NCME national council on measurement in education

VOLUME 21, NUMBER 1 NEWSLETTER March 2013

FROM THE PRESIDENT

Gregory J. Cizek, University of North Carolina at Chapel Hill

Dear Colleagues:

This is the final column I'll be writing as NCME President. I'd like to use this opportunity to highlight other great things that NCME members have been doing, and to express the great honor and privilege it has been to serve. First, some highlights.

NCME 75th Anniversary Celebration

A major event each year for NCME members is the annual meeting. Final preparations are being made for the upcoming conference, to be held April 27–30 in San Francisco. This year, 2013, marks the 75th anniversary of NCME, and the annual meeting this year will include many ways to mark the occasion. The 75th Anniversary Gala is a free event for members of NCME and it promises to be a night of fun, reconnecting with friends and colleagues, and entertainment. The gala will be held:

Sunday, April 28th, 2013
6:00-8:30 p.m.
Grand Ballroom, InterContinental Hotel

Please save this date and join your colleagues as we celebrate together this major milestone. The festivities will include:

- Complimentary hors d'oeuvres and beverages
- Musical Parodies of Measurement
- Psychometric Silliness
- Time Capsule
- NCME Timeline
- Champagne Toast, Anniversary Cake, and more...

Overall, I can tell you that a lot of volunteered time and effort has gone into the planning of the gala, both on the part of the NCME Central Office staff and the 75th Anniversary Planning Committee: Neal Kingston (Chair), Gretchen Anderson, Amy Clark, Linda Cook, Anne Fitzpatrick, David Frisbie, Kris Waltman Frisbie, Karoline Jarr, Susan Loomis, Plumer Lovelace, Kimberly O'Malley, Barbara Plake, W. James Popham, Elaine Rodeck, and Edward Roeber; Master of Ceremonies, Ron Berk. If you know any of these great folks, please let them know that you appreciate the work they have done to make the celebration a memorable one.

From the above list of festivities, I'd like to highlight the NCME Timeline. The timeline will provide a unique opportunity to view a chronological history of NCME. Designed to capture NCME since its founding in 1938, the timeline provides a graphic and pictorial display of our organization's history. It begins with our founding by 45 college teachers of measurement and 12 public schools or state Departments of Educations as the National Association of Teachers of Educational Measurement, to our name change to National Council on Measurement Used in Education in 1940, to our current identification as the National Council on Measurement in Education in 1960. Along the timeline, the names of the organization presidents are shown along with some photos of key figures, important NCME Board decisions, and historically significant educational policy and organizational milestones.

The 18’ Timeline will be displayed throughout the 2013 annual meeting. Its main display area will be at the NCME Headquarters in the InterContinental Hotel, but it will be moved for display at the gala and at the annual breakfast meeting. In addition to viewing the timeline, conference attendees will be encouraged to use sticky notes to interact personally with the timeline, inserting other events that have personal meaning onto the timeline. Two copies of the timeline will be printed, one for display at the conference (and other uses); the other will be placed in a Time Capsule—another special feature of the 75th Anniversary—that will be sealed for opening in 2038 at NCME’s 100-year anniversary. A digital version will be located on the NCME website.

A number of NCME members collaborated to create the timeline. Thanks for all of their efforts are due to the Timeline project committee, made up of Gretchen Anderson, Linda Cook, Plumer Lovelace, Katie McLarty, Deanna Morgan, Barbara Plake, and Rosemary Reshestar.

Finally, I would also like to remind all NCME members attending the conference to be sure and pick up their complimentary commemorative 75th Anniversary computer bag when completing on-site badge pick up and registration.

NCME Board Meetings; Annual Breakfast and Business Meeting

Two board meetings are scheduled to occur in San Francisco. One meeting will be held on Saturday, April 27, from 4:00–7:00pm, and the second on Tuesday, April 30, from 4:00–7:00pm. Both meetings will be held in the Cathedral Hill meeting room of the InterContinental Hotel. Any NCME member may attend as an observer, and all are certainly welcome to do so.

On a somewhat related note, I want to highlight the role that Jennifer Kobrin (JKobrin@collegeboard.org) has played in supporting the work of the NCME Board of Directors. Jennifer is the Recording Secretary of the board—an office that perhaps many of us didn’t even know existed. Jennifer is the master keeper of all records, updater of the NCME Handbook and Policies, recorder of meeting proceedings, and just generally a super on-top-of-things resource that, without her contributions, the work of the board would be significantly impeded. On behalf of all of us, Jennifer, thanks for all of your great work.

The major agenda item of the first meeting will be the board’s final consideration and, ultimately, vote on the level at which NCME supports the revision of the *Standards for Educational and Psychological Testing*. NCME joins the American Educational Research Association and the American Psychological Association boards in considering official recognition of this longstanding and influential compilation of best practices in educational and psychological measurement. The NCME Handbook provides for three possible levels of recognition. From lowest to highest level, they are:

> Level 1 - “Acknowledgment of Receipt”: A document (solicited or unsolicited by NCME), which provides an explication of issues and is prepared for the purpose of informing an issue and/or promoting discussion.
>
> Level 2 - “Board Recognition”: The document is recognized by the Board as reflective of sound measurement practice.
>
> Level 3 - “NCME Endorsement”: Materials are developed by NCME and are prepared for the purpose of defining technically sound professional practice. The Board and NCME membership review these materials, which are endorsed only by a 2/3 vote of the Board. The standards would directly affect the practice of nearly all members, and they would serve as an ethical imperative for professionals in the field.

A highlight of each annual NCME conference is the annual breakfast and business meeting. This year’s event will be held on Monday, April 29, from 8:00-10:30am in Grand Ballroom B of the Intercontinental Hotel. The meeting is a great opportunity to reconnect with colleagues over a meal, to learn about our association’s accomplishments over the last year, and to join in honoring the many NCME members who will be receiving awards at the breakfast. In addition, I have planned a rather non-traditional Presidential Address this year that I hope will provide great encouragement for all of us to press onward with the great work we are doing advancing the science and practice of educational measurement. I won’t say more about the content of the address, except to perhaps pique interest by disclosing its title: *An Unpublishable Presidential Address.*

Plans for an NCME Charitable Giving Arm

One of the projects begun by the Board in the last year has been investigation of a Charitable Giving Arm. Initiated by past president Linda Cook (lcook@ets.org) in early 2012, an ad hoc committee has been working to create a mechanism for NCME members to support the goals of our association.

The committee has developed a draft of a mission statement, which identifies its purpose:

> To provide donors with a means by which they can express tangible support for advancing NCME's mission in the science and practice of measurement in education and for recognizing students and junior scholars, researchers, and practitioners working in the field. This support would be tax-deductible and could consist of cash, book royalties, or other kinds of assets.

The establishment of an NCME charitable giving arm potentially allows for:

- Funding for awards from non-operational revenue sources
- Credible structure and choices for donor contributions
- Opportunities to promote NCME and giving to NCME
- Positive recognition of benefits of donor funds in advancing the field
- Help in further building NCME as a community of scholars

The committee presented an initial plan at the October 2012 Board meeting and a revised plan at the January 2013 Board meeting. Currently, the ad hoc committee is drafting text for the NCME Handbook regarding the establishment of a standing Fund Development Committee. The Board of Directors is scheduled to review these proposals at its April meeting. For more information, contact Linda Hargrove (Linda.Hargrove@THECB.state.tx.us).

Plans for Establishing an NCME Archive

Another initiative underway is planning for an NCME archives. An Archives Committee was formed, and its first task was to develop a proposal for NCME archives policies and practices. That proposal was approved by the NCME Board in April 2012. Stated goals were to (1) preserve information about the organization that is of historical interest and not needed for daily business, and (2) make that information accessible for use. Plans were to establish both physical and digital archives (with digital copies to be made of physical items for accessibility purposes and physical copies to be made of digital items for preservation purposes). In July 2012, the committee submitted an implementation plan for the physical and digital archives, along with preliminary cost estimates. After a lengthy discussion, the NCME Board approved the plan, but requested refined cost estimates for the vendor and IT support that would be required to digitize physical items and add content to the NCME website.

The Archives Committee then commenced to build a small-scale demonstration of the digital archives, to help with the task of refining costs. A representative sampling of content and file types was identified and cost estimates were obtained from various vendors. That exercise demonstrated something that the Archives Committee didn't expect; namely, the cost of digitizing physical materials would be prohibitively expensive (largely due to the perceived sensitivity of the materials). Hence, the committee decided it was necessary to downsize the scope of the project.

Current thinking of the Archives Committee is that the digital archives will be developed from available digital content only, and that materials from the current physical archives (i.e., items stored at the NCME Central Office's storage facility) will not be digitized unless there proves to be a demand or need for it. An index or catalog of what is available in the physical archives will be published on the website, so that people can request access to it. The committee will present its findings and refined plans to identify digital archival material to go on the website and work with the Website Committee to publish it to the web at the upcoming NCME Board meeting. Additional information on this project can be obtained from the committee chair, Mary Pommerich (mary.r.pommerich.civ@mail.mil).

Thanks

To close this final installment of my presidential columns, I want to say how much I have appreciated the trust to serve as president of our association. It has been an honor and a privilege to serve. I am so grateful for the work of so many of us who volunteer our time and expertise to make NCME a truly outstanding organization. I have been blessed that this term of office has given me the opportunity to cross paths with so many colleagues that I might not have gotten to know otherwise. Please accept my most sincere thanks and best wishes for much continuing success....G>

GREETINGS FROM THE EDITOR

Susan Davis-Becker, Alpine Testing Solutions

In this issue we feature the fourth and final column from NCME president, Gregory Cizek, who highlights several of the exciting efforts of our organization and some of the events to look forward to at the upcoming conference. We also welcome Melinda Montgomery as our new graduate student columnist, who shares some insight for graduate students preparing for the annual conference. Our Spotlight Member in this issue is Wim van der Linden, NCME's incoming president. We have also continued our member perspectives on the future of NCME in this issue with contributions from Derek Briggs and Marianne Perie. In continuing our thinking about the future of our field, we asked two researchers to share their thoughts on the current and future states of item and test generation – very interesting thoughts! Next, our NCME training session chairs provide some highlights to the training sessions available at the upcoming annual conference. Finally, Patrick Meyer provides some guidance on how to utilize the NCME website RSS feeds so that we all can stay better connected with what is happening in NCME!

As we all begin to thoughtfully prepare our presentations for the upcoming conference, I want to leave you with a quote shared with me by Jerry Melican (one of our Newsletter Board members):

> "A speech is a solemn responsibility. The man who makes a bad thirty-minute speech to 200 people wastes only a half-hour of his own time. But, he wastes 100 hours of audience's time – more than four days – which should be a hanging offense." - *Jenkin Lloyd Jones*

I look forward to seeing everyone in San Francisco!

GRADUATE STUDENT CORNER

PREPARING FOR THE NCME CONFERENCE: AIMING FOR SUCCESS AND MANAGING COMPLICATIONS

Melinda Montgomery, University of Kansas

This is probably my favorite time of year not only because spring is almost here but also because the NCME conference is approaching. For first time presenters in particular, the upcoming conference can create a combination of excitement mixed with varying degrees of terror. By this point, the initial excitement of learning that your research project was accepted has worn off. Now the research must be completed, the paper written, and the actual presentation prepared. Whether it is a formal presentation or a poster, there are often differences between how the presenter envisions the presentation experience and the reality of the presentation. While this will probably always be the case, preparation can make the process much smoother.

One difficulty when preparing for the conference is knowing how to plan for different types of contingencies. In the discussion that follows, we will consider two common presentation formats—poster presentations and paper presentation—and strategies for handling some of the potential complications inherent in each type of presentation.

Poster Presentations

Complication one: printing the poster

Printing your poster allows you to display your work in the most professional format possible. Because printing a poster can be expensive, some presenters opt for printing off the information and taping it on a board. While that can work, keep in mind that some universities have poster printing services available to faculty and students that are much less expensive than the commercial options. I recommend that presenters investigate their options and arrange for the poster format that best fits their financial constraints while showcasing their work.

Complication two: transporting the poster

Poster displays should be kept in a pristine condition; free from wrinkles and creases. If air travel is required, presenters should consider acquiring an appropriate poster container to protect the poster during transit. Often faculty own containers that they may be willing to loan to student presenters. Since the containers are large enough to accommodate more than one poster, sometimes other graduate students have access to containers and are willing to transport your poster along with theirs. I opted for this method one year when a colleague, who was also presenting a poster, offered to transport my poster along with hers. It is also possible to buy an inexpensive cardboard carrier.

Complication three: the poster presentation

The goal of a poster presentation is to draw people in by your display so that they will show interest and ask questions. One of the best ways to accomplish this is to use pictures and/or graphs in the poster. People will stop and look at a picture or graph as opposed to a mass of text. Few people will stand there and read your poster unless something draws their attention. Remember, "A picture (or graph) is worth a thousand words." It also tends to be more inviting when there is plenty of white space.

Complete the research prior to the conference and have a printed copy of the complete paper with you for reference. Sometimes people will ask specific questions about your sample or analysis that is not on the poster, so having a reference with you will assist you in easily answering any questions that might arise. You might also consider providing a handout, no more than one page, with the name of your presentation, some important details, and your contact information. This way after the conference, people will have something to remind them of your poster and prompt them to read your paper. Another strategy would be to have a sign-up sheet near your poster so people who would like more information can provide their email address. Alternatively, some university career centers provide very inexpensive business cards that you could hand out to interested people so that they could contact you for more information.

Try not to take it personally if someone should question your research. At my first poster presentation, one person stopped and asked me, "Why are you interested in this topic?" I was a little taken aback by this question but explained that my background as an educator had led me to conduct this study. Another person questioned my methodology by suggesting another analysis that I could have or, in their opinion, should have done. I listened to their suggestions and took notes. While I did not implement those suggestions in that project, it was a good opportunity for me to look at the project differently and to consider other approaches.

Paper Presentations

Most first time presenters envision explaining the motivation, the methodology, and their results in detail during their presentation followed by an insightful question and answer period. That is probably not going to happen in the ten to twelve minutes you have for a typical presentation. The truth is there simply is not enough time to explain the project in detail. The reality is that unless your presentation is the first one in the group it is very likely that the previous presenter will take some of your time. Be prepared to cut something if you have to. In my first presentation both of the presenters before me ran over their allotted time. I chose to simply touch on a few slides that I had planned to explain further and focus on the big ideas and results.

Presentation logistics and format

Whether you plan to use your own laptop or a computer provided always have a backup. It is a good idea to have the presentation on a flash drive or in a Dropbox folder. It is also good to have a hard copy of the presentation with notes. Some people prefer to write out a script of what they want to say so that they do not forget important parts of the presentation. That can be effective as long as it does not appear that the presenter is simply reading the script to the audience. The same thing applies to reading the slides to the audience. The slides should contain only the bulleted highlights and your notes should contain everything else that you want to say. Similar to poster presentation, some presenters opt to use primarily images, diagrams, or simple graphs to convey the message visually rather than with text. To avoid reading the presentation, practice, practice, practice. Practice the presentation so much that you no longer need to refer to your notes, but have them with you just in case.

Presentation content

The focus of the presentation should be on the research question(s), the methods, and the results. I received some very good advice from Dr. Skorupski when I was preparing for my first presentation. He said to consider the presentation as an advertisement for the paper. The goal of the presentation is to create interest in your research and encourage participants to get a copy of the paper. I think this is good advice for both posters and oral presentations. This also takes off some of the pressure of providing a thorough explanation of the project in the allotted time.

Begin with a cover page containing the title, author(s), and university or testing company where the research was conducted. If you tend to get nervous, the cover page will remind you to introduce yourself. Personally, I do not think an outline of what will be in the presentation is needed. Going over this will take several minutes that could be used for explaining the methods or results. A literature review is usually not needed in a presentation but it should be in the paper. As with the outline, explaining the review of literature will take up valuable presentation time.

Many advisors recommend that equations not be used on slides unless the presentation is about a new equation. Slides containing derivations of an equation are very difficult to follow in a ten minute presentation. I like to leave time at end of the presentation to list some ideas for future research. This will encourage the audience to get a copy of the paper and maybe work on some of those future research projects.

While I have only attended the last two NCME conferences, I have seen some interesting things particularly during question and answer time. At one presentation last year, by the time the question and answer time came I was the only remaining audience member. I asked a question just to fill the time. When no one has any questions, there is complete silence. Last year I saw a moderator trying his best to get the audience to ask questions, which was very awkward. More than once, I have observed audience members who appear to have attended the presentation with a single goal: questioning a presenter's research. It is very awkward not just for the presenter but for the other audience members. While it is unlikely that any of those things will happen to you, with a conference this large there is always an element of unpredictability.

Final Thoughts

This is your study; you made decisions along the way, so stand by your decisions. Be prepared to defend what you have done. By defend I do not mean to be argumentative. Explain what you did and why you did it, but keep an open mind. Just because someone else might have approached the topic differently, that does not necessarily mean that your approach was wrong. However, listen to the suggestions of other researchers, take notes on their suggestions; you might find them useful in a future project or if you plan to submit this project for publication. You could address the concerns in your paper even if you decided not to implement them. As graduate students, this is all part of our learning process.

If this is your first conference presentation, I hope that my suggestions help you to be better prepared. The process can be a little unnerving the first time but it gets better. Each time I present I have a better understanding of what to include and what to leave out. My final word of advice is to enjoy the presentation and the conference. Go to some presentations that are different for your current research interests. This will not only broaden your knowledge, but possibly expand your research interests.

SPOTLIGHT ON THE PEOPLE WHO MAKE OUR ORGANIZATION GREAT – WIM VAN DER LINDEN, CTB/MCGRAW-HILL

For this issue, we are fortunate to receive some insights and reflections from Dr. Wim van der Linden, Chief Research Scientist at CTB/McGraw-Hill and the vice president of NCME.

How did you get into this field?

Like most of us probably, I made quite a detour. During my years in high school, I was obsessed with chemistry—biochemistry to be more precise. Watson and Crick had just discovered the double helix structure of DNA, and the popular science literature was full of speculation about what would be next. In the Netherlands, where I received all my degrees, you immediately have to choose your major if you go to university. I went to Utrecht, to quickly discover how boring chemistry was. Monotonous lectures every morning and lab work every afternoon, just too much for me. Especially the lab work was a burden; I just couldn't get excited if something happened in my test tube or flask. In my second year, I decided to transfer to psychology, which I immediately liked much better. But it did miss the rigor I had expected to find. It was therefore not too long before I discovered that methodology and statistics were my true passions. But, alas, at the time specialization at the level of a master's degree in this area was impossible. So I began planning a career in social psychology, with as much emphasis on research as possible. Then, during my master's, several things happened almost simultaneously. I picked a class in test theory, which I had never done before, and fell in love with it: Finally, something that had the combination of substance and rigor I had been looking for! At the same time, a new department was established in psychology that did offer specialization in methodology and statistics, including a test theory track. And just when I completed my master's, this department had a faculty opening. At the time, the common route to a PhD was to get an appointment as a faculty member after your master's, and conduct dissertation research while teaching. I applied, was accepted, and began dissertation research on the use of decision theory in mastery testing. Because of massive student revolts at Utrecht (in the wake of the 1960s), my first years as an assistant professor were not so easy. My dissertation supervisor, Gideon J. Mellenbergh, quickly moved to Amsterdam, and I left Utrecht for Twente. I received my PhD from Amsterdam while working at Twente.

If you weren't doing this what would you do?

After so many years, it is hard to imagine anything else than educational measurement. But if I would be able to do everything over again, I would definitely waste less time and pick up the more foundational literature (mathematics, statistics, and computer science) much earlier. Also, rather than my parallel degrees in sociology, which I completed because of my planned specialization in social psychology, I would probably have followed my interest in music much earlier. I just love classical music. When I was a student, the development of baroque music on period instruments had just begun to catch on in Europe. It was fascinating how, all of a sudden, everything sounded much more natural on these instruments. I still follow this development closely, but my taste is now more inclusive. Actually, I like anything classical. That is, as long as it is based on a clear idea and has a clear form; most avant garde music is formless and hopelessly annoying.

What advice would you offer a graduate student who is thinking about psychometrics?

As I suggested in the NCME Newsletter Graduate Student article last year: "Get your degree if you think you need it. But don't graduate, always remain a student". The field of educational measurement is developing so fast right now. I often feel the same sensation as in the early days of my interest in biochemistry, when the nature of DNA had just been discovered and everyone wondered what would be next. If you think a degree in educational measurement is enough, and you stop further familiarizing yourself with its fast growing literature (not only on applications but also on the foundational disciplines), you'll soon fall behind. But the rewards are enormous; we are writing history right now.

When not teaching or researching, what do you do or like doing?

Again, music is my other passion. When I joined the University of Twente, something special happened during my first lunch walk. Twente is a university with a campus in a beautiful, rural setting, which features a carillon tower. I don't know who played, but still remember what was played—Mozart's lovely variations on "Ah-vous dirais-je, Maman." And when I learned that the carillonneur at Twente offered lessons, I immediately joined his small group of students. A few years later, I joined the Netherlands Carillon School, part of the Utrecht Conservatory, as a student. But actually I hardly had the time to study. I was already Chair of my Department at Twente, and two years later became Dean of its Faculty of Education. I had just made it to my fourth year, when I had to stop. Shortly thereafter, however, I got an appointment as the municipal carillonneur of a city close to my home town. My job was to play the carillon in its main tower, one night every week and during official ceremonies. I also had to write new short pieces for its automatic play every six months. At the end of 2000, I had to stop because of chronic injuries. A pity, but live concerts and classical radio stations are great substitutes.

What would you say has been one of the biggest innovations in psychometrics in the last decade or two?

The biggest development I have been able to witness is the increasing importance of computers. Not only because of their important role at testing companies, but equally well as object and tool of research. It was just impossible to foresee their impact when I began my career. The second innovation is no doubt item response theory. It began as a more serious development right when I was a student, and is now applied routinely everywhere in educational measurement. It has begun penetrating other areas of measurement as well.

When you go to conferences, how do you pick what sessions to attend?

I just browse the program book carefully and pick what I believe might be innovative. It does not matter whether it is a new application or a theoretical development, both are equally interesting. At the same time, I try to avoid presentations that only add a little twist to what is already known. In fact, I cherry pick and consequently change sessions with multiple papers a lot. As for our own annual meetings, the lack of abstracts in its program books has always annoyed me. It makes it so difficult to choose, and if you miss something there is no second chance. As of the 2014 annual meeting, our program book will have abstracts of all presentations.

Who has been a significant influence in your professional life?

I have been overwhelmingly blessed with opportunities to learn from the leaders in our field. From Don Mellenbergh I've inherited a special interest in the conceptual aspects of what we do. Technical progress is nice, but only as long as it has a sound conceptual basis. Don is also a prolific writer, who has paved the way for me to international journals and taught me how to deal with reviewers. I learned IRT from Fred Lord, during a three-week summer course at ETS in the late 1970s, where he taught what later turned out to be the first version of his 1980 monograph, and Bayesian statistics from Mel Novick a few years later, during a course in the Netherlands, where he taught his 1974 book with Jackson that had just been published (and, how shameful, never has been reprinted). Lord and Novick, who else can claim that! I missed George Rasch when he visited the Netherlands in the later 1960s, but learned about his models from Gerhard Fischer when he taught his monumental 1974 book (alas, only available in German) during a course he gave in the Netherlands at about the same time. In fact, I can just go on and on. Nearly every leading colleague you know has been a visitor at Twente, usually as the main guest during its annual IRT Workshops my former department there has been organizing (the 28th later this year!). My career has been greatly influenced by opportunities to travel internationally to most important meetings and conferences, for which I'm quite grateful.

CURRENT PERSPECTIVES ON RESEARCH IN ITEM AND TEST GENERATION

Thinking about the future of NCME, we asked researchers in our field to share their perspectives on current and future research on item and test generation.

Perspectives on Item Generation

Susan Embretson, Georgia Institute of Technology

What are biggest challenges to developing and maintaining a testing program to producing high quality items to allow valid interpretations of scores? How are the new models that you are working with addressing some of these challenges?

Test score validity starts with the items. From my perspective, there are three major challenges to producing high quality test items.

The first challenge to producing high quality items is the item writing process itself. Certainly item writers vary in their ability to produce items of a certain type. However, the larger challenge is the quantity of high quality items that can be produced. For complex item types, such as mathematical word problems and paragraph comprehension tasks, only a few items may be produced daily. Thus, item writing often is a slow and expensive process.

The second challenge to producing high quality items is the review process. Certainly review processes are necessary to assure the quality of handcrafted items. However, item review is also a slow and expensive process. Items for high stakes educational achievement tests typically involve several levels of reviews. For example, recently we had items placed for tryout on a state achievement test in middle school mathematics. We reviewed the items initially (at least two independent reviewers) and then the items were submitted to the standard process in the state. First, a mathematical editor reviewed the items (and rejected some). Next, the items were presented to a panel of educators for review as appropriate for the intended indicators of the standards. Some items were also rejected at this stage. Then, the items went to the state board of education for review and, again, more items were rejected. Finally, the items were placed on tests for empirical tryout. Again, some items were rejected. While we had a relatively high rate of acceptance of our (generated) items, nonetheless the process involved the effort of many professionals. And several months of effort were involved.

A third challenge is that support for the response process aspect of test validity is often not available. While extensive review of items as fulfilling content specifications is standard in educational testing, the items must also elicit the intended response processes by examinees. While effort is devoted to the content specifications, with items having multiple levels of review, relevant response process evaluations are often lacking. Thus, for example, a mathematical achievement item may fulfill the content specifications, but the item may also contain content that requires the examinee to engage in construct-irrelevant processes. For example, success on a heavily contextualized mathematical achievement item may depend more on reading than on the intended mathematical skills. Research on items is needed to counter construct-irrelevancy. For example, item difficulty modeling from cognitive complexity variables, think-aloud protocol analysis and so forth are needed to support construct relevancy on similar items. For mathematical achievement items, relevant background research may be especially important for English language learners to assure that it is mathematical skills—not English language skills—that are the main contribution to item solving processes. Thus, an important challenge to test score validity is lack of support for the response process aspect of validity.

These challenges have conflicting impact on the production of high quality items. Item writing and item review are needed to assure item quality but at the expense of item quantity. Support for the response process aspect of validity, would seemingly involve even more extensive review to assure item quality, but again at the expense of item quantity.

Item generation, if based on prior research, has potential to solve these conflicts.

How has research in this area contributed to the feasibility of this approach?

The most apparent contribution of research on item generation is its ability to produce high quality items rapidly. And the lengthy evaluation process can be circumvented under certain conditions. That is, given a research foundation on a particular item type, psychometric properties may be highly predictable. Such research would involve, for example, cognitive modeling of test items to understand item difficulty and discrimination. For item generation based on formal or logical structures, the cognitive models developed could predict the psychometric properties of previously unobserved combinations of item features. Or, for item generation based on the item model approach, in which an existing high quality item is "variabilized," empirical tryout could support the application of the item model psychometric properties to the variants that are generated.

Another important contribution of research on item generation is to highlight the importance of understanding the basis of examinee's responses. To effectively create equivalent generated items, construct-irrelevant substitutions must be understood. Especially for complex items, background research is needed to understand the impact of substitutions on item difficulty and examinees' response processes. It was not so many years ago that these questions were not even asked. That is, it was assumed that if an item correlated sufficiently highly with other items, it met standards for item validity.

What in "classical" test development should/have be retained?

While it may seem that automatic item generation would replace traditional item writing, in fact item writers remain essential in the test development process. However, their role should be upgraded to item designer. That is, while item generation is an excellent means to provide high quality items rapidly, and with predictable properties, the underlying item models and item structures need to be developed. Item designers are needed to initially develop items to fulfill content specifications.

Also, items need to be evaluated for quality. In educational measurement, review of item content by panels of educators and other relevant experts supports the content aspect of validity. Also, empirical tryout to evaluate psychometric qualities is necessary to support the internal structure aspect of validity. I would not eliminate either process; however, I would change the target of these evaluations. Rather than evaluating each generated item, item models and samples of generated items could be evaluated for quality. If the item generator is proven to produce items that meet these quality standards, newly generated items may be able to bypass further evaluation. However, item generation may not yet be at the point where new items should be placed on tests without being viewed at least minimally by human eyes.

What research areas would you suggest for practitioners, especially new researchers?

I would encourage researchers to more thoroughly understand the contributions of construct-relevant and construct-irrelevant processes on item success. That is, the item content that leads to these different types of processes must be identified. This type of research provides the foundation for item generation. However, analyses such as item difficulty modeling require access to item content and to responses by examinees. Unfortunately, such access is becoming increasingly difficult with the proprietary interests for many tests. I believe that fundamental improvements in item development are possible only if item content and data are made more readily available to researchers.

Measurement Design and Development: The Next Generation

Richard M. Luecht, University of North Carolina at Greensboro

What are biggest challenges to developing and maintaining a testing program by producing high quality items to allow for valid interpretations of scores? How are the new models that you are working with addressing some of these challenges?

In my opinion, there are three fundamental challenges facing the measurement research community and testing industry that signal the need for what Thomas Kuhn called a *paradigm shift*: (1) producing and pilot-testing of massive amounts of low-cost items to meet growing demands—especially for on-demand, multidimensional formative tests; (2) efficiently measuring complex cognitive skills; and (3) developing assessments with dedicated purpose, rather than asking more and more of our largely summative test forms and measurement scales. Meeting these challenges requires a long-term commitment to research and change—to iteratively developing practical assessment design and development solutions and then to evaluating and modifying those same solutions under a solid versioning strategy.

For some time now, I have been promoting what I call *assessment engineering* (AE) principles to help effect that change. AE borrows some well-established design principles from modern software- and manufacturing-engineering practices, integrates Mark Wilson's notion of construct mapping, and layers everything on top of Bob Mislevy's evidence-centered design (ECD) framework. Simply put, AE is a way to implement ECD; it is about iteratively developing scalable and replicable principled assessment design solutions that reduce costs and automate as much of the overall testing enterprise as possible. Fundamentally, it is all about *design*. If you do not have a specific design in mind, do not build it. Manufacturing companies have learned that good designs happen in response to careful articulation of very exact specifications. That is part of the problem in testing. Our item writing and test design specifications are not very exact and we seldom have more than correlational evidence that our test forms and scales are what we actually intended to build. It is like an artist randomly slopping paint on a canvas and then declaring, "Yes, that is exactly what I intended to build. Just look how well it hangs together."

Solid design and automation are key components of successful production in most manufacturing settings. Henry Ford taught us that. He changed the world by rejecting the long-held assumption that customized, high-cost craftsmanship was the only way to build carriages. Building an experimental automobile assembly plant in Dearborn Michigan, Ford demonstrated that well-engineered, standardized assemblies built to strict design tolerances could significantly lower costs, improve quality and increase productivity, making it possible for many to purchase high-quality Ford Model-T automobiles. It took no small measure of risk and a great deal of up-front investment to engineer robust production systems and processes. Today, his success is an indisputable fact of history—a success shared by thousands of manufacturing companies since the early 1920's.

Change to more principled assessment design and development approaches will be costly, but ironically, one of the primary reasons why change is needed is COST! Like it or not, it is virtually impossible to talk seriously about making substantial changes in testing practices—to meeting the above challenges—without discussing the costs implications.

Test item production, facilities and testing seat time, and human scoring may be the primary cost drivers in testing, but the real cost driver is humans. There are two near certainties whenever humans are involved in a complex process like a testing enterprise. First, costs will go up. Second, there will be more error and variability in the processes of item development, test assembly and publication, pretesting, test administration, scoring, data processing, and analysis. Humans are useful—I happen to know a few. But, their time is often expensive and the very judgment that helps them solve complex or fuzzy problems also tends to add "noise" to many of their undertakings, especially those like item writing that often to encourage creativity. It is hardly a profound insight to realize that ours is a very expensive industry because it depends too much on human creativity to compensate for design misspecification (or lack of design altogether). I am not saying that item writers can be completely removed from the process or that we can automate everything. Rather, I am implying that the some of the roles that humans play in the process can be altered to form combined human-system hybrid processes where some automation—especially automation that reduces costs and increases quality— is welcome.

Pretesting is another important cost and security issue. Ask yourself how a tablet computer manufacturer could stay in business if EVERY tablet computer coming off the assembly line had to be "field tested" by several hundred or several thousand qualified users to detect potential manufacturing aberrations introduced by creative workers? Test items cost hundreds or even thousands of dollars each and have severe constraints on exposure and reuse, not to mention a limited "shelf life." And, at best, the inferences we can draw about actual proficiency is limited to rather vague alignment between the coded item content and a blueprint or *post hoc* alignment with achievement level standards set long after the test forms are assembled. Maybe it is just me, but that does not sound like a very efficient or cost-effective way to run an industry.

What if we could design systems and procedures to generate thousands of items and hundreds of test forms that ultimately cost only a few dollars each? What if we could produce useful multidimensional formative tests that measure instructional-relevant traits in sufficient quantity and quality to support on-demand, instructionally integrated measurement every day of the school year—rather than relying on teacher-built classroom tests? Meeting those types of massive item-production demands and simultaneously getting costs down seems highly improbable, given traditional item and test development procedures. My argument is that we need to completely redesign those systems and procedures under a completely different paradigm.

Let’s consider the problem of item production. Do we really need highly creative item writers to hand-craft every item? Some of us say, “Probably not!” For example, Mark Gierl, Isaac Bejar Jackie Leighton, and Susan Embretson are some of the more notable researchers (and colleagues) who have made phenomenal strides in automatic item generation (AIG). It seems fair to speculate that today we basically know how to develop item templates, shells, or models to clone or otherwise generate massive numbers of variant or “child” items from a common parent. Creativity is built into the item design, not into the production system that implements AIG. However, those AIG efforts do not necessarily solve another challenge of pretesting all of those items to gather sufficient data to estimate the statistical parameters relative to a common scale for scoring. Referring back to our tablet computer example, it does not completely solve the problem to automate the manufacturing process if we still need to field test every tablet computer to evaluate potential manufacturing flaws at the unit level.

The AE framework more or less assumes some type of AIG. From the AE perspective, AIG is considered a process for building templates. If properly layered onto cognitively based task models, a natural hierarchy is generated where the task model rather precisely describes what the task measures and why it is at that level of complexity, the template defines how the tasks can be consistently rendered and scored—ideally controlling relevant difficulty and complexity factors, and each [AIG-based] template generates a family of items that ideally perform as statistical isomorphs.

I care about content validity, but, after twenty years of designing automated test assembly systems and dealing with real test assembly problems, I also recognize that there are inherent flaws and inconsistencies in many of our content-based coding taxonomies and overall test specifications. Many content blueprints fail miserably as adequate specifications for item writers to create items at a prescribed level of complexity or difficulty, much less to produce consistent replicates of the tasks over time.

Some of the more interesting research I have been involved with over the past few years involves the development of what I call task-model grammars (TMGs). A TMG is an attempt to *replace* traditional content blueprints with a specification that locates an entire family of items on a scale *by design*. The TMGs are domain-specific and provide very detailed cognitive descriptions of the procedural skills and declarative knowledge components that each family of items is designed to measure. They are not easy to develop, but we are making steady progress in areas of reading comprehension, science and mathematics.

A distribution of task models along the scale—called a task-model map (TMM)—reflects the priorities we place on our proficiency claims and evidence requirements, while simultaneously representing a measurement information target. Having task-model locations in mind via the TMGs and TMMs therefore allows us to develop tolerance specifications for the entire family of items. AIG templates or semi-automated item-writing templates are then designed to maintain that location for purposes of calibration as well as for interpretive purposes. An entire family of items is represented by a single task model. If empirical tryouts demonstrate too much aberrance in the statistical item operating characteristics, the task model and template designs are modified to more strictly control the variation. Imagine generating thousands of items for every task model that individually cost only USD$1.00 to USD$2.50 and that do not have to be pretested.

AIG and AE have a very simple goal. To spend the energy, time and money upfront to engineer robust template design specifications that maintain the [intended] statistical operating characteristics for an entire family of items. Task models ground each template on a well-defined cognitive specification for each family of items that also incorporated core content and skills germane to each domain. That is, different item families have different task models and templates. If a particular item family is working within designated variance tolerances, we may be able to calibrate at the template or even task-model levels. In the worst case, we do what we now do, which is to pretest and calibrate using every item as a unique measurement unit. However, if we can capitalize on solid task-model and template designs that hold tolerance, we may be able to significantly reduce or even eliminate pretesting altogether.

The second challenge reflects a growing consensus across the assessment spectrum that we need to measure higher-order cognitive skills applied to complex content. Unfortunately, we are not always clear about exactly what types of higher-order skills or content we wish to measure or how to best measure our moving-target constructs. Are these new higher skills merely more increased-difficulty tasks that we want to add at the upper end an existing scale? Or do we want to measure a completely different constellation of skills and knowledge? There are some very serious dimensionality issues at play that ought to be worked out before we start writing items or even designing test specifications. An ECD- or AE-oriented approach argues that we need to build construct maps to reflect that proficiency claims as a set of ordered design specifications, complete with tangible evidence-based expectations, and then develop task models and templates to achieve that design. For example, it is entirely possible—as my friends and colleagues at the AICPA have shown—to engineer templates for even complex accounting simulations. Their work over the past decade is beyond mere proof of concept and helps demonstrate that AE has promise if the proper resources are invested under a forward-thinking versioning strategy to iteratively build, evaluate and improve the task model and template designs. It starts with a concrete design, not vague standards and expectations.

The final challenge reflects the unrealistic demands that our clients, legislators, administrators, and the lay public expect from any single test score. As my family and many of my friends know, I love collecting (and sometimes even using) woodworking tools. I also realize that each tool as a purpose. A summative, essentially unidimensional scale constructed by selecting 60 items for one or two test forms that meet an average item-difficulty target and a two-way content-by-cognitive level blueprint cannot do very much beyond, perhaps providing general academic proficiency scores for individual students. If we want those same test forms to simultaneously be instructionally sensitive and measure growth, AND to also provide useful, actionable diagnostic information, we will be disappointed—guaranteed! The problem is not that we do not have the right psychometric model(s) or calibration software; the problem is that the measurement information—the data—do not support those multiple purposes. I cannot effectively use my circular saw to drive screws into a wood board. It really is not all that different in testing. Design the tool (scale) for a dedicated purpose. As psychometricians and measurement specialists, we need to take on each purpose and design a dedicated measurement tool. I understand why we do not do that now—costs! But, there appear to be reasonable alternatives to building supposedly omni-purpose tests that are more likely to do nothing well.

How has research in this area contributed to the feasibility of this approach?

It is perhaps a bit melodramatic to say that we are at the edge of a new frontier, but we are. This, folks, is what a paradigm shift feels like. You cannot see it or feel it directly, but you can sense it happening through the measurement research community. Unfortunately, it will NOT be a glorious march forward into that unknown. It will take money, time and enormous amounts of research.

Our AE-related research to date has been in many cases crude and limited in scope. But, we are learning from mistakes and building on our successes, no matter how minor, all the while keeping in mind the intent to build a particular measurement scale for a dedicated purpose. Perhaps the biggest challenge lies in convincing organizations that we need more nimble ways to carry out multiple phases of empirical research. We need real examinees to experimentally evaluate our hypotheses. Adding a few prototype items to pretest slots on an end-of-year summative test is not adequate. In my opinion, we need to approach assessment design research with the same rigor and commitment that pharmaceutical and biotech firms use to develop medicines, therapies and medical technologies. The investment upfront is worth it in the end, especially if we can eventually automate the production procedures, lower costs and reduce item and test development aberrations.

A brief story from history may help illustrate my point. I recently visited the Wright Brothers National Memorial in North Carolina. I came away awed and inspired by Orville and Wilber Wright, two self-taught engineers who changed the course of human history by a series of human-controlled, engine-powered flights on December 17, 1903. They knew about Daniel Bernoulli's law of hydrodynamic pressure, published 165 years earlier, and took advantage of a century of hot-air balloon research, the development of glider wings, and wind tunnel technologies that scientists had invented to study aerodynamics. But ultimately, it was their belief in what they were doing and a long-term dedication to hands-on research, experimentation, and empirically informed iterative design principles that led to the birth of modern aviation on the dunes near Kitty Hawk.

That same level of dedication seems needed as we move toward evidence-based, principled design in assessment. We can certainly take advantage of over a century of mathematical and statistical theory and model-based research to guide the development of hopefully useful measurement scales. But, statistical theories of test scores are not sufficient to generate valid explanations of how or why examinees respond the way they do—any more than Daniel Bernoulli's work on fluid dynamics formulas magically showed him how to design a wing capable of flight in the 1730's. We need a paradigm shift in measurement that moves beyond content-validity arguments and statistically driven approaches to scale development. We need an empirically based science of test design for specific measurement purposes that further incorporates carefully engineered implementation solutions and strong quality control mechanisms to ensure that the intentional design is achieved. It is no longer sufficient to develop increasingly more sophisticated statistical models and then build scales by simply throwing massive amounts of the same old data at the model—even with a passing nod to content validity. Nor can we rely on achievement level descriptors or other *ad hoc* score interpretations to a add semblance of substantive meaning to the scores, long after we have made critical design decisions about how the items are actually written and why they are ultimately included on each test form. Concrete, useful and actionable meaning needs to be designed into the score scale from the onset.

I have been fortunate to work with some excellent people at The College Board, the AICPA, the DMDC and HUMRRO on various smaller scale, proof-of-concept applications of AE to traditional and novel item types in domains ranging from mathematics and biological sciences to accounting. Some of this work has started with building task model grammars (TMGs) from practice or domain analysis statements; other work has involved basically reverse-engineer existing items to "jump start" the TMG development. It is tedious and time-consuming, but fascinating to see the grammars emerge. Our next steps will be to develop template-based rendering components, data structures, scoring evaluators and software tools that will ideally lead to lower-cost, replicable and scalable item families that behave within established statistical tolerances as intended. Fortunately, as already mentioned, we can draw on some excellent AIG research to-date.

Psychometrically, we actually have a pretty good idea as to how to begin developing quality control statistics and using hierarchical models to calibrate task models and templates. Some excellent work by Wim van der Linden, Cees Glas, Heneke Geerlings, and Sandip Sinharay has provided a solid theoretical foundation for new approaches to calibration, scale maintenance and score equating. Instead of relying on strictly lateral connectivity among data sets (e.g., common items or persons within and across limited numbers of test administrations) these statistically sophisticated approaches taks advantage of Bayesian-oriented, hierarchical connectivity within and between item families. To quote Oscar Goldman in the 1970s television series, *The Six Million Dollar Man,* "We have the technology."

It is important to realize, however, that AE requires a fundamental restructuring of the manner by which we design, develop and evaluate assessment tasks and ultimately, the quality of our measurement scales. It is not about recoding existing item banks using more cognitively oriented taxonomies—it is about redesigning those item banks to provide tangible evidence to justify proficiency-based claims along a scale. The transition to the next generation of test design and development will not be easy or inexpensive—at least not in the beginning. It is also unlikely that we can shortcut or circumvent the important research and development steps that lie ahead. More likely, like the Wright brothers, success will demand many painstaking years of hands-on hard work and empirically driven refinements. We are either committed to change or we will be forced to stay with the *status quo* of producing limited quantities of high-cost items for limited-engagement testing (e.g., annual summative exams) and further have to depend on data-hungry equating and calibration models to generate and maintain serviceable score scales over time.

What in "classical" test development should/have be retained?

Statistical indices will always have a place in measurement. However, rather than being the primary source of information for determining the properties of the scale (e.g., extracting the strongest statistical signal from a factor analysis, considering that the "true score scale" of interest and then retaining items that correlate highly with that factor or a total score based on the factor), item-total correlations, item difficulty indices, and patterns of residual covariances can be effectively used in a strong quality control system to signal aberrant patterns of variance or covariance that might require engineering re-design interventions at the level of templates or database controls used in generating items.

The roles of subject-matter experts (SMEs) and item writers will likewise change. Working with cognitive specialists and item designers, SMEs will be asked to better articulate what they want to measure—the proficiency-based claims and acceptable performance evidence for making those claims. The how (item types, scoring procedures) can be designed around those needs via cognitively oriented task models and templates, informed by empirical research and engineering trials. We do not yet have a clear picture what an item designer for AE or AIG will need to do, any more than we know what a "construct mapper" will do. But, we know what needs to be done. What is clear is that the days of creative aberrance in item writing and test design are numbered if we seriously want to meet the challenges the future.

What research areas would you suggest for practitioners, especially new researchers?

It will take time to enact our *paradigm shift*. It will also be expensive; change almost always is expensive in the beginning. But in addition to just committing research funds, organizations need to understand that empirically informed iterative design is an integral part of assessment engineering. For researchers, the amount of research needed for AE-related test design and development is enormous. I can highlight at least four areas of promising and needed research in the short-term.

The first area is developing practical methods for generating concrete construct maps for multiple-dimensions within a formative assessment setting. How do we envision and document our intended inferences along one or more scales? How can we effectively lay out one or more measurement *schematics* as (a) slice(s) through the domain space? We seem comfortable using phrases like learning progressions, but it is not always clear as to what that means from the perspective of designing one or more scales to show progress. We need researchers who can help us articulate the "*what*" [is expected and what evidence will provide substantive proof of proficiency] as we move from low to medium to high proficiency—but to do so in a manner that recognizes that construct map also directly drives the evidence models, task models, templates and even the psychometric modeling decisions we need to subsequently make. We may need to draw on areas such as program evaluation and qualitative research to help capture and document the vision of each scale—the construct map. If we ever get serious about designing useful formative assessments for integrated instructional use in the classroom, it further seems prudent to point out that we will actually need to design entire *series* of multidimensional scales that allow teachers and students to decide when students may have mastered relevant content, as well as t identify and remediate those who have not. What happens if the dimensionality of these measurement slices through a domain space is actually intended to change as the students progress from one ordered learning state to another along a defined and well-articulated learning progression?

A second area of important research is in developing task model grammars or other ways to represent content in a cognitively relevant way that maps the intended complexity of item and task families to the ordered evidence required as we progress along a scale—that is, the tasks that evidence about the proficiency claims in the construct map. This research can range from reverse-engineering items to using read-aloud protocols to document and capture the problem-solving steps needed to solve prototype items. There has been promising research in this area, but we need more research to better understand the notion of intentional cognitive complexity as we progress along a scale. For the naysayers, I like to point out that video game designers appear to have mastered the science of *designing for intended complexity*. Perhaps we can, too.

AIG is a third important area for more research. I already mentioned some of the work that has been done, much of which has been documented in conference papers, technical reports, journal articles, and recently published book, *Automatic Item Generation*, edited by Mark Gierl and Tom Haladyna. When I said earlier that we *know* how to automatically generate items, however, that does not imply that we know everything or even what else is possible.

Finally, for the mathematical statisticians in some of us, we need effective statistical quality control mechanisms that tell use when the templates are not functioning within tolerance. We need to understand the relationship between task-model or template design features and extraneous variation or covariation in responses or residuals so that the design or data-based controls can be modified to reduce or eliminate it. We need configurable hierarchical calibration engines that let users calibrate ANY type of data, with multiple latent traits or categorical classes, simultaneously using a task model, template or item as the primary unit of analysis. We also need effective mechanisms for calibrating and linking multidimensional metrics over time.

To close, there is a lot of work for everyone. I certainly do not have all of the answers. And, I realize that there is a healthy skepticism out there expressed by some who question whether any of this is possible. I believe that it is not only possible but inevitable. It is more a matter of when and how long it will take for us to realize that.

THE FUTURE OF NCME: MEMBER PERSPECTIVES, PART 4

Later this month we will be celebrating NCME's 75th Anniversary. As we reflect on the lessons we have learned in the past we are also inclined to think about our future. In the past three issues of the NCME Newsletter, we have included a series featuring member perspectives on future of NCME. In this issue, we present additional perspectives from Derek Briggs and Marianne Perie.

Are We Learning From the Past?

Derek C. Briggs, University of Colorado

In the 2012 Volume 20(3) issue of the NCME Newsletter, Leslie Keng and Laurie Davis provided a perspective on the future of NCME entitled "Expanding the 'M' in NCME." In their essay they argue, in a nutshell, that the demand for new and improved large-scale assessments can only be met if psychometricians respond to this demand by expanding their knowledge and skills beyond the "traditional" topics found in a measurement textbook: reliability, validity, equating, and scaling. Although I agree with the spirit of most of what they have to say, I am going to pick on (and potentially even misinterpret) one aspect of their thesis if only for the reason that it gives me an excuse to stand on a rhetorical soapbox.

In my view, before we call on psychometricians to expand their knowledge and skills *beyond* traditional concepts in psychometric theory, I would argue that as much or more could be gained by having psychometricians revisit these core concepts with an eye toward understanding them at a deeper level. In the same issue of the NCME Newsletter in which Keng and Davis's essay appeared, I was struck by some of the responses Wim van der Linden gave as part of his interview with Jerome Clauser. Wim was asked "How do you foresee the field changing over the next decade?". His response was so good that it is worth repeating word for word:

> I might be wrong, but for the next decade or so I don't foresee any revolution or major change of paradigm. Instead, I expect further technical sophistication, both of our research and applications. More of the same thus, but introduced at a much faster speed and higher level of complexity. Am I happy with this? Yes and no. I'm certainly supportive of technical perfection and will keep trying to contribute to it. *But the real danger exists in the neglect of our conceptual basis* [emphasis added]. I'm sometimes shocked by the blind applications of techniques I see. The NCME annual meetings have definitely shown a trend toward more papers with minor technical tweaks. The same holds for our journal articles. If you ask their authors for a motivation, you hardly get an answer. Some of our research areas, I believe, would definitely benefit from a conceptual cleanup. We need technical progress to find solutions to practical problems. But each good solution begins with a correct conceptualization.

To this I can only add "Hear, hear!" I have had the same reaction at many recent NCME conferences when it appears to me that some of my fellow psychometricians act as little more than glorified "test technicians." A test technician knows exactly how to apply textbook recipes for the purpose of maximizing information functions, estimating linking constants, and conducting DIF analyses. We need people who can do these things, to be sure. But I worry that we do not have enough psychometricians who are capable of also wearing the hat of an "assessment engineer" (a term I am borrowing from my friend Ric Luecht). For an assessment engineer, the front-end design of an assessment is just as important as the back-end analysis of item response data. In fact, without clarity about the aims of the front-end design, the back-end analysis will often be misguided. Being a capable engineer means that you can see the big picture about the intended interpretations and uses of an assessment system. It means being able to specify claims and assertions about test scores and their transformations that are testable (ideally, falsifiable) via empirical analysis. Yet it is impossible to be more than a technician if you are incapable of being meta-cognitive about *why* you have decided to apply a technical solution to a given testing problem in the first place.

Over the past 10 years, I have had the luxury of working in academia as a professor at the University of Colorado. One of my responsibilities has been to teach courses in psychometrics to graduate students. Each time that I teach these courses I try to deepen my understanding of foundational concepts. Like Wim, I have about 10 to 15 books and articles that I find myself constantly revisiting, and each time I do I come away with new insights that I try to use to strengthen the way that I think about psychometrics and teach it to my students. Here are some examples of foundational questions that I have wrestled with in the years <u>following</u> my graduate training:

- How do I define the term "measurement?" How do others define measurement? Why might our definitions be different?
- To what extent is measurement in the social sciences different from measurement as practiced in the physical sciences?
- Where does measurement "error" come from?
- What is the difference between reliability and generalizability?
- Are generalizability theory and item response theory compatible?
- If all models are wrong but some are useful, who establishes the criteria for usefulness?
- If vertical scales do not have equal-interval properties, what is the point of having them at all?
- If Frederic Lord and Georg Rasch had met and gotten into a fistfight, who would have won?

The conceptual basis for psychometric research is not something that can easily be picked up while reading chapters in a textbook as a graduate student and then stored away for future reference. Indeed, there is no book out there that will give you straightforward answers to the questions above, because these questions are representative of some fundamental disagreements in the field! I don't find that troubling at all; what would worry me is having a preponderance of psychometricians in the NCME community who think that these disagreements have all been settled or that they would be irrelevant for practitioners. An even greater worry is about those who are unaware that the disagreements exist at all.

In looking forward to the next 25 years of NCME it is important that we take the time to appreciate that many of the psychometric issues that seem so new (e.g., assessments for the "next generation") are, at heart, variants of the same basic issues that NCME members have been debating for decades. Consider the following quote:

> State-administered tests are not closely tied to student learning. Those who believe that external tests will have a salutary effect—by focusing instruction, increasing student motivation, and providing valuable feedback—base their claims on a classroom-level model of teaching and learning. There is no evidence that a shift to the system level will preserve the psychological relations in the model. There is no basis for the belief that an external test can act the part of the stern but loving teacher who exhorted you to learn in your youth. At the very least, if the state replaces the teacher in checking for student learning, the role of teacher is unknown. Surely in trying to anticipate what the effects of centralized testing programs will be, it is more reasonable to invoke our understanding of the sociology of school systems and the workings of both political and fiscal incentives than a psychological model of learning that takes place between individual student and teacher.

This quote, if found in an article published this year, would be entirely relevant to the current policy climate around the use of large-scale assessments for high-stakes accountability. It comes from a commentary written by my colleague Lorrie Shepard for the journal *Educational Measurement: Issues & Practice*. The year? 1985.

The computers have gotten faster, the software more sophisticated, and theories of learning are more elaborate, but in the end we're all still trying to figure out the best way to characterize and draw inferences about the interaction between an assessor and assessee, a student and a test item, an observed score and a latent variable. And those who refuse to learn from the past are doomed to repeat it.

When Testing as the Solution Leads to Testing as the Problem

Marianne Perie, Center for Education Testing and Evaluation, University of Kansas

As we near the 75th Anniversary of NCME, it is interesting to look back on what has changed over the years. Certainly one of the biggest changes has been the growth of the measurement field and of testing in general. Testing has been used in American schools at least since the 19th century, when they were used to determine if students had mastered what they had been taught (Ravitch, 2002). These tests had the student level consequences of passing or failing a class. Teachers, too, had to pass a test to become certified, but then were never tested again.

Certification testing persists today with similar goals with which it began: to determine if an individual has the requisite knowledge and skills to hold a particular title or job. Student testing, however, has expanded enormously, largely due to social and political influences. In the past 40 years large-scale student testing has moved from norm-referenced to criterion-referenced. The accountability movement that began in the 1960s continues to expand and has brought a new focus on cut scores. The most recent reauthorization of the Elementary and Secondary Education Act, known as *No Child Left Behind*, greatly expanded the role of student testing, focused attention on "proficiency", and held schools and districts accountability for student performance.

During that time, many measurement professionals expressed concern about the disconnect between the purposes and design of the assessments. For example, several authors (e.g., Childs & Jaciw, 2003) argued that for the purpose of school accountability, the tests did not need to be given to every child every year. Instead a matrix sampling approach could provide stable results and decrease both testing time and expense. Likewise, measurement experts at the Center for Assessment wrote about and presented on the reliability of school-level test scores and implications for making accountability determinations at the school level (Hill, 2002; Hill & Depascale, 2002). Yet, none of these scholarly works had an impact on the implementation of the law.

More recently, prior to the release of the Notice of Intent to Award funds to develop cross-state assessments aligned to the Common Core State Standards, the U. S. Department of Education held a series of open meetings and invited members of the measurement community and others to provide advice. They received written testimony from over 60 members of the public, the majority of whom were in the testing community and members of NCME. In oral testimony, a predominant theme was to use the Race to the Top funds to test out new concepts in multiple, smaller multi-state consortia. For example, one group of states could test the use of innovative technology-enhanced items, while another could test the effectiveness of performance assessments embedded in curricular units. Other speakers focused on the need to assess all students fairly, including English language learners and students with disabilities. While all the advice appeared to be well-received, ultimately, the Department of Education released their notice of intent to award funding to two large, multi-state consortia to fully develop assessments ready to be operational that included all the innovation and fairness discussed in the testimony. In other words, they missed the point distinguishing researching innovation from operational testing.

So, as we get ready to realize the consortia vision of multi-state testing, how do we ensure that the tests—currently seen as the solution to the problem of comparability across states—do not instead become a bigger problem?

I see the issue as one primarily related to validity. A test cannot be valid in and of itself. Scores have validity for a *particular use or purpose*. The biggest disconnect between the measurement community and the policy world has occurred when a test is built for one purpose and then used for another. For example, most tests built in the NCLB era were constructed to clearly define proficient versus non-proficient performance, meaning the power of the assessment was centralized around that cut score. Those same tests are now being used to measure growth across years, even though there is often very little information about student performance at the high or low ends of the assessment. Tests were built knowing that scores would be used in the aggregate to evaluate schools; those same scores are now used to evaluate teacher effectiveness.

When asked how the new consortia assessments will be used, there is often a plethora of purposes listed: to track student learning toward college-and-career readiness, to measure understanding of the common core state standards, to track growth over time, to evaluate teacher effectiveness, to evaluate school and district progress over time. As my colleague at the Center for Assessment, Scott Marion, is fond of saying, "Tests that purport to serve too many purposes, serve none well." So what do we, as measurement professionals, do?

I believe, first and foremost, that we have a responsibility to lead the validity evaluation of all of these purposes and uses. Validity evaluations do not need to be done—and often should not be done—solely by those developing the test. Those developing the tests certainly have a responsibility to validate the scores for the stated purposes and uses, but all of us in the field have a responsibility to develop studies to verify the claims made, particularly if we are using the data from any of those assessments.

I believe the future of NCME should be driven by education. We have a responsibility to educate both the future generation of measurement professionals and the end users of the assessments. Measurement programs from 20 years ago will not serve graduate students well today. More emphasis needs to be given to understanding all aspects of validity and how to evaluate the utility of scores generated by the assessments. Future measurement professionals need to know that one size of assessment does not fit all uses and be taught to ask first, how will this test be used? The design then follows that purpose and the test is validated for that purpose. A new purpose or use requires a new validation, and perhaps a new assessment. We also need to educate future classroom teachers and school administrators on interpreting test results and, most importantly, understanding their limitations. Finally, we need to make inroads with policymakers that put NCME in the position of driving better testing policy as opposed to reacting to policy handed down from above. The timing is crucial on this education, as the past has shown us that when policymakers have their minds made up, measurement professionals may be invited in only as a courtesy, not as an influence.

References

Childs, R. A., & Jaciw, A. P. (2003). Matrix Sampling of Test Items. ERIC Digest.

Hill, R. (2002). Examining the Reliability of Accountability Systems. Presented at the Annual AERA meeting in New Orleans, LA, April 2002.

Hill, R. & Depascale, C. (2002). Determining the Reliability of School Scores. Dover, NH: Center for Assessment. Retrieved March 14, 2012 from www.nciea.org.

Ravitch, D. (2002). Testing and Accountability, Historically Considered in Williamson M. Evers and Herbert J. Walberg (eds.) *School Accountability*. Stanford, CA: Hoover Press

TRAINING SESSIONS AT 2013 NCME ANNUAL CONFERENCE

Leslie Keng & Ye Tong, Pearson

The NCME Training and Professional Development Committee is excited about the two days of training sessions for researchers and practitioners that will be offered during the 2013 NCME Annual Meeting. This year's program includes 19 sessions on a variety of topics including growth modeling, adaptive testing, generalizability theory, vertical scaling, item response theory (IRT) software, and diagnostic measurement. This year, we are continuing or expanding on the following recent additions to the program:

- Based on positive feedback from the past three years, we will once again be webcasting live training sessions. Four of the training sessions will be webcast live to over 25 international sites.
- In line with this year's conference theme, "Building on the Past, Reaching for the Future: Innovative Ideas to Design Next Gen Assessments", several first-time sessions that focus on next generation topics and/or offered in innovative formats were chosen. Examples of such sessions include:
    - *Crash Course in Hot Psychometric Topics*. Four experts in areas of performance assessments, automated scoring, college and career readiness, and growth and value-added modeling will provide level information, direction, and research summaries on their respective topics. This provides a nice professional development opportunity for aspiring and long-time measurement professionals.
    - *Using Visual Displays to Inform Assessment Design and Development*. This workshop focuses on an increasingly important but understudied area in our field—the graphical displays of assessment information. The session will illustrate how visual displays can help inform steps of the test development and validation process, from program design to item writing and review to communicating results through score reporting.
    - *NCME's Got Talent! How to Write, Present, and Tweet Like a Star*. This training session will help participants develop a comprehensive plan for communicating their measurement-related work through presentations, publications, social media, and working with reporters.

The 19 sessions were chosen by the committee from 28 proposed and invited sessions. Successful proposals were ones on which the committee judged that the topic is important to measurement theory or practice in educational settings, that the presenters were highly qualified for the session, that had high enrollments and positive reviews when previously offered by NCME, that the topic was desired by past sessions' attendees, and that the presenters planned on including hands-on or engaging activities during the training session.

The complete listing of training sessions appears below. The sessions designated with an asterisk (*) are those that will be webcast live to international sites. You can find more information about and register for the sessions at: http://ncme.org/annual-meeting/next-meeting/

- Crash Course in Hot Psychometric Topics*
- Using Visual Displays to Inform Assessment Design and Development*
- An Introduction to R for Teachers of Quantitative Methods*
- An Overview of Psychometric Work at Testing Organizations*
- NCME's Got Talent! How to Write, Present, and Tweet Like a Star
- Multidimensional Item Response Theory: Theory and Applications Using BMIRT, LinkMIRT, and SimuMIRT software
- Language in Assessment—Approaches for Distinguishing Between and Addressing Construct- Irrelevant and Construct-Relevant Language
- IRT-Based Test Linking in R
- Diagnostic Measurement: Theory, Methods, and Applications
- Generalizability Theory and Applications
- Vertical Scaling Methodologies, Applications, and Research
- Setting Cut Scores on 21st Century Tests
- Bayesian Networks in Educational Assessment
- Advice for Graduate Students: Making the Most of Grad School, Obtaining a Job, and Starting a Career
- Introduction to the Multidimensional Adaptive Testing Environment: Test Specification, Simulation Studies, and Operational Testing
- A Practitioner's Guide to Growth Models
- An Introduction to the Measurement and Analysis of Video Game Interaction Data
- Item Response Theory Linking and Equating with jMetrik
- Applications of Evidence-centered Design (ECD) in Large-scale Assessment

NCME WEBSITE FEATURES AND FUNCTIONALITY: RSS FEEDS

J. Patrick Meyer, University of Virginia

The new NCME website is more than just a new design. It includes a number of new features and tools that members will find useful. Among the new features is the ability of the website to publish RSS feeds for organization news and messages. RSS feeds are a form of website syndication that allows you to receive news and other information as soon as it is published online. Through a feed reader you can organize RSS feeds from multiple sources and have all of your latest news and information in a single location.

Why should you care about RSS feeds? Blogs, news websites, and scholarly journals use RSS feeds to broadcast breaking news and recent publications. I have found RSS feeds particularly useful for keeping track of the latest publications in scholarly journals. You not only receive access to article titles and abstracts as soon as they are published online but you can get them all in one place. You no longer have to visit each journal separately to see the latest publications. You can see them all, organized by topic, in a feed reader. The new NCME website makes it easy for you to subscribe to RSS feeds for the *Journal of Educational Measurement* and *Educational Measurement: Issues and Practice*. It also provides two news-related feeds for receiving the latest information and announcements from NCME.

To find links to NCME RSS feeds, go to the website and click "News" in the main menu. There is a link to the RSS feed page in the opening paragraph. You can also find the RSS feed page by directing your web browser to http://ncme.org/news/ncme-rss-feeds. Links to four RSS feeds are listed on that page: (a) Main NCME news, (b) President's message, (c) *Journal of Educational Measurement*, and (d) *Educational Measurement: Issues and Practice*. You must subscribe to each feed separately, and there are various ways to subscribe. You will find it best to use a feed reader to subscribe and organize multiple feeds.

You can choose from a variety of software when it comes to managing feeds but I recommend using one that has both a web-based interface and a smartphone app. That way, you can read RSS feeds while sitting at your desk at work or while waiting in line at the grocery store. I have always used Google Reader to manage my feeds, but Google will discontinue the software on July 1, 2013. In search of a new feed reader, I have found Pulse and Feedly to be my favorites. They are both well-designed applications that allow you to easily import feeds from Google Reader, which is a plus for anyone who will soon need to migrate to a new reader.

Pulse (https://www.pulse.me/) is a popular option for managing feeds. It has a magazine-like appearance with a simple user interface. Pulse is available as a web browser plugin and iOS or Android smart phone app. You can register for Pulse and install it online or install it directly on your smart phone through the Apple App Store or Google Play. Pulse allows you to subscribe to RSS feeds by using their search feature, adding the RSS feed link directly, or importing feeds from Google Reader. Pulse organizes feeds by page. For example, I have a page for "Journals" and it contains all of my academic journal feeds. I have another page for "News" and another for "Sports." When you subscribe to a feed, you add it to an existing page or create a new one. I first began using Pulse as an app for my Android-based smart phone and now that I have discovered the web browser plug-in I am even more of a fan.

Feedly (www.Feedly.com) is another popular reader that provides a nice magazine-like interface to your feeds. It has many of the same features as pulse—a web browser plugin and smart phone app for iOS and Android. Feedly has the added benefit of an app for Kindle. You can install Feedly online through the web browser plugin or as an app from the Apple App Store or Google Play. If installing online, the Feedly website will identify the make of your web browser and offer you different options accordingly. I used Google Chrome to register online and then clicked a link to have the app installed on my Android-based smart phone.

Pulse and Feedly appear to be equally good feed readers in terms of their features. One notable difference is the ease by which you can add a feed for NCME journals. I used the same method of adding a feed in each program, but Pulse did a better job of finding the feed for the *Journal of Educational Measurement* than did Feedly. In fact, it took me several attempts to add the *JEM* feed to Feedly.

After you select a feed reader, you can add the NCME feeds via a web browser using these steps:

1. Go to Pulse or Feedly website and login to your account.
2. Go to http://ncme.org/news/ncme-rss-feeds/ to find the feed URLs.
3. Right click on one of the feed links and select "Copy link address" (Google Chrome) or "Copy shortcut" (Internet Explorer).

4. In Pulse, click the “Add Content” button and paste the feed URL into the search bar (CTRL+V on a PC). Click the “+” sign to add the RSS feed. In Feedly, click “Add Website” and paste the feed URL into the search bar. You may have to remove the leading http:// from the feed link in order for Feedly to find the right feed. Click the “+” sign to add one of the search results to your reader.
5. Repeat steps 2 through 4 for each NCME RSS feed.

Please forgive my PC-centric instructions, but I hope they contain enough information that Mac users can easily follow them too.

The steps for adding RSS feeds via a smart phone app are similar but you will not be able to copy and paste the URL from the NCME website. Rather, use your smart phone’s web browser to navigate to the NCME website. Go to the News page, click the link for the NCME RSS feed page, and then click one of the RSS feed links. Your smart phone may then give you shortcuts for adding the feed to a feed reader. My Android phone currently attempts to add the link to Google Reader, but this option will soon be obsolete. You may find it easiest to use a web browser to subscribe to NCME feeds because you can get the exact feed link from the NCME website and there is no reason to sort through a long list of search results. Moreover, any subscription you make through the Pulse or Feedly website will automatically sync with your smart phone. The opposite is true as well. Subscriptions made through the smart phone will sync with the browser plugins. Once you have subscribed to all of the NCME feeds, don’t forget to add feeds from your favorite blog, CNN, the *Wall Street Journal*, and other websites. You will then be able to read the latest news and journal offerings in one place.

OTHER CONFERENCES OF INTEREST

CASMA Summer Equating Workshop, June 24 – 28, 2013 --- Iowa City, IA

A more intensive and extensive workshop than a one-day session on equating will be offered by CASMA, June 24-28, 2013. Four similar workshops were held previously. They were well received and well attended. For the 2013 workshop, attendance will be limited to 18 persons. If registration is too low to justify holding the workshop, fees will be returned. No workshop can replace a full-length course, but this particular workshop should provide participants with a good working knowledge of basic equating designs, statistical procedures, and applications.

Visit the workshop web page for details: http://www.education.uiowa.edu/centers/casma/conferences/2013-equating-workshop. Should you have any procedural, housing, or registration questions, contact Jennifer Jones at 319-335-5439. Workshop content questions can be directed to Bob Brennan, University of Iowa, at 319-335-5405.

To get the NCME Newsletter four times a year (March, June, September, and December) go to
http://ncme.org/publications/newsletter/

Send articles or information for this newsletter to:

Susan Davis-Becker
Alpine Testing Solutions

Phone: 402-483-5898
e-mail: susan.davisbecker@alpinetesting.com