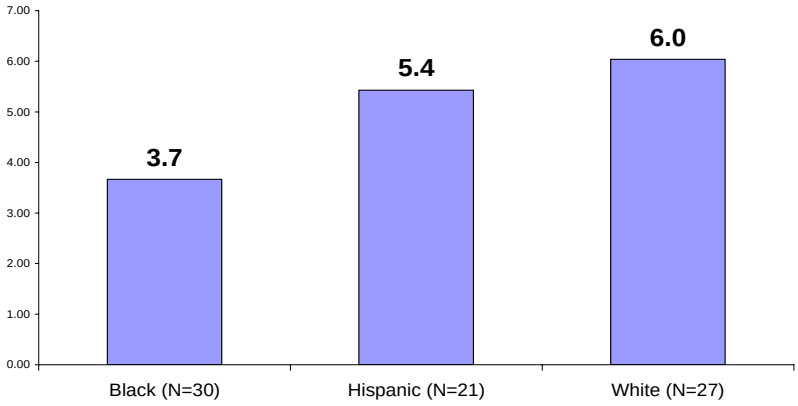
Figure 1. WASL Math Level 1 - Two Consecutive Years (Grade 6, 2006 and Grade 7, 2007)



Continued from previous page...

2. Within the same poverty level and WASL performance, rates of growth differ by ethnicity (Table 2). White and Hispanic students made more gains than Black students. Due to the small number of American Indian and Asian students (fewer than 10), their results are not included.

Table 2. WASL Math Level 1 - Two Consecutive Years
MAP Fall to Spring Growth by Ethnicity



We further examined how group membership factors such as schools, gender, and ethnicity impact the achievement and growth of FRL students. Univariate Analysis of Variance (ANOVA) analysis was conducted. Attending schools is not a significant factor in terms of FRL students' performance and growth, so gender and ethnicity are included in the final analysis.

Table 3 – We used Grade 7 WASL math scale score as the dependent variable and included ethnicity, gender, and the interaction term of ethnicity and gender in the model. The overall F test is significant ($F=3.27$, $p<0.01$). Ethnicity is statistically significant ($F=5.1$, $p<0.01$); however, gender and the interaction term are not significant.

Table 3. 7th-Grade WASL Math

Univariate Analysis of Variance (ANOVA)

	Mean Scale Score	N
Asian/Pacific Islander	389	26
Black	360	50
Hispanic	363	42
White	387	84
Total	376	210

Design: Ethnicity + Gender + Ethnicity*Gender

Results: Significant Effect on Ethnicity, F=5.1, p < .05

Post Hoc Tests - Tukey HSD: Asian vs. Black; White vs. Black; White vs. Hispanic

Table 4 – We used Grade 7 Spring MAP RIT score as the dependent variable and included ethnicity, gender, and the interaction term of ethnicity and gender in the model. The overall F test is significant ($F=2.83$, $p<0.01$). Ethnicity is statistically significant ($F=4.1$, $p<0.01$); however, gender and the interaction term are not significant.

Continued from previous page...

Table 4. 7th- Grade MAP Math

Univariate Analysis of Variance (ANOVA)

	Mean RIT Score	N
Asian/Pacific Islander	225	26
Black	216	50
Hispanic	217	42
White	226	84
Total	222	210

Design: Ethnicity + Gender + Ethnicity*Gender

Results: Significant Effect on Ethnicity, F=4.1, p < .05

Post Hoc Tests - Tukey HSD: White vs. Black & White vs. Hispanic

Table 5 – We used MAP Fall to Spring growth as the dependent variable and included ethnicity, gender, and the interaction term of ethnicity and gender in the model. The overall F test, ethnicity, gender or the interaction between these two variables are not statistically significant. Descriptive statistics showed interesting growth results for different ethnicity groups within FRL student population. White students had made the most gains (6.4 RIT points), followed by Hispanic (5.5 RIT points) and Black (4.5 RIT points) students. Asian students have made less than 50% of RIT points gains (2.7 RIT points), compared to their White and Hispanic counterparts. Due to the small size of Asian FRL students in this study, results should be treated cautiously.

Table 5. 7th- Grade MAP Math Gains Fall to Spring

Univariate Analysis of Variance (ANOVA)

	Gains: RIT Score	N
Asian/Pacific Islander	2.7	26
Black	4.6	50
Hispanic	5.5	42
White	6.4	84
Total	5.4	210

Design: Ethnicity + Gender + Ethnicity*Gender

Results: No Significant Effect

Discussion

Students in poverty are faced with multiple challenges: poverty, violence, victimization, family instability, and the perils of negative stereotyping. The impact of these social conditions, hardships, and stereotypes can extend into the actual classroom setting and is evident in student achievement.

The results of this study bring us one step closer to understanding the achievement gap in math as measured by WASL and MAP assessments. Improving student learning in math is a highly complex challenge, especially for students in

Continued from previous page…

poverty. The finding that schools and gender are not significant factors in impacting FRL student performance and growth over time is important. At the same time, it highlights the sensitive and critical issue of ethnicity for students in poverty. Consistent with NWEA research, White and Hispanic FRL students made more academic growth in math than their counterparts. Although these differences are not statistically significant, it provides meaningful information in terms of instructional support for diverse learners in the high-poverty school setting.

This study also shows that one assessment result will not provide the full picture of student learning and progress. Students in poverty need assessment that shows both, how they have mastered the state learning targets and how far they have come to achieve the standards.

References

Dahlin, M. P. (2007). A Study of the Alignment of the NWEA RIT Scale with the Washington Assessment System. Northwest Evaluation Association. Lake Oswego, OR.

McCall, M., Hauser, C., Cronin, J., Kingsbury, G., & Houser, R. (2006). Achievement Gaps: An Examination of Differences in Student Achievement and Growth. Northwest Evaluation Association. Lake Oswego, OR.

NWEA (2005). Normative Data: Monitoring Growth in student Achievement. Lake Oswego, OR.

McKinney, S., Frazier, W., & Abrams, L. Responding to the Needs of At-Risk Students in Poverty. From http://www.usca.edu/essays/vol172006/mckinney.pdf.

OSPI. (2002). Addressing the Achievement Gap: A Challenge for Washington State Educators. Olympia, WA.

OSPI. (2007). Teaching Math in Washington's High Schools: Insights from a Survey of Teachers in High Performing or Improving Schools. Olympia, WA.

Link to PowerPoint Presentation

-Feng-Yi Hung, Ph.D. *is Director of Assessment and Evaluation in Clover Park School District. Contact information: fhung@cloverpark.k12.wa.us*

Q& A on Logic Modeling

(Prepared for OPEN September 2007)
Kari Greene, a program administrator/evaluator in Oregon's Public Health Division Program Design & Evaluation Services, prepared this Question and Answer article for the Oregon Program Evaluation Network (OPEN) fall 2007 newsletter.

Q: What is logic modeling?
A: Logic modeling is the process of creating a visual representation of a program or intervention. This process – broken down to the most basic level – addresses the questions, "What are you doing? Who are you doing it with? What resources do you have to do it? And what are the effects of what you're doing?" Relationships are drawn between these questions, helping identify the theory or logic behind the program inputs, elements and the intended outcomes.

Q: Why do you think it's so important?
A: I think many of us – whether we're evaluators, program managers, researchers or funders – don't take time to sit back and look at the big picture in a reflective, thoughtful way. Instead, we're busy getting the work done related to the program. Yet we fail to realize that while we might be working very hard, we might not be achieving what we set out to achieve. Logic modeling can help identify why a program is not reaching the intended goals or what is missing in order to achieve the expected outcomes.

By gathering together the main stakeholders on a program, logic modeling facilitates focused and productive communication about the program. At the end of a logic modeling session, I've heard people say, "this is the first time we've actually discussed what 'success' is for our program and what we think we're doing all day." That's why I refer to this as logic model*ing* – which emphasizes the process – instead of focusing solely on the logic model as a product.

Q: What's the most common mistake people make regarding logic modeling (if any)?
A: People often get hung up on creating the perfect logic model that's "right" but they don't see that it's usually at the sacrifice of a meaningful process. To me, the logic model itself is not as important as the process of logic modeling. I think that people sitting down together, working as a group to identify a shared vision and the key programmatic components and outcomes is much more important than a pretty picture that may or may not accurately describe the program. And a logic model will never be "right" because it will change and shift depending upon who is helping to create it, what that group's shared vision is, and the motivation behind creating it at that time.

Q: Why should program evaluators understand logic modeling (how does it link to the evaluation)?
A: The first step in most evaluations is program definition – logic modeling is an easy, valuable process to use when defining a program to evaluate. Evaluators can use logic modeling as a planning and communication tool throughout the evaluation process. It's possible to have multiple logic models for a single program "nested" on top of each other and each used for a different purpose. One might be to uncover the program theory while another can be used to detail the evaluation design and data collection process. In fact, a direct link is evident between many of the Program Evaluation Standards (as identified by the Joint Committee on Standards) and logic modeling, from Stakeholder Identification to Context Analysis.

Q: What are some great books on logic modeling?
A: OK, I'm kind of dorky and overly enthusiastic about logic modeling, but even I don't sit around reading books on logic modeling!!! However, I always direct people to the Kellogg Foundation's Logic Model Development Guide and I have a resource page that I hand out in my logic modeling presentations that lists helpful websites and reports so it's not so overwhelming for people new to logic modeling. And actually, there's a book that I haven't read but have heard is good by Joy Frechtling called "Logic Modeling Methods in Program Evaluation" so someone could read it & give us a little book report in our next OPEN newsletter!

Continued from previous page...

Q: Any final thoughts on the subject?

A: I guess I'm reminded that even when I'm starting an evaluation project on little time, little money and little sleep, I still find logic modeling helpful to orient myself to the project. It doesn't have to be a big deal and it's a kind of logical thinking process that comes very naturally to most evaluators. I just need to remind myself not to make it into a bigger deal than it needs to be...

References

--, 2004. *Logic Model Development Guide*. Battle Creek, MI: W. K. Kellogg Foundation.
This document is available online at http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf.

Frechtling, J.A. (2007). *Logic Modeling Methods in Program Evaluation*. San Francisco, CA: Jossey-Bass.

Oregon Program Evaluators Network (OPEN) is an organization for residents of Oregon and Washington who are involved with or interested in program evaluation. OPEN allows for the exchange of ideas and information to promote and encourage high-quality evaluation practices. The members of OPEN represent government agencies, universities, and private consulting firms. Web address is http://www.oregoneval.org.

-Kari Greene, MPH, *is active in Oregon public health program evaluation circles and is a long-time OPEN member. She presents frequently on the nuts and bolts of program evaluation.*

WERA/OSPI Winter Conference 2007 lunch time conference goers.

SPSS Tips and Tricks and Beyond: For Beginning and Intermediate Users

-By Andrea Meld, Ph.D.

Introduction

SPSS, then know as the "Statistical Package for the Social Sciences," was first developed back in the days of punch cards and mainframe computers. Computers came in a very limited color scheme: white, green or amber on black. One small error in your inches-thick stack of cards and it was literally back to the drawing board. Then again, you could only have 80 columns of data, which also limited a data set to 80 variables. In my computer applications class at the University of Washington, we learned FORTRAN before venturing into the brave new world of SPSS commands. I read instructions from a maroon SPSS manual, about the size of a phone book, which also helped me learn statistics, but there was no such thing as on-line help.

Today's SPSS is installed on a desktop or laptop computer, with a dizzying assortment of drop-down menus and point-and-click commands. The number of variables can exceed 500, for example, in state-wide assessment data sets. Help is available in a number of different ways.

Syntax has many advantages over menu commands, which will be explored in this article. I use SPSS almost daily in my work, but I am still expanding my syntax-writing abilities. For the new and intermediate user, SPSS may loom as an entirely unexplored universe, especially when it comes to preparing and running syntax. It really helps to discuss things with other SPSS users, thus the motto:

Share what you know, learn what you don't.
http://www.spsstools.net/index.html#Share

Raynald Levesque, *Raynald's SPSS Tools*
http://www.spsstools.net/

What is SPSS?

SPSS is a software system for data analysis and management. You can use SPSS to import data from various types of files and create customized tables, reports, and graphs. SPSS can also be used to produce frequency distributions and data trends, descriptive statistics, such as mean, median, mode and standard deviation, and complex statistical analysis, such as analysis of variance and regression analysis. It is used in over 100 countries and in all 50 U.S. state governments, according to SPSS, Inc., as well as by universities, some of the largest U.S. newspapers, and other business enterprises, where it is often used to analyze survey and marketing data.

What does SPSS stand for?

When SPSS was founded in 1968, the developers named it "Statistical Package for the Social Sciences," or "SPSS." At first, it was used mostly in colleges and universities. SPSS is now widely used in business and other settings for various types of analysis. Today, "SPSS Inc." refers to the company and "SPSS" to the product.

Syntax vs. Drop-down menus (Point-and-Click).

Most SPSS procedures can be done either through syntax or drop-down menus. Using syntax has many advantages. Here are some comparisons of these methods:

Continued from previous page…

1. Drop-down menus are easier to learn initially.
2. They may be adequate for every day or one-time use.
3. Undetected errors are more likely using menus.
4. Results can easily be reproduced using syntax, which is useful for work that is repeated or done on a regular basis, such as annually.
5. Some procedures and manipulations are only available through syntax.
6. Syntax allows you to document your work.
7. Syntax can be big time saver and enhance productivity and efficiency.
8. It works better for complex data management and lengthy analysis.
9. It allows you can communicate effectively with other SPSS users, so it can serve as a universal language, although SPSS is available in different languages.

Tips on Learning Syntax:

1. Study existing syntax created by other users.
2. Share successful syntax with co-workers and friends.
3. You can cut and paste menu commands and start using these for syntax – although it may require some tweaking.
4. Save syntax that works – keep a log.
5. Refer to books and websites.
6. Although SPSS is not case sensitive, if you capitalize SPSS commands and write variable names in lower case, syntax is easier to read.
7. Put spaces before and after slashes (/) and apostrophes (‘) to enhance readability.
8. Save your syntax file frequently – SPSS does not save syntax automatically.

Steps in Using Syntax:

Using syntax involves an iterative process, as follows:

1. Record what you want the program to do. For example, run FREQUENCIES for a reading test that was administered from 2004 to 2007 at grade levels 3, 4, and 5. This can go in the COMMENTS line.
2. Write the syntax – either in the syntax file or editor.
3. Run your syntax.
4. Check for errors in your output.
5. Try to fix the syntax.
6. Run your program again.
7. Repeat steps 4 – 6 until your output is error-free.
8. Debugging errors can take longer than writing your program.

Avoiding the Most Common Syntax Errors:

1. Make sure your file and path names are correct.
2. Avoid typing letter **O** instead of numeric **0** and letter **l** for numeric **1**. Find and correct all typos and spelling errors.
3. Avoid long variable names – it’s easier to make typos. Use variable labels and value labels to describe your data.
4. Be sure to start a comment line with an asterisk ***** and end with a period**.**
5. Indent the second line of each command statement.

Continued from previous page...

6. Close and balance parentheses ( ) and nested parentheses.
7. Use apostrophes ' ' or quotation marks " " to enclose string variables.
8. Don't wipe out your data with SELECT IF statements. This can happen if your selection criteria are misspelled, for example, or are logically impossible.
9. Close command statements with a period, but don't use a period just because you've come to the end of a line.
10. Most commands can be abbreviated to 3 or 4 letters, for example, FREQ for FREQUENCIES. However, COMPUTE cannot be shortened to COMP.

Further Debugging

1. Become familiar with your data. Run FREQUENCIES and CROSSTABS.
2. Check your output, and if your results seem too strange to be true, look at your data and syntax and check for errors.
3. Use the DESCRIPTIVES command to make sure that the minimum and maximum data values are within the expected range.
4. Compare mean scores for different groups. Are the results what you would expect?
5. Use IF statements to find contradictions, for example, if students can't be both absent and tested, you can use this syntax to flag cases:

```
IF absent = 1 and tested = 1          FLAG     abtest = 1.
 FREQ abtest.
```

Steps in Using SPSS with Some Examples:

Read data Translate raw data or data in another form into SPSS and SAVE.

```
GET DATA / type = txt
        /FILE = 'C:\test\newdata.dat.'
     SAVE OUTFILE = 'C:\test\newdata.sav.'
```

Open an SPSS file

```
GET FILE = 'C:\test\newdata.sav.'
```

Define variables Put labels onto variables and values so that SPSS knows how to read it properly, and it will make sense to other users.. For example:

```
    VARIABLE FORMATS name 1-26 A25   grade 26-27 F2.0.

VARIABLE LABELS ennrol07   '2007 school enrollment'.

VALUE LABELS          gender 1 'male' 2 'female'.

MISSING VALUES       exam1 to exam3 (99).
```

Transform data Create new variables or change the values of existing variables.

Continued from previous page...

For example, you can use SPSS to:

RECODE
COMPUTE
EXECUTE
SORT
IF (THEN)
DO IF
SELECT IF, ETC.

```
RECODE age (10 thru 20 = 1) (21 thru 30 = 2).
COMPUTE total = exam1 + exam2 + exam3.
SELECT IF (speced = 'Y').
```

Create Tables, for example:

FREQUENCIES Run to get score distribution:

```
FREQUENCIES VARIABLES=mattempt
      /ORDER= ANALYSIS .
```

CROSSTABS
MEANS TEST
ANOVA
REGRESSION
MULTIPLE REGRESSIONS

Save Your SPSS File

There are several options when it comes to saving your SPSS file. You may want to DROP or KEEP certain variables. You can keep the SPSS format or change formats using SAVE commands and extensions:

```
SAVE OUTFILE="C:\TEST\NEWDATA.XLS"
```
will produce an SPSS file.

```
SAVE TRANSLATE OUTFILE="C:\TEST\NEWDATA.XLS"/TYPE=XLS    /MAP /REPLACE /FIELDNAMES.
```
will produce an Excel file.

Create Graphs to Visualize Data

Graphs provide an overview of the data and possible insights that may not be apparent from tables. SPSS's graphs can be a very powerful and useful way to visualize your data. (See "Using SPSS to Visualize Data," WERA Newsletter, *Spring 2007 Standard Deviation*, pages 26-31, at http://www.wera-web.org/pages/publications.php). For syntax and menu procedures to produce tables and graphs, please go to:
http://docushare.everett.k12.wa.us/docushare/dsweb/Get/Document-10657/Meld+StDevMay07+spss+syntax+link+rev4.24.07.pdf

Continued from previous page...

You can use either menu commands or syntax (see Appendix to this article) to produce:

- Frequency distributions of scores
- Population pyramids, which show the distribution of a variable such as test scores or age vertically, split by variables such as gender
- Crosstabs tables showing the relationship between two categorical variables (gender and ethnicity) by counts, percents, and chi-squared analysis.
- Histograms displaying data distribution grouped in intervals
- Box plots showing numeric values across categories in terms of percentiles, medians, and outliers
- Scatter grams illustrating the correlation between two numeric variables.

Final Suggestions:

- Run the syntax in the Appendix by substituting your own path and file names for those shown. Use the variables of interest to you. Save your version of the syntax for future use.
- To produce a data dictionary listing all your variable and value names and labels, run:

DISPLAY DATA.

- From here, you can either cut-and-paste the data dictionary into Word or Excel, or EXPORT to Word, Excel or HTML formats using the menu commands.
- Set your printer to landscape if you are running wide tables. That way, they are less likely to get split over two pages and will be much easier to work with.

References: Where to Go for Further Help

Books

An Intermediate Guide to SPSS Programming: Using Syntax for Data Management
(Paperback) by Sarah Boslaugh, Sage, 2004. For more information:
http://books.google.com/books?id=mf8pEAAzN4gC&dq=an+intermediate+guide+to+spss+programming&sa=X&oi=print&ct=book-ref-page-link&cad=one-book-with-thumbnail&hl=en

SPSS Programming and Data Management: A Guide for SPSS and SAS Users, by
Raynald Levesque, SPSS, Inc., 2005. For more information: http://www.spss.com/spss/data_management_book.htm
If you would like to download a PDF file of this book (540 pages) to save on-screen or print out:
http://www.spss.com/spss/SPSSdatamgmt_4e.pdf

SPSS 11.5 Syntax Reference Guide, Volumes I and II, SPSS Inc., 2002.
An electronic version of the base reference guide, the spssbase.pdf file, can be found in SPSS by going to the Help menu. Select Syntax Guide and then Base.

SPSS Data Analysis and Statistics Books and Manuals by Marija Norusis in cooperation with
SPSS, Inc. Several books dealing with statistics and data analysis can be found at: http://www.norusis.com/

Continued from previous page...

Additional Websites

East Carolina University:

- SPSS Lessons: Univariate Analysis http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Lessons.htm
- SPSS Lessons: Multivariate Analysis: http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-MV.htm
- SPSS Links: http://core.ecu.edu/psyc/wuenschk/spss.htm

Harvard-MIT Data Center, Guide to SPSS
http://www.hmdc.harvard.edu/projects/SPSS_Tutorial/spsstut.shtml

Raynald's SPSS Tools: http://www.spsstools.net/
(If you want to go to only one website, go to this one.)

SPSS Homepage http://www.spss.com/corpinfo/source=homepage&hpzone=nav_bar

UCLA Academic Technology Services Resources to help you learn and use SPSS
http://www.ats.ucla.edu/stat/spss/

WERA Newsletter, ***The Standard Deviation,*** at
http://www.wera-web.org/pages/homepage.php?page=homepage (See Spring 2007.)

SPSS built-in tutorial and help menu

APPENDIX – Sample Syntax

```
***sample syntax**********************************************************.

GET
  FILE='C:\ASSESSMENT\SPSS\YOUR_DATAFILE.sav'.

**frequencies and crosstabs with row percent and chi-square analysis********.

FREQUENCIES
  VARIABLES=writescale
  /ORDER=  ANALYSIS .

CROSSTABS
  /TABLES=gender  BY writelev
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT ROW
  /COUNT ROUND CELL .
```

Continued from previous page...

```
**histogram with normal curve, population pyramid ************************.

GRAPH
  /HISTOGRAM(NORMAL)=sciescale .

XGRAPH CHART=([COUNT][BAR]) BY scielev [c] BY gender[c]
   /COORDINATE SPLIT=YES .

**correlations (bivariate)********compare means and anova******************.

CORRELATIONS
  /VARIABLES=readscale writescale mathscale sciescale
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .

MEANS
  TABLES=writescale  BY ethnic
  /CELLS MEAN COUNT STDDEV
  /STATISTICS ANOVA .

**scatterplot and boxplots*************************************************.

IGRAPH /VIEWNAME='Scatterplot' /X1 = VAR(mathscale) TYPE = SCALE /Y = VAR
 (sciescale) TYPE = SCALE /COORDINATE = VERTICAL  /X1LENGTH=3.0
 /YLENGTH=3.0 /X2LENGTH=3.0 /CHARTLOOK='NONE' /SCATTER COINCIDENT = NONE.

EXAMINE
  VARIABLES=readscale BY ethnic /PLOT=BOXPLOT/STATISTICS=NONE/NOTOTAL.

EXAMINE
  VARIABLES=readscale BY gender BY ethnic          /PLOT=BOXPLOT/STATISTICS=NONE/NOTOTAL.
```

-Andrea Meld, Ph.D., *is a Senior Data Analyst at OSPI and WERA Board Member. Contact information:* *andrea.meld@k12.wa.us*

Stupid Excel Tricks

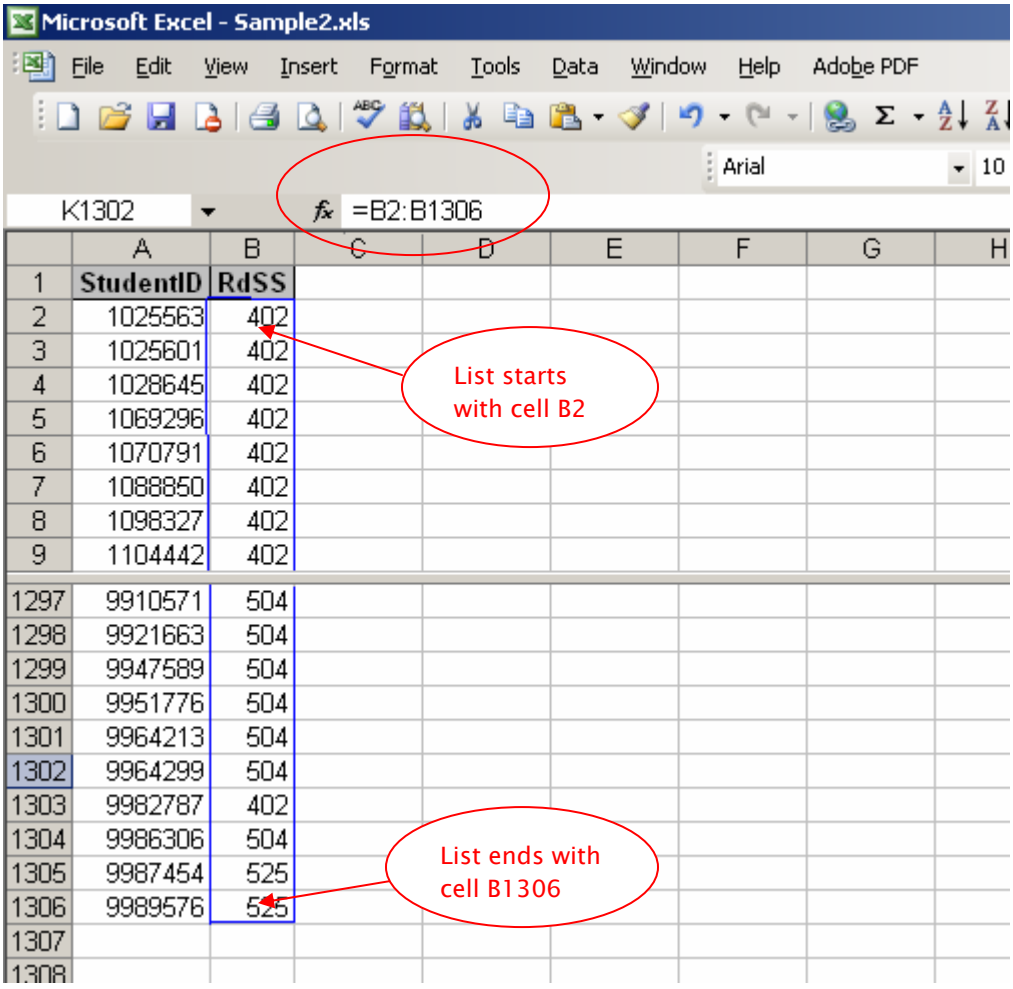
By Patrick Cummings

Introduction

Excel has many statistical functions available and, often it is easier to crunch the data in a spreadsheet rather than bothering with importing to SPSS. Using a simple list of WASL Reading Scale Scores, let's take a look as some of the crunching available to us.

Start with WASL scores:

Let's begin by opening a list of test scores. Below is a sample data set of two rows, **StudentID** and **WASL Reading Scale Scores**: I split the screen so you could see the beginning and end of the list, starting with cell **B2** and ending with **B1306.**



	A	B	C	D	E	F	G	H
1	StudentID	RdSS						
2	1025563	402						
3	1025601	402						
4	1028645	402						
5	1069296	402						
6	1070791	402						
7	1088850	402						
8	1098327	402						
9	1104442	402						
1297	9910571	504						
1298	9921663	504						
1299	9947589	504						
1300	9951776	504						
1301	9964213	504						
1302	9964299	504						
1303	9982787	402						
1304	9986306	504						
1305	9987454	525						
1306	9989576	525						
1307								
1308								

Continued from previous page...

Absolute Value

Sometimes we copy and past cell ranges and want to make sure the data list remains fixed. By using a "**$**" we can lock the array of numbers and make them "absolute values". Therefore, the list of test scores in the array **B2:B1306** can be fixed or made stable by adding "**$**". It then looks like this: **$B$2;$B$1306**.

Let's Roll

Below is a list of common descriptive statistical functions that we can run on the list of test scores or student IDs. Here goes:

	A	B	C	D	E
1	StudentID	RdSS			
2	1025563	402			
3	1025601	402			
4	1028645	402		Lowest Score	315
5	1069296	402		Hightest Score	525
6	1070791	402		Mean	424.5
7	1088850	402		Median	428.0
8	1098327	402		Deviation	28.1
9	1104442	402		Student Count	1305
10	1105370	315		Number Met Standard	1112
11	1115022	315		Number Did Not Mt Std	191
12	1117259	402		Blank Records	2
13	1117961	404			
14	1118341	404			

=**MIN**($B$2:$B$1306)

=**MAX**($B$2:$B$1306)

=**AVERAGE**($B$2:$B$1306)

=**MEDIAN**($B$2:$B$1306)

=**STDEV**($B$2:$B$1306)

=**COUNT**($A$2:$A$1306)
Note: we count studentID here

=**COUNTIF**($B$2:$B$1306,">399.9")

=**COUNTIF**($B$2:$B$1306,"<400")

=**COUNTBLANK**($B$2:$B$1306)

Continued from previous page...

Wait, there's more!

By taking the lowest scores, **315**, and adding **10 Scale Score Point** until get to the highest score, 525, we can break the list of reading scores to 22 different levels.

	G	H
1		Cumulative Frequency
2	315	6
3	325	6
4	335	6
5	345	21
6	355	27
7	365	57
8	375	71
9	385	107
10	395	155
11	405	290
12	415	440
13	425	575
14	435	805
15	445	971
16	455	1187
17	465	1250
18	475	1283
19	485	1292
20	495	1292
21	505	1301
22	515	1301
23	525	1303

=**FREQUENCY**($B$2:$B$1306,G2)

This asks "what is the frequency of scores that equal **315**, which is cell G2)

=**FREQUENCY**($B$2:$B$1306,G15)

This asks "what is the frequency of scores that equal **445**, which is cell G15)

We can add a Normal Distribution like this:

=H2
This is only for the first cell

=H3-H2
Or 6
- 6
0

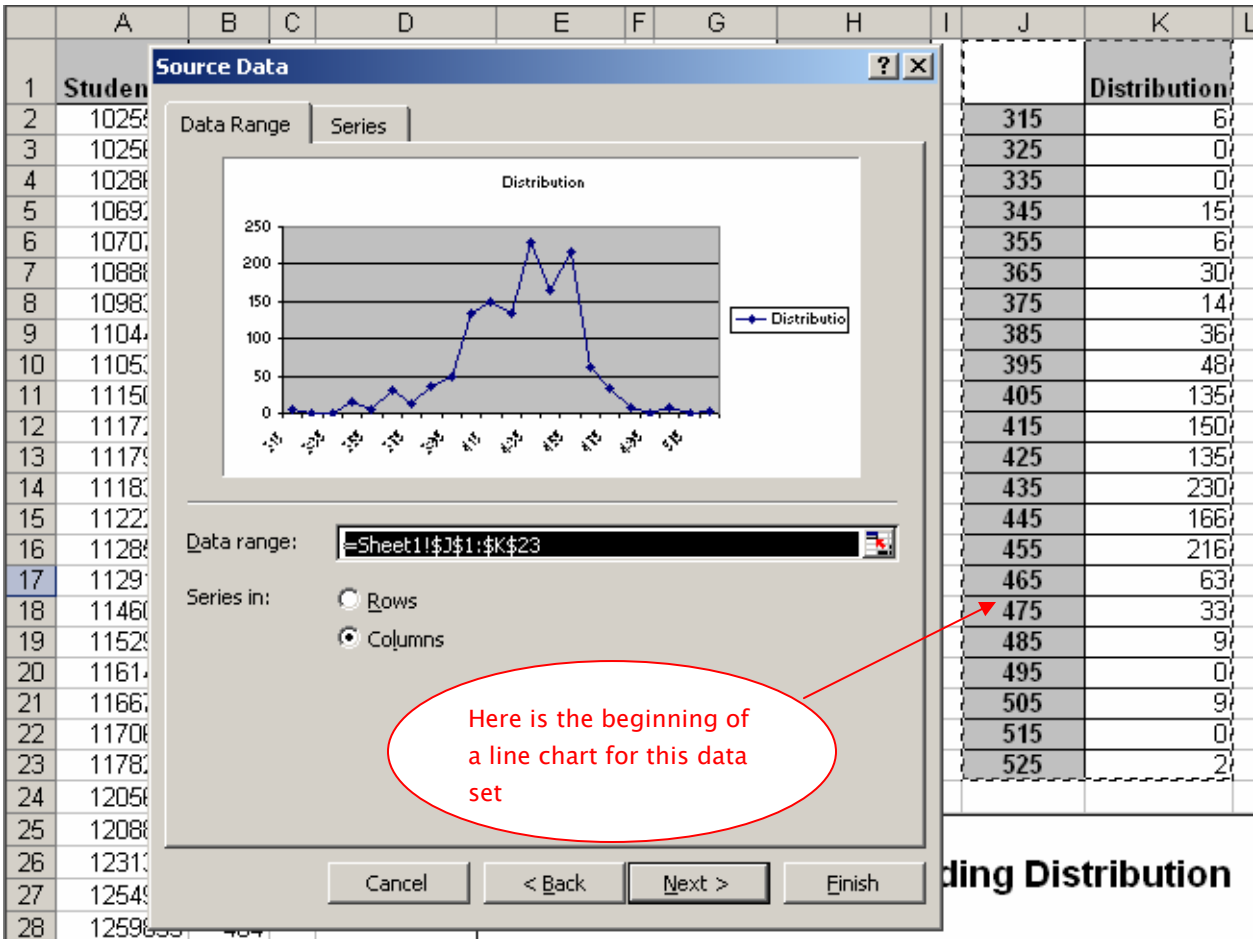
fill down to

=H23-H22
or
1303
- 1301
2

	G	H	J	K
1		Cumulative Frequency		Distribution
2	315	6	315	6
3	325	6	325	0
4	335	6	335	0
5	345	21	345	15
6	355	27	355	6
7	365	57	365	30
8	375	71	375	14
9	385	107	385	36
10	395	155	395	48
11	405	290	405	135
12	415	440	415	150
13	425	575	425	135
14	435	805	435	230
15	445	971	445	166
16	455	1187	455	216
17	465	1250	465	63
18	475	1283	475	33
19	485	1292	485	9
20	495	1292	495	0
21	505	1301	505	9
22	515	1301	515	0
23	525	1303	525	2

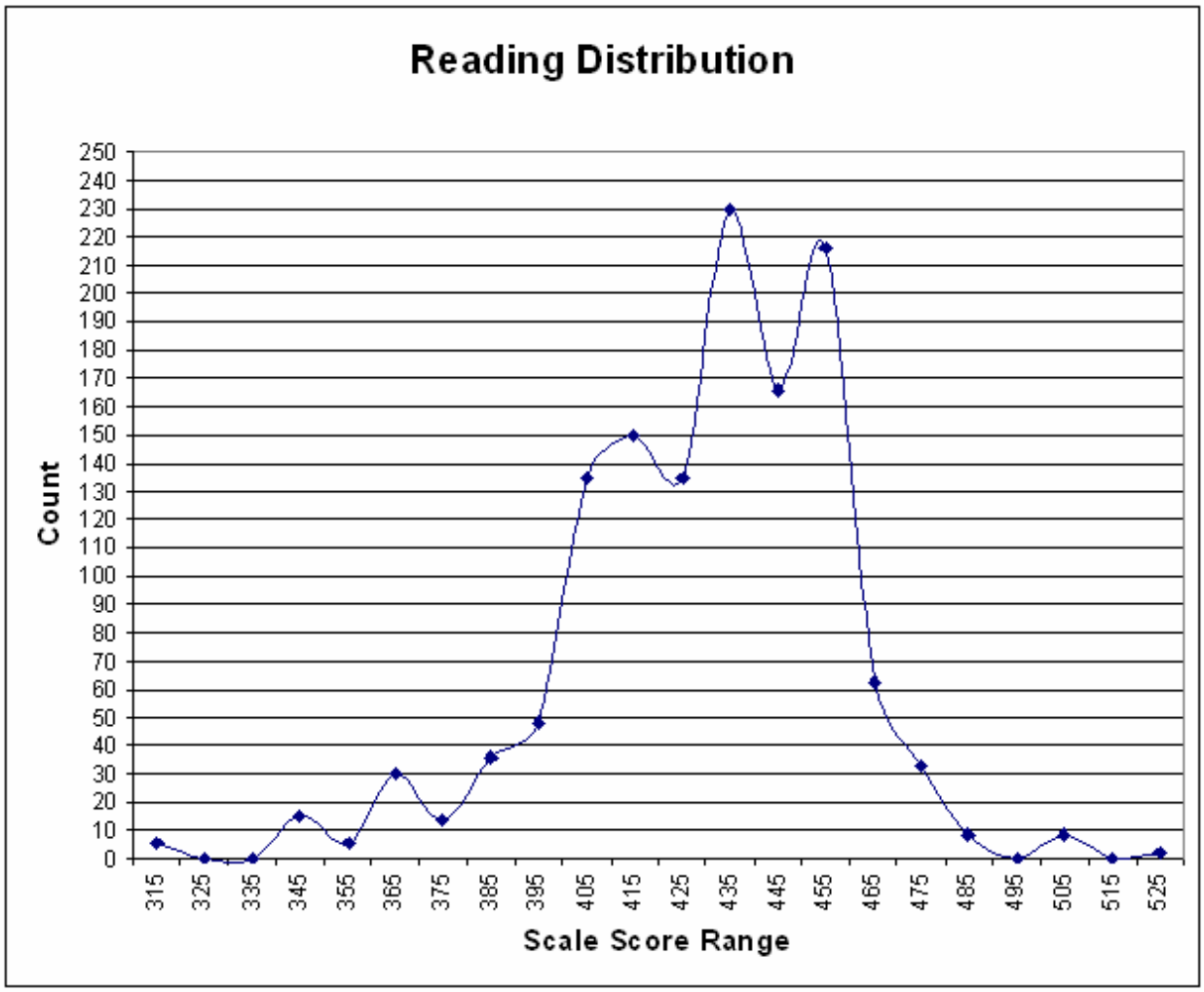
Continued from previous page...

Once we have a normal distribution list we can display a bell shaped curve distribution of the reading scores.



Continued from previous page...

After cleaning up the chart and reformatting we get a nice display of the data:



Now who knew data displaying could be so much darn fun!

-Patrick Cummings is Director of Research and Evaluation for Tacoma Public Schools and is a regular contributor. Contact him at pcummins@tacoma.k12.wa.us

H. M. Jackson High School Band (Everett Public Schools) performs at the Winter Conference

Facing a challenging accountability dilemma? Need expert advice?

Do you have a question about assessment or testing? Let noted assessment author W. James Popham have a crack at it. Author of more than 20 books on education assessment, Popham is well known for his insightful analyses of sticky assessment issues as well as for his colorful commentary. Send your questions by e-mail to el@ascd.org with "Ask About Accountability" in the subject line. Responses to selected questions will appear in Popham's monthly column in Educational Leadership.

WERA Board Goals

- Visibly and purposefully advance a social justice agenda specifically highlighting the research and strategies that result in all children achieving or exceeding standard.
- Increase the participation and membership of Washington State higher education professors and students in WERA and the sharing of their research at conferences.
- Increase the membership's awareness of current research taking place in Washington State through conference sessions, the Standard Deviation and the WERA website.
- Increase the applications and nominations for the awards and grants.

WERA Board Members

Lorna Spear, President
Executive Director, Teaching & Learning
Spokane School District
200 North Bernard Street
Spokane WA 99201
509-354-7339 phone
509-354-5965 fax
lornas@spokaneschools.org

Nancy Arnold, President-Elect
Assistant Director, Special Services
Puyallup School District
214 West Main
Puyallup, WA 98371
253-435-6532 phone
253-841-8655 fax
arnoldnl@puyallup.k12.wa.us

Pete Bylsma, Past President
Education Consultant
8332 New Holland Court NE
Bainbridge Island, WA 98110
206-201-3074 phone
bylsmapj@comcast.net

Executive Secretary
Leonard Winchell
Washington Educational Research Association
Po Box 64489
University Place, WA 98464
253-564-4816 phone
253-564-4816 fax
lenwwa@aol.com

Members-at-Large

Phil Dommes
Director of Assessment
North Thurston School District
305 College Street NE
Lacey, WA 98516
360-412-4465 phone
360-412-4555 fax
pdommes@nthurston.k12.wa.us
(Term expires April 30, 2008)

Emilie Hard
Principal
Glacier Park Elementary School
Tahoma School District
23700 SE 280th Street
Maple Valley, WA 98038
425-432-7294 phone
425-432-6795 fax
ehard@tahoma.wednet.edu
(Term expires April 30, 2008)

Andrea Meld
Research/Assessment Analyst
OSPI
P.O. Box 47200
Olympia, WA 98504
360-725-6438 phone
360-725-6333 fax
andrea.meld@k12.wa.us
(Term expires April 30, 2009)

James Leffler
Program Director, Services to the Field
Northwest Regional Educational Laboratory
101 SW Main Street, Suite 500
Portland OR 97204
800-547-6339 ext. 649 phone
503-275-9584 fax
lefflerj@nwrel.org
(Term expires April 30, 2009)

Washington Educational Research Association
PO Box 64489
University Place, WA 98464

We're on the Web!

Visit us at:

www.wera-web.org

The Standard Deviation
January 2008

The Standard Deviation is published spring, winter, and fall as an online newsletter. Submissions are welcomed from WERA members and others.