From: NCME Task Force on Foundational Competencies in Educational Measurement

To: NCME members providing public comment on the September 2022 draft report

Re: Task Force response to comments from NCME members

Date: March 2023

The NCME Task Force on Foundational Competencies in Educational Measurement is grateful for the engagement and comments from NCME membership on the September 2022 draft report. We have revised the report in response to your feedback.

We have collected your feedback in the following pages and appended our brief responses in **bold**. We look forward to continuing to engage with you and NCME membership at the Annual Meeting this April and beyond.

With respect and appreciation,

Terry Ackerman, University of Iowa; University of North Carolina Greensboro, emeritus
Debbi Bandalos, James Madison University
Derek Briggs, University of Colorado Boulder (ex officio)
Howard Everson, SRI International and CUNY
Andrew Ho, Harvard University (Chair)
Sue Lottridge, Cambium Assessment, Inc.
Matthew Madison, University of Georgia
Sandip Sinharay, Educational Testing Service
Michael Rodriguez, University of Minnesota
Michael Russell, Boston College
Alina Von Davier, Duolingo
Stefanie Wind, University of Alabama

From: Hao, Jiangang <jhao@ets.org>
Sent: Friday, September 16, 2022 6:36 PM
To: NCME <ncme@talley.com>
Cc: Alina von Davier <avondavier@duolingo.com>
Subject: Re: NCME Task Force Draft Report for Release (Friday)

Hi,

Alina contacted me for feedback on the report on foundational competencies as listed below.

https://higherlogicdownload.s3.amazonaws.com/NCME/4b7590fc-3903-444d-b89d-c45b7fa3da3f/UploadedImages/NCME_Task_Force_on_Foundational_Competencies_-_Draft_Report_Release_9_16_21.pdf

I am sharing some thoughts after reading the report. A major observation I noticed is that there seems to be a lack of discussion of skills related to data science, machine learning/AI, and natural language processing (NLP). In a world of digital learning and assessment, the next generation psychometric researchers really need to have these skills to survive and thrive in their future career.

Over the past few years, we (with Alina and many others) all went through some painstaking efforts to hire people with the right combination of skills to meet the challenge of digital learning and assessment. We observed that many applicants from psychometrics programs do not have the needed data science/machine learning/NLP skills (and mindsets) to process and model complex data from digital tasks. In contrast, applicants with data science/machine learning skills from other disciplines, such as computer science, generally know very little about the core values of psychometrics and measurement. In practice, hiring people who do not know the core values of the substantive area poses a big retention challenge for organizations, as they may quickly move on if they find they are not interested in the area at all after a few months. Therefore, we feel it is imperative to prioritize a set of new methodologies and integrate them with the core values of psychometrics in a principled manner to help prepare a stable workforce for digital learning and assessment in the future. As such, Alina, Bob and I edited a volume, computational psychometrics (https://link.springer.com/book/10.1007/978-3-030-74394-9), by which we tried to carve out a suitable subset of these skills to help psychometrics researchers/students to getting them. Early this year, in Derek's NCME presidential address, he nicely summarized four aspects of measurement, and computational psychometrics is one of them.

So, if the report aims at laying out the foundational competencies needed for the future generation of psychometrics researchers, I would strongly suggest including these new skills. I hope this feedback could be helpful. Thanks.

The Task Force agrees that these skills are important and appreciate that you noticed that we describe the foundations for these skills in Domain 2: Technical, Statistical, and Computational Competencies. We edited the description to be clearer that these are foundational competencies for more advanced skills like AI and NLP. We also edited the section on Theory and Instrumentation to acknowledge how these skills and this field may transform instrumentation in the years to come.

Robert C. Shaw, Jr., PhD
Senior Vice President, Examinations
Direct: 913.788.2547
Robert.Shaw@nbrc.org
10801 Mastin Street, Suite 300
Overland Park, KS 66210

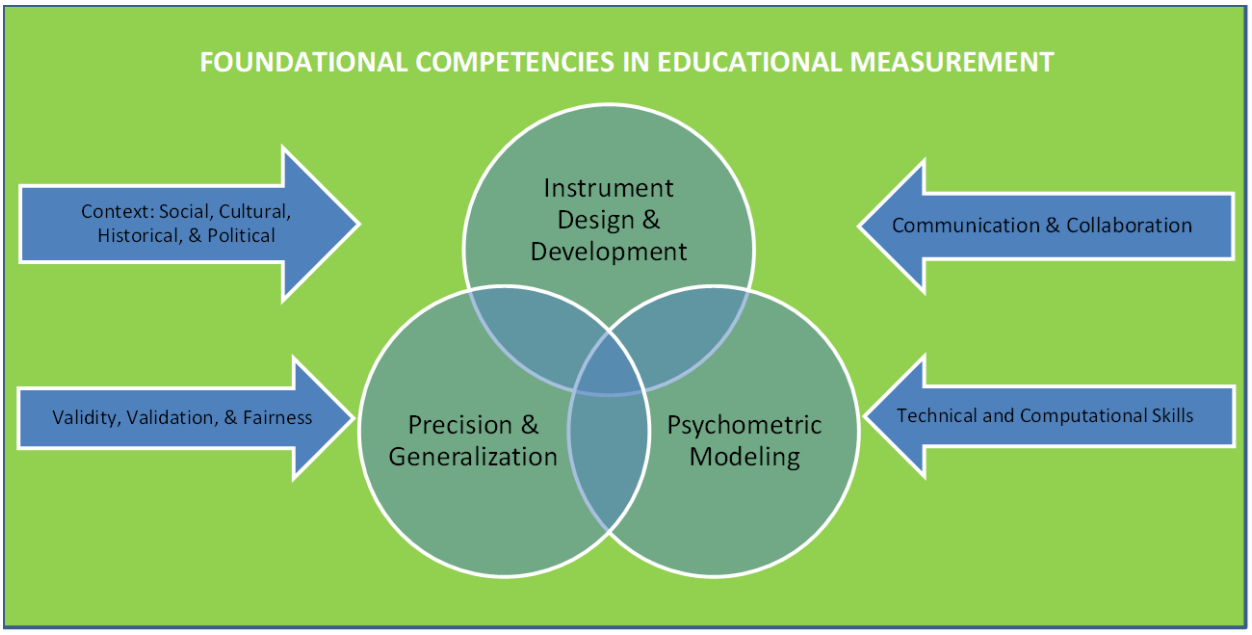
While reviewing content on page 15 of the document within the section called Common Duties of Psychometricians in Licensure and Certification Organizations, I expected to see a bullet point describing the activity in which potential stimuli for examination content are scrutinized, often through survey responses, and examination design specifications are created. If 'standard setting' can be called out in bullet 5, then it seems that 'job analysis' or 'role delineation' can be called out as well. Perhaps such activity is supposed to be among the 'psychometric activities' in bullet 1, 'research issues' in bullet 3, 'presentations at' 'exam related conferences' in bullet 4, or 'other related meetings' in bullet 5, but omission of this specific set of activities seems to leave a significant gap. I suggest adding 'job analysis' before 'standard setting' within bullet 5 as a solution.

The Task Force agrees that this is important. We have added this duty to Example 3.

Hi Derek,

Thanks for including me in the reviewer group. The draft report is excellent, very well-written, and should be extremely valuable to graduate programs in trying to optimize success for their students. Similarly, I imagine perspective and current students would also benefit from this summary, even touches on the old 'academic vs industry' career path dilemma. Just a few comments/suggestions noted below:

1) I was wondering if the graphic on pages 1 and 6 could be simplified through a little restructuring. Given the focus is on educational measurement competencies, maybe it could be combined as such where those 3 primary overlapping domains are in the center and being influenced by the slightly broader competencies/skills/considerations that are somewhat less specific to our field. Below is something I quickly put together to illustrate.



The Task Force appreciates this alternative framework and acknowledges that the current figure is complex. However, members prefer a visual that acknowledges that Technical, Computational, Communication, and Collaboration competencies exist outside of educational measurement and that context and validation are overarching and undergirding. We have attempted to improve the graphic for clarity, however.

2) I think the label 'Validity, Validation, and Fairness needs a bit more clarity to help the reader understand the difference between validity and validation. Are they really different? If so I think that's important to articulate. If not, then maybe just validity and fairness?
The Task Force appreciates this point. Members wished to emphasize both the activities we associate with validation and the concept of validity. We attempt to distinguish these on page 12, "…an activity known as validation."

3) Page 13 refers to figures 1 – 5, but I didn't see those figures in the document.
4) Page 14 has 2 example 1's. You'll see
Thank you for catching that these were typos and should have read, for example, "Example 1." We have corrected these.

5) On pages 14-16, where the report details common responsibilities for different careers, I'm wondering if it would be helpful to try and summarize this part as a compare/contrast. For instance, individuals who work in the Educational Measurement field, regardless of career, will likely need to collaborate with non-technical colleagues and stay abreast of newer methodologies and applications. Likewise, a career as a research associate in a non-profit might require more competence in data collection, coding, and analysis, perhaps because you are less likely to work within a team (just guessing here). This could help readers distinguish what's unique about different accessible jobs in the field
The Task Force emphasized that there is healthy overlap among common duties. In response to this comment and others, we also emphasized that the listed duties are illustrative and not exhaustive.

Thanks again for including me and I hope this is helpful

Rich

Richard A. Feinberg, Ph.D.
Senior Psychometrician

T +1 215-590-9553
E rfeinberg@nbme.org
W nbme.org

---

via Terry Ackerman, written by Ric Luecht:

I like the document and its organization. I'm less clear about the intended audience (e.g., faculty at graduate measurement programs?). I may have missed that, but then that should be on page 1: who this is intended for? It's on shakier ground if the intended audience is hiring managers and researchers at testing organizations. I doubt that there is a strong demand for anyone to want to "certify psychometric competencies".
Thank you for this comment. Per our charge listed on page 4, the primary audience is NCME membership and members of the educational measurement field. Past-President Derek Briggs has also provided a foreward clarify the purpose and context.

Beyond that, there were three things I noticed. **First**, I believe that the document should explicitly focus on "entry-level' competencies. Be very explicit here--much like we would when describing the target population for a standard setting study (not the superstars or highly experienced--the competent individuals who can get the job done. Toward that end, you might start by inserting some **narratives** or brief stories of the experiences of a real (entry-level) or fictional psychometricians at some of those types of organizations. This would be in addition to Figures 1-4 to more elaborately illustrate some of the prototypical roles of new hires in K-12 organizations, admissions like LSAC, ACT or College Board/ETS, NBME and various licensure/certification agencies like Amazon and Microsoft, psychological testing organizations, etc.. Listing, clustering and then sorting job description bullets isn't really all that informative, in my opinion. I'm willing to risk saying out loud that most faculty members have no real clue what goes on a testing company re the day-to-day or season operational work, beyond hearing summaries of results from analytical studies (if they happen to be on a TAC). The narratives might help make that real. I'm sure that you could get some volunteers at various companies to write a paragraph describing what they do, how they do it, and (in some cases, the "why" conditions and "when").
The Task Force appreciates the recommendation to include "narratives" and considers this a possible task for a future Task Force or standing committee that continues this work.

Second, I'm not sure what types of ***evidence*** would be needed to help confirm progress or attainment of the more esoteric competencies? Nor is it clear as to the value-added of those competencies, IF one could actually verify them. This is especially the case for the "social, cultural, historical, and political context of measurement". How does knowing the historical context (e.g., going back to Pearson, Binet, or even Cronbach, Lord, Bock and others) make one a more competent psychometrician--other than being well read and possibly realizing that what we often think of as "new" was probably already consider by some pretty smart people in the past? How much of that history makes us "competent"? What evidence would I need to see to support the competency claim? Re "socio-cognitive" aspects of measurement, what and how do we infuse the knowledge and skills our students learn with an on-going

awareness of socio-cognitive threats to validity and reliability. For example, how does the i**nvariance issue** re the GRE you discussed with Brian apply toward recognizing that analytical methods should probably consider as many social and cultural conditions as possible and be willing to empirically show where a model or set of results breakdown or should not be applied (e.g., when residual patterns suggest non-trivial fit issues)?
The Task Force appreciates this point and welcomes next steps beyond its charge that can address this aim.

Third and rather glaring is the rather antiquated view of construct development and test design articulated under "Instrument Design and Development". There is absolutely no mention of ***modern practices of assessment design and test development*** (like ECD, AE, PADI, etc.). I can tell you from personal experience that I get the question once a month as to whether we have ANY students versed in Assessment Engineering (AE)? I'm sure that Bob Mislevy got the same questions at U. of Maryland re ECD. The merger of validation by design, cognitively based task design and modeling, and psychometrics for quality control is becoming the "modern" view of test design. Contrast that with the "old way" of construct design and instrument development: (a) developing blueprints from vague notions of the construct(s) or a practice analysis, (b) writing items from said blueprints and then pilot testing items and keeping items that correlate well with other items; (c) building test forms and then building a statistical scale from the resutls; and (d) spending the rest of our time trying to figure what the numbers mean and how to interpret (usually normatively) various points along the scale). My point is not to be overly critical of the traditional methods*; rather, to point out that entry-level competencies should also point to where the field of measurement seems to be headed (i.e., validity and interpretation baked into the recipe for every item).
Thank you for this point. We have edited to emphasize that these foundational competencies support these modern practices.

Hope this helps.

Ric
* We once "traditionally" thought the world was flat. As Mislevy said circa 1989, "*Modern psychometrics is the application of 20th century statistics to 19th century psychology*."

Dear Derek, Andrew, and Foundational Competencies in Educational Measurement Task Force:

This e-mail is a response to your request for comments about the NCME Foundational Competencies Draft Report. I appreciate the idea and the task force's efforts, and I do believe that such a document will be useful to NCME and to the field. I see that care was taken in the preparation of this document.

I provide two types of comments. First, in the following paragraph I comment in general about how this is not the document I wished (and, frankly, expected) it to be. For that, it would need a major revision—and that's my main first comment. Second, assuming perhaps too pessimistically that such a major revision will not happen, I provide some comments that I think would at least make the draft document better.

First, major comment: Given my own work on the NCME Task Force for Classroom Assessment and especially Derek's (brilliant, in my opinion) theorizing about the domain(s) represented by our field in his Presidential Address (which had a place for classroom assessment in it), I was disappointed to see that

the definition of the field of educational measurement in this document is much less inclusive than what I thought NCME has been aiming for over the past decade. We have tried to encourage membership, for example, from people who do assessment in states or school districts but who would not recognize many of their own competencies in this list. We have tried, as per Mark Wilson's directive, to demonstrate how conventional educational measurement people could learn some things from classroom assessment people, and vice versa—currently neither has an inclusive enough view of what it means to take the pulse of student achievement. Particularly ironic from my point of view is that, while not mentioning classroom assessment or related competencies, when it comes to desired curriculum and program experiences for educational measurement professionals in training, the authors do want high quality classroom assessment for their educational measurement (defined as in this document) students. The document calls for, among other things (p. 19), through-course, end-of-course, and comprehensive assessments, and (later, same page) lots of feedback. Someone is going to have to have the competencies to provide those things. So in a perfect world, I'd like a revised document more inclusive of a rapidly evolving educational measurement field.

The Task Force agrees that assessment warrants greater attention. We have added a definition of assessment as well as acknowledged classroom assessment as an example of a context in which educational measurement can occur.

Second, if we're just revising the draft we have, here are some comments to consider. These tiny changes will at least infuse some "education" in with the "measurement" and, by inference, communicate that educators have a place in educational measurement.

- P. 8, Subdomain A: given that we are talking about educational measurement (that is, measurement for educational purposes and contexts, see your definition on p. 12), I would like to see "educational" added to the list of contexts (social, cultural, historical, political, and educational contexts…). If EM professionals need to know about social, cultural, historical and political contexts (which they do—I'm not contesting that), they should surely know a little something about the context in which students learn the constructs (learning domains however defined, by state standards or otherwise) they are going to be assessed on. What someone knows about something carries within it some aspects of how it was learned.
- P. 14, the first Example 1, Domain 1, bullet 4: add "educators" to the list (Work with educators and members of testing services….teams, etc.). For example, I serve (as does Derek) on the Smarter Balanced TAC, and every meeting includes interface with educators.
- P. 14, the second Example 1, Domain 1, bullet 1: add "educators" (Collaborate with educators and internal teams…) You have certainly, as have I, seen some commercial ed-tech offerings on the internet that have no relation to school and learning as we know it (or, for that matter, in some cases no reference to quality educational measurement either).
- P. 19, Domain 3. Add "and educational" to the title list of contexts, and revise the paragraph description accordingly.
- P. 21, Course 1, Week 1, bullet 2. Add "educational" to the list of contexts.

The Task Force has incorporated many of these changes and reinforced the importance of educational contexts, particularly in Subdomain A.

Thank you for your consideration of these comments.

Best wishes,
Sue

Susan M. Brookhart, Ph.D.

Professor Emerita, Duquesne University
5129 Randall Street
Culver City, CA 90230
406-431-7746
suebrookhart@gmail.com
brookhart@duq.edu
http://susanbrookhart.com

Hi,
I just watched the video from last week's session, and I would like to give a few comments on the work you presented.

First, I would like to thank the work done so far. This is very relevant. One big part of my interest in writing a comment is related to the fact that a definition of foundational competencies will sketch the boundaries of what is our "jurisdiction" both as individual professionals and also for NCME as a professional group.

In this line, I think one problem I have with the document, which a participant of the session also mentioned, is that the domains in the document seem to target what we have been as a field and not necessarily what we want to be in the future. There will be some external influences in what a professional in the field will need to know, however, there is also some space in this initiative to think more proactively so we can shape that future.

I like to think about the future by thinking about sustainability. How sustainable is our field? For how long will our area stay as a field, given the competencies we acknowledge as foundational? I am not sure if this question has appeared in the task force's discussions, but I would suggest you could reflect on that.
Thank you. It is our hope that by identifying foundational competencies we can advance and sustain our field.

One point I am missing in the document and proposal is some section about the ethics of our profession. We have a code of conduct at NCME, some Standards get close to norms, and in general, I think we need professionals capable of reflecting on the history of educational measurement and be prepared to own that legacy and lead us to better practices. I imagine both to know the general principles of our codes of conduct and enough history knowledge plus reflection on it, something I would consider foundational for a sustainable field over time.

This is something I think many programs have missed. If you see any regular certification profession, a chapter about that profession's ethics is usually part of a certification's content domain. It is disturbing, in fact, given the criticisms that educational and psychological testing has received over the years, that the word "ethics" does not appear not even once in the document proposed.

Now, if the purpose is to reflect on what we have been as a field, I think the document fairly represents that. However, this is not enough, in my opinion. As a student soon to graduate, it is sometimes depressing to see all the harsh criticism the field receives. It makes you wonder if you are in the right place or on the right side of history.

So, how did we get there? "we" as all generations in the field before us. The lack of concern and further discussion about the ethics of our profession is part of that.
Thank you for this comment. We believe that our section on Fairness address some of the dimensions of ethics that you highlight. As important as a code of ethics is, it is beyond the scope of our charge to provide one here.

Another concept I am not sure how much you have discussed is the role of math modeling as a competency part of what is foundational. As a mathematician, I have seen a certain rigidity in many students and young professionals in the field when analyzing unexpected situations. I think the field of psychometrics has been the heir of the methods but not of the spirit. The spirit that originated many of the methods a psychometrician usually knows, was the spirit of modeling phenomena relevant to psychology and education.

The inclusion of data sciences in our field, and the displacement the traditional psychometrician will face from that, is fueled by this lack of modeling skills. That is a problem because the future seems to bring much more complex modeling needs, which will require modifying a psychometric model a lot or applying a completely custom model. This tendency to honor the methods rather than the spirit also manifests in a certain tendency to apply psychometric models and statistics everywhere, even when you need the expertise of some other type of models. To know the boundaries of the methods used in any profession is something I would consider foundational.

So far from what I read in the document, the Technical, Statistical, & Computational Competencies part will continue with this legacy of being the heirs of the methods but not creators of new ones. I would suggest that including more core math can help to form creators and not just administrators. Topics like optimization, information theory, measure theory, and stochastic processes can serve pretty well to expand the types of models in the future of our field.

One lens that may be helpful for your task is to think about those situations when you see a professional, and you realize "they do not know that they do not know." Which blind spots are dangerous in an educational measurement professional? From what I mentioned, I see two blind spots not covered yet in your document: professional ethics and modern maths/math modeling.
This is helpful, and although our section on technical, statistical, and computational competencies alludes to this, we also mention creativity in our new Principle 6 on page 6.

Thanks for opening this chance to send our comments! I hope this helps!

Sergio Araneda
PhD student in Research, Educational measurement and Psychometrics (REMP) at Umass Amherst,
Ingeniero Civil Matemático Universidad de Chile
https://calendly.com/sondaxius/30min
https://umass-amherst.zoom.us/j/6144481969

Michael Peabody <michael.peabody77@gmail.com>

My primary comment mostly addresses the process. The group formulated to address this issue is primarily composed of faculty members and there is no representation from the certification & licensure

community. If there were representation from the certification & licensure community, I'm sure it would have been noted that the activity of determining competencies is essentially a job task analysis or practice analysis and should have been conducted as such. As it stands, this report reads like the opinions of a few faculty who have mostly never worked at a testing organization. I can appreciate the addition of job descriptions to serve as examples, but that really only further highlights the confusion around whether this is an aspirational document of what faculty members think that students should know or something grounded in real-world expectations of entry-level professionals.
Thank you for this comment. Half of our Task Force is either a current standing professional or served for decades as a professional.

I'm struggling with the inclusion of the subdomain on Social, Cultural, Historical, and Political Context. In many instances, psychometricians are not fully able to understand these aspects of the construct. The example on page 9 does little to help as it almost sounds like it's just about reducing construct-irrelevant bias. Perhaps I'm simply misinterpreting this subdomain, but I would never expect this to be a foundational competency because it changes based on the field of application and over time. This is learned by someone once they enter a field.
Thank you for this comment. We clarified Subdomain A based on your and other feedback. Although particular contextual features may change and evolve, we believe that an understanding of the ways in which context affects measurement is foundational.

Comments on NCME Task Force on Foundational Competencies in Educational Measurement Draft Report
Drew GItomer
Rutgers University
October 24, 2022

Thanks so much for giving me the opportunity to review the Draft Report. I think it is excellent and will be a very important and useful contribution. Indeed, we are looking to revitalize our own program, and I just shared the document with colleagues at Rutgers.

As I read through this, I had a few thoughts, many of which I assume the Committee debated. It is a short list for your consideration:

p. 5: Principles
"the design, use, and evaluation of measures of cognitive, affective, and psychological constructs *that individuals*" I was struck that there is no consideration of any type of group/team assessments. Perhaps that was not seen as foundational, but I think the next generation of measurement folks is going to have to engage with this—whether in job performance, gaming/multi-user contexts, etc. Obviously, there has been work in this area, so the question is whether only individual assessment is deemed foundational at this point.
Thank you for this comment. We have broadened this to "individuals and groups."

p. 7: Communication and Collaboration
I thought of several other potential collaborators in addition to the list you provided. First, I

think “other professionals” could include policymakers and practitioners. Measurement people should be able to communicate clearly with these groups. A second group would be to broaden the list of researchers to include policy researchers and other social sciences, including but not limited to policy research, sociology, economics, as well as communication/information sciences.
Thank you for this recommendation. We added practitioners.

pp. 8-9: Subdomain A
The focus in this section is on understanding how contextual factors can influence the design, enactment, interpretation, reporting, and use of assessments. Kudos for including this—all are critically important. What is not in here, though, is an understanding of debates and issues regarding the impact of assessment within society. There is relatively little treatment in the document of consequential aspects of assessment as compared to interpretive aspects. Measurement professionals should understand arguments made concerning perceptions of certain assessment practices as doing harm to particular groups of test-takers, for example.
Thank you for this recommendation. We have attempted to respond to it in Subdomain A.

Final Suggestion
There was one broad competency that I think is addressed only tangentially. This has to do with becoming a critical reader and interrogator of analyses—either one’s own or those of others. I think this is both dispositional and skill based. Measurement people need to examine.
We have acknowledged this in a sixth principle on page 6.

NCEO Contact: Sheryl Lazarus, Director (laza0019@umn.edu)

Foundational Competencies in Educational Measurement: A Presidential Task Force Draft Report

National Center on Educational Outcomes (NCEO) Comments

This report provides an important contribution to the field of educational measurement by clearly laying out NCME’s view of essential knowledge and skills for individuals in the field of educational measurement. NCEO appreciated the clarity of the report. Nevertheless, we identified two areas in which the report could be strengthened.

First, the report should more clearly lay out its intended audience. Much of the report seems to be speaking to higher education professors who are teaching educational measurement courses. Yet, there are several references in the document to individuals who are in the educational measurement field, for example, in test companies or state and district assessment offices. These “measurement professionals” likely comprise the largest number of individuals to whom the competencies apply. Clarifying a primary audience (higher education professors) and secondary audiences (educational measurement practitioners such as school or district professionals) would be helpful if those are indeed the intended audiences and the “primary” and “secondary” separation is what you intend. If a separation into primary and secondary audiences is not

intended, we recommend that the entire document be revisited so that it truly speaks to the entire intended audience.

Thank you. Past-President Derek Briggs has added a foreward that provides additional context about the audience, intention, and charge.

Second, another competency – knowledge of the population to be tested – seems to be missing from Domain 3, either in Subdomain A (Social, Cultural, Historical, and Political Context) or Subdomain B (Validity, Validation, and Fairness). Understanding the population to be tested necessarily affects other aspects of Domain 3, particularly Subdomain C (Instrument Design and Development). We recognize that the report does include a section on understanding culture. This is very important. But, it does not address understanding the test-taking population in ways that would clearly include disability or second language learning. Although the example of a career in which the professional is helping to solve an operational test design issue is about English language proficiency assessments, there is no indication that the professional should have some knowledge or skills in the population to be tested and the content of the assessment. It is difficult to share a view on the difficulty of English proficiency speaking items (as in Example 3 on page 17) if you do not know much about how children learn language or what the speaking items are intended to measure.

Thank you for this. We have emphasized that social structures can include those of language and ability.

We would also like NCME's Presidential Task Force members to reconsider the report's use of the term "ablemess." A common definition of this term is "able in body or mind." Is this the definition the Task Force is using? We see the term used occasionally by other organizations (e.g., Center for Assessment, Boston College), but in those cases, it is not defined. If the term is used in the report, it should be clearly defined. That said, it is not clear that the definition would encompass English learners or many individuals with disabilities (e.g., individuals with learning disabilities). We think it is likely better to simply call out English learners and students with disabilities whenever groups of individuals are listed, rather than assuming that because "ableness" has been mentioned readers will know that English learners and individuals with disabilities are included.

Thank you. We have edited this to "disability" to use a more familiar term.

First, I want to thank Derek Briggs for making this an initiative of his presidency and a big thanks to all of the task members for volunteering what clearly was/is a considerable amount of time, energy, and thoughtfulness to developing this report. The bullet points that follow include my reaction, musings, and thoughts about the document laid out. Before going into these, I also want to commend the idea behind pairing the written report with the webinar(s). Although I was only able to view the first (as the second led by the educators of measurement SIGIMIE hadn't been made available online at the time of writing this), it was enlightening to hear points of tension, agreement, and struggle among and within task force members, and how they relate to the given charge of the committee. The following bullets are not in any specific order, nor are they critiques of the committee who had given more thought and had more discussion surrounding these ideas.

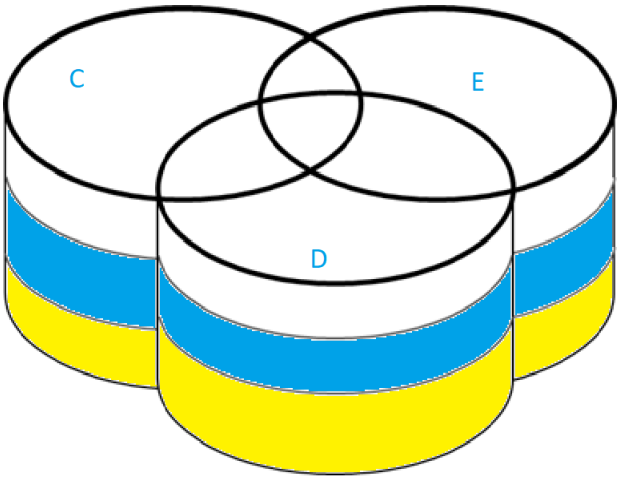
- Since reading/listening I have wrestled with the meaning of 'foundational' in 'foundational competencies'. I am grateful for the guiding principles laid out in the report which have helped frame the background that the task force worked through and situate my own thoughts. The webinar also provided insights to the thought process and a slight peak at the man behind the curtain regarding the work of the committee. When thinking about foundational competencies I have found myself pondering synonyms – fundamentals, building blocks. Although in my view each has slightly different meanings when it comes to foundational competencies, they have helped frame my thinking. I commend the task force diagrams, which have left me thinking about the parallel to quantitative psychology.

  Quantitative psychologists tend to have a degree of expertise in statistics, design, evaluation, and measurement. Most master level students will be exposed to each area and develop knowledge in order to understand, know what questions to ask for in a given application, and, in most scenarios, know what they don't know in the necessary area to learn more if needed. It is then in a PhD program that a quantitative psychologist develops a deep expertise in one area. I see a similar framing in the work being presented in this document. The term foundational tells me that once a student (using for ease, but this should be considered a student of educational measurement, not student in the traditional sense) achieves competency in these areas, they know the questions that should be asked, can do some applications, understand what methods they can and can't do (or how to identify the shortcomings), but also, and I'd argue most importantly, know what they don't know. This is not a negative, as exposure or introduction is essential. In educational measurement, for example, if all the student learned was CTT and never a mention of IRT, then they are not competent to know that there is another approach 'out there'. However, if they develop familiarity to CTT but only exposure to IRT, then at some level this foundation sets up a student to engage in future learning.

  As noted in the webinar and throughout the text, the concepts of validity, validation, and fairness are critical to educational measurement. In fact these, in combination with context, are what sets educational measurement apart from other fields (e.g., data science). Knowledge of these components are the building blocks, so that without competency, one can argue there is no way to have competency in C: instrument design and development, psychometric modeling, and precision and generalization. Furthermore, without context including the social, cultural, historical, and political backgrounds, an educational measurement student cannot truly understand validity, validation, and fairness. How can one understand bias, be critical of bias, or broaden fairness beyond bias, without understanding the social, cultural, historical and political nature of the measure/construct/etc?

  It is painful for me to suggest adding three dimensions to a graphic, but alas, here it is. To help me frame these ideas, I show an alternative version to the primary visual:

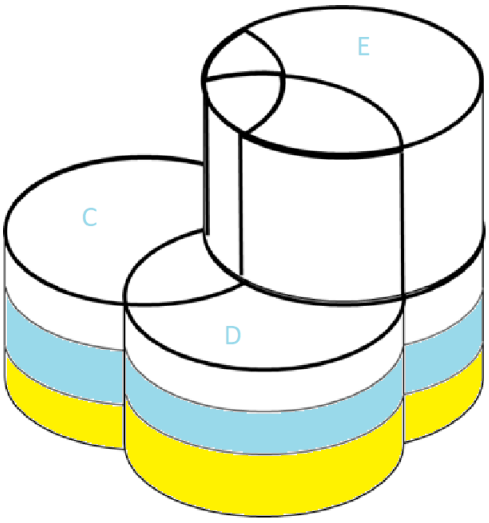
Yellow: Context: Social, Cultural, Historical, and Political
Turquoise: Validity, validation, and fairness



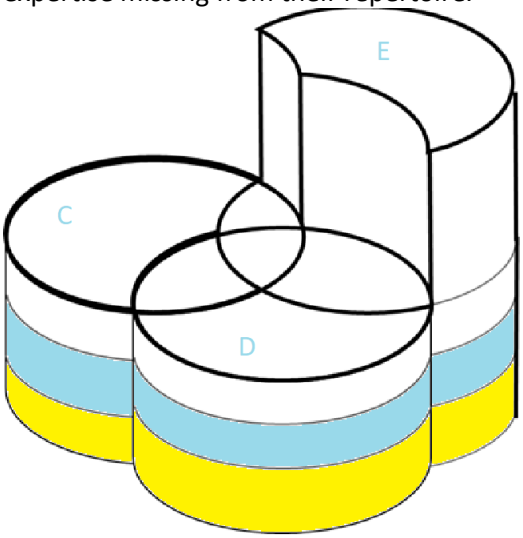
You will likely first notice the 3-d nature of the graph. This is to match the term foundation – or building blocks that are fundamental or are foundationally essential. Without building this 'amount' or 'foundation' of knowledge, a student of educational measurement does not have ground to stand on. You may also immediately notice the two colors. Note that C, D, and E match those on the original graphic, representing instrument design and development, psychometric modeling, and precision and generalization. The two colors, yellow and turquoise, represent context and validity/validation/fairness, respectively. To have the foundation in any of the three areas, there must be a foundational layer of context followed by validity/validation/fairness. These are critical to C, D, and E and without either, each section would 'collapse'. By collapse, I mean the knowledge base doesn't have a leg to stand on and would crumble when challenged or critiqued[1] (mixing metaphors, I know).
Above the two colors, there is white space representing the additional foundational components in each of the three areas. These represent the knowledge above and beyond the context and validity, validation, and fairness that result in competency in each area. For example, learning the statistical modeling principles of item response theory 'out of context'. While some may argue knowing the mathematics behind a 2 -pl is suitable to knowing IRT, without contextualizing nor understanding how it can be used/applied in context results in it being a technical/statistical competency, outside of the educational measurement competency circle [perhaps in the interaction piece on the left figure of figure 1 -interaction between 3) educational measurement competency and 2) statistical competency].

Also, using this frame can also lead to extensions, where someone can then go on to develop an expertise in one of these areas (e.g., getting a PhD specializing in an area; advancing learning to develop expertise in equating). So, for example, an educational measurement professional's diagram may look like the following where they have expertise in psychometrics/psychometric modeling:

[1] Educational measurement professionals are constantly challenging and being challenged (e.g., public on testing; administrators; etc.), but with a 'validity argument' in the 'context', we can overcome the challenge.



However, educational measurement professionals must still build on their understanding of instrument design and development as well as precision and generalization or else, their large E circle would look like the following with a significant piece of their psychometric modeling expertise missing from their repertoire.



The Task Force appreciates this elegant extension of its framework and welcomes further elaboration of the framework and its implications in possible companion contributions as our engagement with membership continues.

- As spoken about during the webinar and noted here, these are foundational to develop expertise for people who want to continue to learn more and are motivated to being educational measurement professionals. Thus, to me an essential foundational competency that appears missing is being able to identify, seek out, and learn from professional development resources. This should include knowing where and how to find new, novel, or deeper information. In a course (Section 3), this may be to read articles from ITEMS; EM:IP) and not just

reading a single text. I do think it is a competency to be able to have information literacy skills within the educational measurement domain.
Thank you for this suggestion. We have added this to our list of principles as a dispositional competency that we think is important but distinct.

- Related, in the fifth guiding principle, "Educational measurement competencies are those that support the design, use, and evaluation of measures of cognitive, affective, and psychological constructs". Should this include knowing how and where to find developed measures, how to evaluate the quality for an intended use/interpretation, and what questions to ask to understand the applicability of the measure? Although that may be embedded in the validity competency, this seems foundational for anyone being an educational measurement professional.
  Thank you for this suggestion. We have added this to our list of principles as a dispositional competency that we think is important but distinct.

- This principle also blurs the line between different types of measurement. There are times throughout this report where there are instances of the phrases "educational measurement' and "measurement". Are these being used interchangeably? Is the term 'educational' too limiting here for real-life application in graduate programs? Given that the case is made the educational measurement professionals do not live in a vacuum and in order to be true educational measurement scientist they will have interdisciplinary interactions, if someone were to achieve these competencies, are they set up for success in either psychological or educational measurement? Given that some journals and resources (e.g., *Educational and Psychological Measurement*) blur the line, is this distinction essential? Perhaps as educational measurement expands to include more affective and psychological measures, educational measurement competencies must include psychological measurement competency. Understanding learning theory as well as cognitive psychology (or having a foundation to each) are now likely critical to making a response process argument for validation – should educational measurement foundations be the same as psychological measurement foundational competency and then we can just safely use the term measurement?

  From a practical standpoint is it too limiting for a graduate program that is situated in a psychology department, but may yield educational measurement professionals? This is such a great document that I wouldn't want it to be disregarded if it is too limited in scope. This is just a musing.
  Thank you for this observation. Our charge was to engage an NCME audience that is inclined toward education, so we oriented our framework accordingly. Subsequent engagement by NCME members may be able to broaden this audience.

- I am incredibly grateful that collaboration and communication skills are not only in this but are given substantial description.
- Although this is a minor topic, and I believe one of the activities in section 3 gets to this point, we use terms that are often mistakenly treated synonymously by students: assessment, testing, measurement, etc. Do you see one of the foundational competencies as understanding the difference here?

- I have limited comments toward section 3 as I primarily see that section as examples for courses/sequencing. Given that these introductory type courses are rarely (if ever) only offered for students who are exploring educational measurement as a profession, it is difficult to provide insight due to the many situational factors. For example, our introduction to measurement theory course is a required course for students in a school psychology program. Students in this program must meet requirements for boards and therefore course content is not 100% for educational/psychological measurement students. I do believe that the framework emphasizing context/validity theory as critical components is helpful. I do want to emphasize that I find the information in section 3 incredibly value. As noted, collaboration and communication are essential for educational measurement professionals, this is a prime example of how educational measurement educators can and should communicate to improve curriculums. Sharing syllabi, curricular activities and having intimate discussions about courses can improve the graduate school experience across the board. So, thank you! As an aside, it was with this motivation that I started the initiative in the Educators of Measurement SIGIMIE to create a syllabus repository – but unfortunately, it never took off although it is not too late and these kinds of discussions (as noted here linking it to the foundational competencies) can be very valuable for educators!
  Thank you, and the Task Force is also grateful for your contributions to the Educators of Measurement Special Interest Group.

- Brian C Leventhal, PhD
- James Madison University
- Director, Assessment and Measurement PhD Program
- Assistant Professor, Department of Graduate Psychology
- 298 Port Republic Road, MSC 6806
- Harrisonburg, VA 22807
- 540-568-5004 (Office)
-

Hi Derek,

Hope you're well! First off, a huge thanks to you, Andrew, and the team for putting together this document. I think it's an important contribution to the field, and I will certainly use it to guide my teaching. In case helpful (and in the spirit of constructive feedback/improving on the margins), I have a few thoughts on the document in its current form. Please do let me know if I can be helpful or clarify anything as you move forward.

Broad Comments

- Perhaps my biggest comment is that the current draft makes it seem like educational assessment is synonymous with/limited to educational testing. While that's the bulk of the field, given the growing emphasis on SEL-related survey outcomes and the proliferation of teacher observational protocols (among other instrument types), I think the framing could be a touch broader/more inclusive. As examples of the testing focus, the draft often mentions things like "score reporting". Another concrete example: "relevant social, historical, and political factors in common *testing applications including accountability testing, admissions testing, certification exams, or classroom assessment.*" As another, the ***description*** subsection of the validity section, the first sentence mentions only test scores.

  Thank you for this observation. We consider SEL-related survey scores to be test scores. This prompted us to add a definition of assessment and testing to our glossary.

- A more minor point related to psychometric modeling: I think it might be worth calling out that IRT, SEM, G-theory, etc., are all just different flavors of latent variable models. In my experience, providing that connective tissue, and understanding the similarities/differences among these types of models, can be invaluable.

  Thank you. We have added this idea to Subdomain E.

- Finally, at some point in the document, it seems worth pointing out that many of these technical skills and competencies may look different in important ways when aggregate inferences are desired compared to individual scores. For example, building latent regressions into an IRT model and using EAP scoring may be the appropriate course of action for aggregate inferences, but would not be justifiable for individual uses. In addition to pointing to NAEP and PISA, you could also call out growing work on the importance of these scoring decisions to aggregate inferences related to understanding growth/treatment effects (happy to provide some of my favorite references if useful).

  Thank you for this important observation. Although plausible values methodology is beyond what we consider a foundational competency, we have added acknowledgments about aggregate-level inferences in the precision and generalization section.

Specific Comments

- The logic of Figure 1 and why A sits above, B below is still a bit murky—could be spelled out more

  Thank you for this observation. We have expanded the Domain 3 description to spell this out.

- In general stats, it seems there's lots of emphasis on software, less on the nuts and bolts of probability

  Thank you for this observation. We added a nod to probability.

- I think it's worth pointing out in the validity section that part of this competency is knowing that validity is not a property or trait of a measure; it is specific to the intended use and group being measured. The language used is absolutely in line with this sentiment, I just think it could be more explicit.

  Thank you for this observation. We have tried to improve this.

- In the precision section, possibly also call out that reliability changes for particular contexts and issues, such as when examining change over time, and that precision can be person-specific/dependent on where the individual is on the scale

  Thank you for this. We have attempted to address this in Subdomain D.

- In subdomain E, also mentioning the importance of measurement model misspecification, and the ways it can introduce bias not only into scores, but into downstream estimands of interest (e.g., treatment effect estimates)

  Thank you for this observation. We have made minor edits to Subdomain E.

- Also in E, understanding the assumptions implicit in sum scoring and when sum scores are/are not justifiable might be worth a call out?

  This is rather specific, and we hope it is captured in existing sections at a higher level of generality. We look forward to your ongoing research that continues to deepen appreciation of this issue.

Thanks so much again!

Jim

James Soland
Assistant Professor
Research, Statistics, & Evaluation
School of Education & Human Dev.
University of Virginia
e-mail: jgs8e@virginia.edu

Dear Task Force,

Thank you for the thoughtful effort you have placed into this process of developing foundational competencies in educational measurement. I have long believed this to be necessary, and that the field will greatly benefit from having such a [routinely updated] document.

I will focus my comments in response to the framework as all other components of the document are conditioned upon that framework. My first comment refers to Figure 1 and the decision to make Subdomain A (Context: Social, Cultural, Historical, and Political) a subcomponent of educational measurement competencies. I want to thank the task force for naming the importance of this subdomain explicitly. However, I would argue that this subdomain should, in fact, be the "it" that surrounds all three domains and their subdomains. To be sure the ways in which the sociopolitical, cultural, and historical context play out in the ways in which we communicate/collaborate and in the field of statistics are profound. For example, with respect to communication and collaboration, the document reads "Collaboration skills also include the ability to understand a variety of perspectives, manage priorities of all participants, and meet expectations as a member of a team" (p.7). Expressions like "understand a variety of perspectives imply that every perspective is one worth considering, which is not - in fact - the case if that perspective is rooted in white supremacist logics or if the perspective fails to even acknowledge the impact of the conversation/decision on minoritize groups. A stronger - more justice-oriented- stance would make a note of the importance of students/learners being supported in their development of cultural competence and/or humility (which looks like subdomain A).

Similarly with Domain 2 (Technical, Statistical, Computational Competencies), I want to call out the importance of students of statistics understanding the history of statistics and the ways in which that historical context cannot - or at least should not- be ignored (Zuberi, 2003). Add to this in the Justification I would like to see the authors call out the limitations of most statistical methods we employ in that they focus on the center and ignore those on the margins.

This is an important and astute observation, and certainly the Task Force recognizes that context is an overarching frame that suffuses all of these domains. We have added text to acknowledge this in our description of Subdomain A. We explain there that we locate Context over Educational Measurement competencies to emphasize the responsibility of learners and professionals to advance these competencies within this field as well as beyond it.

Moving on two the examples used to illustrate Subdomain A…The document reads "This includes understanding the importance of designing items and scoring procedures that adhere to bias, sensitivity, and accessibility guidelines to maximize test-taker engagement and minimize potential construct-irrelevant bias" (p. 9). I argue that we (as a field) already do this and not much has come from it in the last 20 years (other than the removal of items about skis and European vacations). Indeed, this document brings up bias and sensitivity

guidelines more than once - each time assuming a level of quality that I argue is not deserved. Instead, we must do more than teach our students to avoid bias. We must teach them to seek justice in the review process. Moreover, the document reads "They can understand the importance of political context by, for example, anticipating how test-based accountability policies for teachers and schools may inflate aggregate trends over time" (p. 9) - And I would add to this that students need to understand how results can further support negative stereotypical narratives of interiority about minoritized students. And that students should engage in pedagogical conversations about ways to present data results that take into the current context and the negative consequences of their reporting decisions (e.g., how they report the data)

With respect to Subdomain B (Validity, Validation, and Fairness), I want to call a particularly problematic example used ("why the correlation between test scores and college grades is important evidence to support the use of an admissions test…"). If, in fact, both of these measures privilege whiteness, then the appropriateness of that interpretation is called into question. I am not suggesting that this example simply be removed. I am suggesting that this subdomain also take into consideration what you laid out in Subdomain A - in that context matters. Instead, students should be encouraged to interrogate the measures used when establishing the *relation to external variables* validity argument to ensure assessments are not simply further perpetuating existing systems of repression (i.e., correlating measures of white supremacy).

This is an important observation, and we adjusted the text to make it clear that our goal is not just to seek evidence but understand whether, when, and why this evidence is supportive of score use in context.

Did anyone notice that despite the inclusion of Subdomain 3A in the framework, the ways in which it would show up in the four examples of career scenarios is not at all included? Yet, 3B, 3C, 3D, and 3E are called out at least once in the examples. This should be addressed; and I would argue it should be addressed in each of the four examples.

Thank you. We have tried to improve this in the more specific examples 3 and 4, and we believe there are many other improvements and additions to contribute.

Best,
Jennifer Randall

Hello -

I wanted to share a comment on the Foundational Competencies in Educational Measurement report. First, I wanted to thank the authors for their excellent work drafting this report. One suggestion that I have is to add competency in open science to the list of foundational competencies. I think that transparency is an important aspect of the educational measurement work and is valued more and more in the modern world. Hence, knowledge and

skills in open science (open data, open code, preregistration, etc.) will be necessary for students to develop.
Thank you for this comment. We have edited Domain 2 to acknowledge this.

Thank you,
Daria

Daria Gerasimova, Ph.D.
Assistant Research Professor
Data Science Team
Kansas University Center on Developmental Disabilities (KUCDD)
University of Kansas

Comments from ACS on NCME Draft Educational Measurement Competency Document

The team at ACS Ventures collaborated to create a collective set of feedback on the NCME Draft Educational Measurement Competency Documents. The comments below represent our collective perspectives.

1. Overall, we recommend beginning the document by addressing the following questions:
    a. What is the purpose of the document?
    b. Who is the intended audience?
    c. What is the intended use(s) of this information?
    d. What is the gap being addressed with this information?
    Thank you for this suggestion. We have added a foreward to the document from Past-President Derek Briggs to provide additional context about the context, audience, and charge.
2. We respect the expertise of the Task Force members but do not feel this group was representative of the range of stakeholders in the educational measurement community.
    We respect this and hope to engage stakeholders who you feel were not represented.
3. An analysis of competencies or practice should follow a systematic process that is broadly inclusive of the profession (see *Standards for Educational and Psychological Testing*). This would involve a survey of stakeholders to get representative information from membership at a minimum.
    This is a wonderful idea for testing and extending this report and similar efforts in the future.
4. The document purports to identify foundational competencies, but it is difficult to determine what those competencies are. There appear to be descriptions of the competencies, but no explicit statements of expected competence. We anticipated seeing something like statements identifying what knowledge or skills are involved and then what individuals are expected to then do with those knowledge and skills (e.g., tasks).
    This is a level of specificity that future efforts could aim to achieve.
5. We do not think Figure 1 clearly conveys the relationship among the domains and subdomains. Additionally, it does not appear to represent the educational measurement field based on the emphasis (or lack thereof) of the included concepts.

6. With respect to the organization of the document, consider leading with what is unique to the field and then follow that with the generalizability of things like communication, technical etc.
7. Testing and measurement have a long history. Much of this history is relevant to current practice and the historical context provided here seems limited and, to some extent, dishonoring of its contributions to equity. Some of the context also appears to reflect current social contexts that have yet to be established as historically impactful.
   Thank you for your observation. We hope to engage you further to understand what you feel is missing and how we could improve our framework.
8. We aren't aware of many K-12 practitioners who engage in job analysis on a regular basis. This appears to be out of place but rather should be in the licensure and certification list.
   Thank you for this observation. We agree and have shifted job analysis to Example 3.
9. In practice, many educational measurement professionals are leaders within their organization or considered to be thought leaders for the programs when serving as an external expert. However, we do not feel this perspective is represented in the current framework but rather there is more emphasis on data collection, modeling, and analysis. There are many programs where one psychometrician is responsible for all activities rather than just serving in an analyst role.
   This is a great point, and we have added this to Section 2.
10. There should be more emphasis on the fundamental aspects of program design, test design, and test development phases. For example, where does the data come from and what does it mean before starting to apply the range of modeling and statistical methods to the data?
    Thank you for this observation. We hope that our revisions to Subdomains A, B, and C capture the importance of data context, purpose, construct definition, and instrumentation.
11. The section on precision and generalizability should be expanded to include decision consistency, human judgment, and computer-based automated (algorithmic, AI, ML) methods.
    We have emphasized rater error in Subdomain D. We also now elevate computational methods into our subdomain related to instrumentation.
12. Subdomains C, D, and E can stand alone as distinct, additional domains.
    We acknowledge that this is possible, but we focus on their development and applications for educational measurement.
13. Measurement experts are often asked to interpret data for a variety of audiences. More emphasis on this competency is needed.
    Thank you for this comment. We have tried to emphasize this further in Domain 1.

Thank you for this opportunity to review the draft document and provide feedback.
Andrew Wiley, Chad Buckendahl, Susan Davis-Becker, Deborah Schnipke, Russell Keglovits, and Kelley Wheeler

---

Thank you for the opportunity to review the draft report on the Foundational Competencies in Educational Measurement. It is an excellent document and users should hind it helpful. I have provided suggestions below.

Page 5
The definition for fairness should be broader, including not only score use but test use. Tests used as levers for educational change should be evaluated for their use and resulting consequences, both intended

and unintended. Fairness issues arise when there are negative consequences for subgroups of students (such as narrowing of curriculum to focus on lower-level skills).
We have expanded our discussion of these issues in Subdomains A and B.

A definition for precision, generalization, and validity should be provided.
We have attempted to clarify what we mean by these terms in Subdomains B and D respectively.

Page 6
In figure 1, The inclusion of Subdomains A and B is much needed and should be made more prominent.

Page 8 – Domain 3 paragraph
The last sentence should be changed to " validity, reliability, and fairness of score for their intended purposes and their uses, validity and fairness should also be evaluated in terms of enacted uses and consequences of use". The document should not portray the narrow view of validity and fairness as being relevant only for test scores and not their uses. This was a mistake of the *Standards* that should not be perpetuated.
We have added "enacted purposes."

Last sentence in first paragraph of Description
In addition, "to improve the likelihood of valid interpretations and intended effects', it should also include "and to minimize the likelihood of unintended, negative effects".
Thank you for this suggestion. We have incorporated it.

Social Context
I am not sure what is meant by ableness.
We have edited this to "disability."

Page 9 Subdomain A: Justification
The way in which constructs are conceptualized and defined is influenced by the social, cultural, and historical influences of those involved in defining them. This notion should be included in the document.
We have attempted to address this in Subdomain A.

Subdomain A: Example
Bias should be defined. And the guidelines should refer to fairness as well as to bias, sensitivity and accessibility.

Subdomain B – Description, Justification, and Examples
I am glad that the reference to interpretations and uses of test score is here, but it should also include enacted uses and intended and unintended consequences. Validity should also refer to test use, not only test score use, such as when tests used as educational change agents.
Thank you. We have added tests as well as test scores.

Page 10 - Subdomain C- Description
It should also include expertise in designing" task and scoring models/templates" that can generate items that have some common features, but also unique features.

Subdomain C- Example
The learner should also work with experts from different social and cultural groups when defining the construct.
We have added the importance of including those with relevant experience and expertise.

Page 10 - Subdomain D
Description – In addition to generalizing to items, raters, and occasions, which are commonly used, include contexts and groups.
We now mention aggregate scores.

Page 13 – 1st paragraph
When discussing the effects of advances, I would also include the need to learn new models and methods.

2nd paragraph- For communication competencies, include communicating with different stakeholders, such as state agencies, public, reporters, etc.

Page 14- Example 1- Domain 2
Don't psychometricians also perform statistical analysis and develop statistical programs?

Page 16 – Example 4 -Domain 1
This domain should include the ability to independently and/or collaboratively write grants to support faculty research and students.

Page 16 Example 1
I would also include internship programs and collaboration with other researchers.
For faculty, I would also include fund students as graduate student researchers.

Page 17 – Example 2
Many of the items are important for all faculty in research universities, such as mentoring students through dissertation research, pursuing external funding, reviewing for journals. It would be inaccurate to send a message that these activities are only for senior faculty.

Page 18 – Example 4
I would include the need to ensure that the sample the engine was trained on was representative of the population.

Page 19 –

In all discussions regarding statistical and psychometric competencies, it is important to stress the need that they should focus on both theory and application.

Domain 2
I would include it may require courses in statistical theory not just applied statistics.

Domain 3
Investigating fairness should begin with defining the construct and end in methods for evaluating the consequences of testing in addition to equating and setting performance standards.

Page 20 Subdomain C
I would indicate that some training in cognitive science (how students learn) would benefit instrument design.

Page 21 Course 1
Why limit to a 'consensus' definition of validity and fairness in the Standards? Other views should be presented. OK, I see it in week 3, but why wait so long. Validity, Validation and Fairness should be in the beginning since it is pervasive in all of measurement and assessment.

There seems to be quite a bit in this first class, and I am concerned that the depth of any one topic will be sacrificed for breadth. I would indicate that additional courses are needed in areas such as Design of Achievement Measures, Survey Design, Fairness in Testing, Sampling, etc.

For course 2, I would also suggest that there could be separate classes on all of the topics. Again, breadth is emphasized at the expense of depth.

Some programs have a cognate of 6-9 credits where students can get more training such as in cognitive science, statistical theory, data mining….

Thank you for this collection of feedback. We have attempted to respond to the collection in our revision.

Great set of curricular activities

Suzanne Lane
Professor Emeritus
Research Methodology
University of Pittsburgh

Castellano, Katherine KEcastellano@ets.org

- P. 14: There are two "Example 1"s
- Pp. 14-16: Only faculty members are given the duties of "developing new measurement models that advance scholarship and practice" and "demonstrating new applications of existing measurement models for problems of practice", but these are tasks done by psychometricians and researchers at testing companies as well.
  Thank you for this observation. We agree, and we have added these to duties to the list. We have also emphasized that these descriptions are illustrative of common duties and not exhaustive.
- "Machine learning" is only mentioned once in the document and "computational psychometrics" is never mentioned, but the field is moving in this direction and as Alina mentioned in her talk at the 2022 MARC Conference, education/ed assessment is already behind in this phase of the industrial revolution. There is a brief example of such a task on p. 7 and taking classes in AI, ML is mentioned on p. 19, but more attention to this skillset should likely be included in the report.
  Thank you. We agree and have emphasized that statistical, technical, and computational competencies are foundational for these methods.
- Last year I was part of a hiring committee for a research scientist, and it was very difficult to find someone with the skills we needed – someone with both strong training in foundational psychometrics and thought deeply on issues like reliability AND was proficient in AI/ML models and able to develop ways of applying them to new psychometric issues. We didn't quite find this unicorn, but I'm guessing this job description will be more the norm than the exception as we move into more computationally intensive modeling in psychometrics so training new psychometricians/measurement specialists with these skills is important.

To: NCME Task Force on Foundational Competencies in Educational

Measurement From: Stephen Sireci

Re: Draft Report
Date: November 25, 2022

I am writing to provide some comments on your excellent draft report. I realize these comments are coming in almost one month past the deadline, but I am sending them in case there is still time to consider them.

1) *My first comment is, Thank you for the excellent draft report!* The amount of work accomplished by the Task Force is truly amazing. When looking at the members of the Task Force, that is not surprising. What a great group. We are all indebted to you. This is an important endeavor for NCME and you have already provided great material.
Thank you. We enjoyed the effort and look forward to future engagement.

2) *I think the timeline for comments should be extended.* The draft report was released on September 16, with a deadline for comments by October 31. On one hand, 6 weeks seems like a long time for a comment period, and I know the Task Force wants to release a final report ASAP. On the other hand, September and October is an insane time of the year for university professors and graduate students—perhaps the two most important stakeholder groups for this report. It is only because I was able to fill my family with tryptophan and lots of wine that I had a little time to quickly review the report. As soon as I started to read it, I realized I should have made this a UMass Center for Educational Assessment response involving all faculty and students, instead of a Steve Sireci response. For that reason, I am copying my colleagues and will discuss with them whether any of them have also submitted comments. However, if you could extend the comment period until January 31, that would give faculty and students the January break to really absorb the proposals and provide feedback. If that extension will not work, another option would be to consider these "version 1" for a year, with a timeline to revise after one year of public comment.
Yes, as the report emphasizes, we expect to continue to engage NCME membership and look forward to future efforts to come to broader consensus on foundational competencies.

3) I thought *the framework is brilliant*. The foundations are comprehensive and their overlap is well displayed.

4) In “Subdomain B: Validity, Validation, and Fairness” (p. 9), I think *thorough familiarity with the AERA/APA/NCME Standards should be explicitly stated* in the description, justification, and examples.

We appreciate this point. The Standards are essential, but not every standard is what we might consider “foundational.” The Standards are also 8 years old and about to enter a period of revision. It may be useful to consider these two efforts as in conversation with each other rather than having one simply point to the other.

5) pp. 13-16: I think the reference to Figures 1-5 on page 13 must refer to the examples labeled 1-4 (there are two example 1s), but I am sure you have already caught those typos. More important for this section is **to *add a description of where the information came from*. Were practice (job) analyses d**one to come up with these lists, or was it something the Task Force came up with on its own? If the former, how the data were collected should be described. If the latter, that should be added so its subjective nature will be more clear.
This is a good point, and we can stress that these are reflect the experiences and perspectives of the Task Force rather than a more systematic analysis.

6) *The section from page 19-22 seems to be the one section that needs revision*. For example, on page 19 it reads, “Traditional course sequences in educational measurement often begin with a treatment of validity and defer fairness and methods for detecting differential item or test functioning until later in curricular sequences.” I do not think that is a good description of “traditional course sequences,” if such a sequence exists. For example, we run our courses on a two-year sequence and so it is quite likely a student will learn about DIF the first semester. Validity is covered in several courses, and we also have a full-semester course on validity. Item response theory (IRT) is an entire one-semester required course, and we offer a second IRT course as an elective.

Similarly, I thought the section on pages 20-22 “What are examples of course topic sequences in a first-year educational measurement sequence?”, was unrealistic. It seems to assume a university teaches measurement in two courses, rather than a comprehensive series of courses over a two-year sequence. I realize the idea is to propose “one possible design for a two- semester sequence in educational measurement” (p. 19), but two courses are not two semesters and what is presented makes it sound like you can teach psychometric foundations in two classes. Some may agree, but I do not. It probably depends on what kind of program you are referring to. My bet is the *Institute for Credentialing Excellence* is certifying assessment expertise with far fewer competencies, but for a masters or doctoral program it would be hard to be comfortable with two courses in measurement. Moreover, many of the topics listed in one week of the course outline is a full semester course at UMass.

In revising this section, it may be helpful to avoid prescribing how to offer the competencies, but rather stress they should be covered over a two or three-year course sequence. If you want to "illustrate one or more curricular models for a graduate program in educational measurement" (p. 4), I suggest you do that. *Provide the course sequences from a few universities that are doing a good job graduating relatively large numbers of competent measurement specialists* (e.g., Iowa, UMass, JMU). Both masters and doctoral programs could be illustrated and connected with NCME's list of programs (https://higherlogicdownload.s3.amazonaws.com/NCME/4b7590fc-3903-444d-b89d-c45b7fa3da3f/UploadedImages/Documents/Educational_Measurement_Program_Descriptions____2020-06-24_.pdf).

The Task Force may also find it helpful to review the International Test Commission's "Learning Center," which has 7 modules on testing fundamentals: https://www.intestcom.org/page/81.

These are helpful references, and perhaps a supplementary document could provide pointers to these programs.

7) There are *3 definitions I think can be improved*: *Statistics*, *psychometrics*, and *education*. Table 1 presents the definitions from the Task Force report and some alternatives. I am not suggesting the alternatives be directly substituted, rather, I think they illustrate the current definitions in the Task Force report are a bit too limited and can use revision.

I hope these comments indicate my strong support and appreciation of the work of the Task Force, and also lead to improvements in the revision. Thank you for your dedication to NCME and the field of educational measurement.

Sincerely,

Stephen G.
Sireci, Ph.D.
Distinguished
Professor
Director, Center for Educational Assessment

Cc: Lisa Keller, Javier Suárez, Scott Monroe, Craig Wells, April Zenisky

Term	Task force Definition	Suggested Revisions
Statistics	the science of describing and modeling physical and social phenomena using data to improve prediction and understanding	“the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements” (dictionary.com)
Psychometrics	a field of study in psychology and education characterized by statistical modeling of latent variables motivated by psychological theory	“the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality traits, skills, and knowledge" (APA, AERA, & NCME, 1985; p. 93).
Education	a process or system for improving human competencies through learning	“a purposeful activity directed at achieving certain aims, such as transmitting knowledge or fostering skills and character traits” (Wikipedia)

Thank you for these alternative definitions. The Task Force took these definitions into account in our revision.