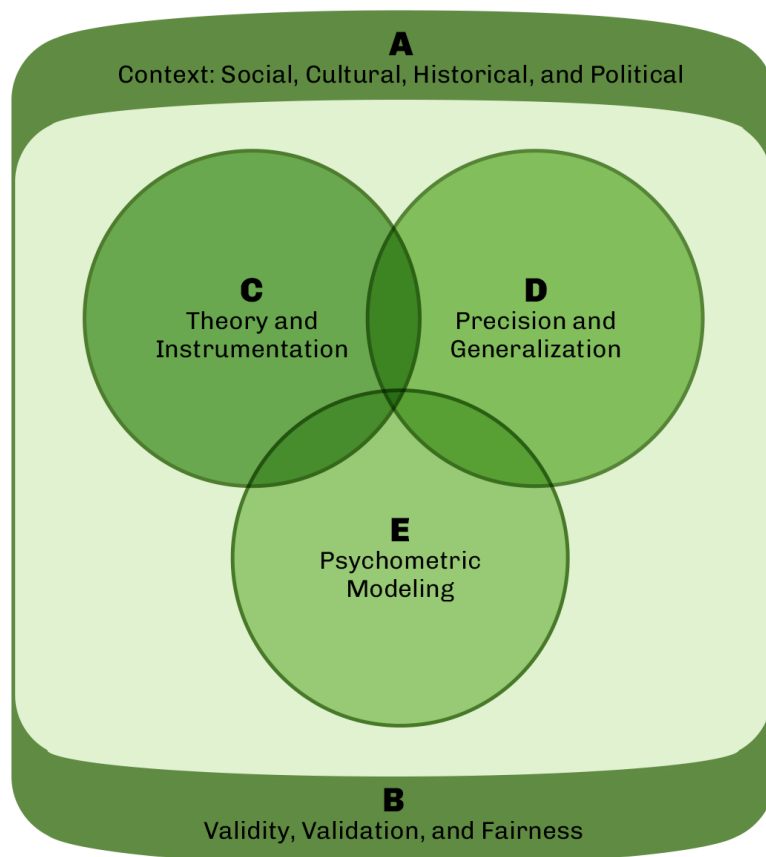


## FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT



### A Presidential Task Force Report of the National Council on Measurement in Education

March 2023

<b>Foreword by NCME Past-President Derek Briggs</b>	<b>4</b>
<b>NCME Task Force on Foundational Competencies in Educational Measurement</b>	<b>6</b>
<b>Process</b>	<b>6</b>
<b>Guiding Principles: What does “foundational competencies in educational measurement” mean?</b>	<b>6</b>
<b>TASK FORCE PROVISIONAL DEFINITIONS OF TERMS</b>	<b>7</b>
<b>PART 1: A FRAMEWORK FOR FOUNDATIONAL COMPETENCIES</b>	<b>9</b>
<b>Domain 1: Communication &amp; Collaboration</b>	<b>10</b>
<b>Domain 2: Technical, Statistical, &amp; Computational Competencies</b>	<b>10</b>
<b>Domain 3: Educational Measurement Competencies</b>	<b>11</b>
Subdomain A: Social, Cultural, Historical, Political, and Learning Context	11
Subdomain B: Validity, Validation, and Fairness	12
Subdomain C: Theory and Instrumentation	13
Subdomain D: Precision and Generalization	14
Subdomain E: Psychometric Modeling	14
<b>PART 2: FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT CAREERS</b>	<b>16</b>
<b>What defines a career in educational measurement?</b>	<b>16</b>
<b>How do foundational competencies manifest in educational measurement careers?</b>	<b>16</b>
<b>How do competencies continue to develop throughout educational measurement careers?</b>	<b>17</b>
<b>How can different educational measurement careers require foundational competencies?</b>	<b>17</b>
Example 1: Illustrative Duties of Psychometricians in K-12 Assessment Companies	18
Example 2: Illustrative Duties of Psychometricians/ Researchers in K-12 Educational Technology Companies	18
Example 3: Illustrative Duties of Psychometricians in Licensure and Certification Organizations	19
Example 4: Illustrative Duties of Research Associates in Educational Research and Consulting Companies	19
Example 5: Illustrative Duties of University Faculty in Educational Measurement	20
<b>How can specific educational measurement career scenarios require foundational competencies?</b>	<b>20</b>
Example 1: Mentoring Graduate Students to Become Educational Measurement Professionals	21
Example 2: Senior Faculty Contributions to Measurement	21
Example 3: Helping to Solve an Operational Test Design Issue	22
Example 4: Helping to Solve an Operational Item Scoring Issue	22
<b>PART 3: FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT COURSES, PROGRAMS, AND ACTIVITIES</b>	<b>23</b>

<b>Where in a curriculum can programs develop students' foundational competencies?</b>	<b>23</b>
<b>What are examples of sequences of course topics in first-year educational measurement courses?</b>	<b>24</b>
Course 1: An Introduction to Educational Measurement or Measurement in Survey Research	25
Course 2: Item Response Theory and its Applications	26
<b>How can curricular activities develop foundational competencies?</b>	<b>27</b>
Curricular Activity 1: The Meaning of Measurement	27
Curricular Activity 2: Sampling Foundations of Classical Test Theory	27
Curricular Activity 3: Contrasting Classical Test Theory and Item Response Theory	28
Curricular Activity 4: Theory and Instrumentation	29
Curricular Activity 5: Psychometric Properties of Gain Scores	30
<b>AFTERWORD: A CONCLUDING REFLECTION FROM TASK FORCE CHAIR ANDREW HO</b>	<b>31</b>

## **Foreword by NCME Past-President Derek Briggs**

As I was beginning my term as president of the National Council on Measurement in Education in the spring of 2021, it struck me that there has never been a better time to learn about the methods and practices of educational measurement. There are so many things worth measuring, so many new ways to think about the instrumentation of measurement, and so much data that in many instances might just be a few clicks of a button away. Indeed, as I write this, our field is grappling with national and international debates about high-stakes testing, artificial intelligence (e.g., chatGPT), and whether, how, and what it means to measure new constructs that are more purposefully situated in cultural context.

But these new opportunities require some caution. Many people who self-identify as psychometricians or measurement specialists tend to arrive in the field through idiosyncratic channels. Few people have a disciplinary background in psychology; some may know very little about education. And many of the kinds of people that might have previously self-identified as psychometricians or statisticians may now see themselves as “data scientists.” Perhaps most importantly, we are at a time when the context of educational measurement is undergoing renewed scrutiny. So, it stands to reason that we should be asking ourselves some important questions:

What are the foundational competencies that we would expect any newly arriving member to the educational measurement profession to eventually know and be able to do? How can we build consensus around these foundational competencies and help to foster them among students and new entrants to the field?

It was with these questions in mind that I began to imagine the possibility of producing an NCME consensus document that would address these questions. It could be as broad as a framework or as detailed as a curriculum. Such a document could help to improve the visibility and standing of the field. It could attract talented and committed undergraduate students to important work. It could cohere and improve instruction among and within graduate programs. It could improve the skills and readiness of incoming professionals to measurement organizations. And it could even serve as the basis for a license and certification program in educational measurement.

I arrived at the following three-part charge for a hypothetical task force that could get this ball rolling:

1. To develop a set of foundational competencies for the field of educational measurement.
2. To illustrate one or more curricular models for a graduate program in educational measurement.
3. To engage NCME membership and the field with Task Force findings through conference presentations and published journal articles.

I set about identifying a small task force that would help me to accomplish this charge. I asked for and received nominations from NCME’s Board of Directors and from the co-chairs of NCME’s Educators of Measurement (EoM) Special Interest Group in Measurement in Education (SIGIMIE). From these nominations, I identified 11 individuals who were diverse in terms of their role in the field of educational measurement, their areas of expertise, their years of experience, and their gender and racial/ethnic identity. I asked Andrew Ho if he would be willing to serve as the task force chair, and he agreed.

Indeed, I was heartened that almost everyone I approached to participate in this project agreed to do so with considerable enthusiasm. The full membership of the task force, in alphabetical order by last name, follows:

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*

It was a true pleasure to work with this group to produce this document. What I think you will find is that it not only meets my motivating charge by delineating a set of foundational competencies and illustrating how these competencies can be realized in graduate training, but it also takes up the role of foundational competencies and their development in a wide variety of career possibilities in educational measurement. It is also important to emphasize that what this task force has produced is a *consensus* report for this particular moment in time. Each task force member has their own perspective on what should or should not be a foundational competency, how these competencies are manifested professionally, and how they can be best developed through formal educational opportunities. However, to our surprise, we found that we agreed far more than we disagreed. I am proud of the consensus we were able to build, and I want to acknowledge and thank Andrew Ho for the important role he played in helping to facilitate this process as task force chair.

I hope this report will prompt fruitful discussion and debate about what it means to be a professional in the field of educational measurement. I hope it will also prompt some reflection and introspection about *what, who, how, why, and when* we measure. And I hope it will help the field evolve in a way that makes all of us proud to be a part of it.



Derek Briggs  
Past President, National Council on Measurement in Education  
(Presidential Term: 2021-22)

## **NCME Task Force on Foundational Competencies in Educational Measurement**

### **Task Force Charge**

1. To develop a set of foundational competencies for the field of educational measurement.
2. To illustrate one or more curricular models for a graduate program in educational measurement.
3. To engage NCME membership and the field with Task Force findings through conference presentations and published journal articles.

### **Process**

Task Force members met monthly from October 2021 through July 2022 to develop a draft consensus framework for foundational competencies, discuss curricular models, and consider how different career paths require foundational competencies. The Task Force released a draft report for member comment in September 2022 and held webinars soliciting feedback from NCME members. The Task Force met in late 2022 and early 2023 to respond to feedback. This report reflects the consensus of the Task Force.

### **Guiding Principles: What does “foundational competencies in educational measurement” mean?**

Through its discussions, the Task Force came to consensus about the nature of its charge under six broad principles:

1. Foundational competencies support future development of additional professional and disciplinary competencies in educational measurement. Foundational competencies need not be an exhaustive list of competencies. They are a foundational subset of a fuller set of competencies that educational measurement experts can possess.
2. Foundational competencies in educational measurement overlap and interact with competencies in other professions and disciplines. Some foundational competencies in educational measurement may also be foundational competencies in other professions and disciplines.
3. Foundational competencies overlap and interact with each other. They are not a discrete list; many are characterized by intersections and interactions.
4. Foundational competencies are both descriptive of the profession and discipline, and aspirational about the future of the profession and discipline.
5. Foundational competencies support educational measurement, broadly conceived. Educational measurement competencies are those that support the design, use, and evaluation of measures of cognitive, affective, and psychological constructs that individuals and groups develop in formal schooling, training, and other learning environments.
6. Foundational competencies in educational measurement intersect with important general dispositions and mindsets for learners and professionals that are not unique to educational measurement, including critical thinking, intellectual humility, meta-cognition, creativity, flexibility, and openness and willingness to critique.

Part 1 of this report provides description, justification, and examples of competency domains and subdomains. Part 2 provides examples of how careers in educational measurement both require and develop these foundational competencies. Part 3 proposes a curriculum and illustrates how activities within a curriculum can help students to develop foundational competencies.

## Task Force Provisional Definitions of Terms

To achieve its charge, the Task Force found it helpful to provide provisional definitions of common terms. While there remains debate among Task Force members about these definitions (reflecting further debate in the field), this glossary may nonetheless provide clarity about Task Force intentions.

**Statistics** is the science of describing and modeling physical and social phenomena using data to improve prediction and understanding.

**Psychometrics** is a field of study in psychology and education characterized by statistical modeling of latent variables motivated by psychological theory.

**Measurement** is a systematic process of data collection using instrumentation that results in a quantity intended to support inferences about an attribute or property of an object, event, or phenomenon.<sup>1</sup>

**Assessment** is the process and outcome of collecting and analyzing data to inform an interpretation, judgment, or decision about an attribute or property of an object, event, or phenomenon. Assessment can involve but does not necessarily require measurement.

**Testing** is the development and deployment of an instrument and scoring procedure that results in a categorization or quantity that may support inferences about an attribute or property of an object, event, or phenomenon. Testing can sometimes but not always produce a measurement.

**Education** is a process or system for improving human competencies through learning.

**Educational measurement** involves measurement of knowledge, skills, dispositions, and abilities for some educational purpose, such as supporting learning, certifying learning, or identifying policies and practices that improve learning.

**Educational measurement careers** are careers that include professional responsibilities distinguished by expertise in educational measurement.

**Educational measurement programs** are formal academic programs that develop and certify educational measurement competencies.

**Fairness** is the extent to which a measurement process and score use maximizes opportunity for all respondents to demonstrate their capabilities with respect to the construct of measurement.

---

<sup>1</sup> Other published definitions of measurement include “the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity” (*Measurement in Psychology* by Joel Michell, 1999); “the assignment of numerals to objects or events according to rules” (*On the Theory of Scales of Measurement* by S. S. Stevens, 1946); and “an activity of classification, ordination, or quantification of a set of elements according to a model of a relevant attribute in service of a larger goal” (*A Pragmatic Perspective of Measurement* by David Torres Iribara, 2021).

A **learner** is a student or professional in the field of educational measurement who is in the process of improving their knowledge, skills, and abilities. We are all learners.

An **educational measurement student** is a learner who is enrolled formally in an educational measurement program.

An **educational measurement professional** is a learner who is employed in an educational measurement career.

## Part 1: A Framework for Foundational Competencies

The Task Force identified three unordered and overlapping competency domains:

- Domain 1: Communication and Collaboration Competencies
- Domain 2: Technical, Statistical, and Computational Competencies
- Domain 3: Educational Measurement Competencies

The first two domains are not unique to educational measurement. However, developing competencies in these two domains that are relevant to the field of educational measurement requires special training and experiences.

The third domain includes five competency subdomains that are more particular to educational measurement:

- Subdomain A is an *overarching* competency subdomain related to social, cultural, historical, and political contexts in which educational measurement occurs.
- Subdomain B is an *undergirding* competency subdomain related to validity, validation, and fairness.

There are three additional unordered and overlapping subdomains that address the following theoretical concepts and technical skills within educational measurement:

- Subdomain C: Theory and Instrumentation
- Subdomain D: Precision and Generalization
- Subdomain E: Psychometric Modeling

Figure 1 illustrates these unordered and overlapping competency domains (1-3) and subdomains (3A-3E). Appendix A includes an alternative visualization that emphasizes overlap and intersections.

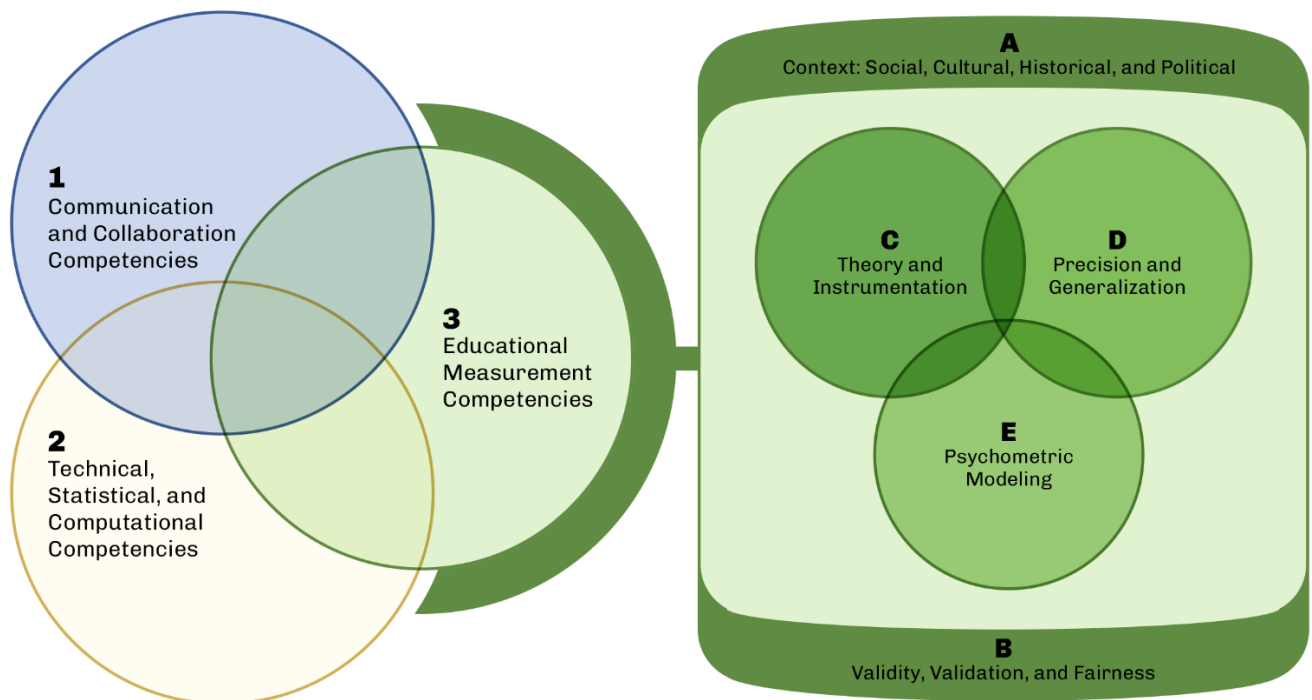


Figure 1. A framework for foundational competencies in educational measurement

## Domain 1: Communication & Collaboration

**Description:** Communication competencies in educational measurement refer to the ability to describe measurement processes and procedures; present findings from psychometric analyses, statistical analyses, and validation studies; and share interpretations of score reports through multiple media to a variety of audiences. Collaboration competencies include the skills required to work in a constructive manner with other professionals and practitioners in education, psychology, computational sciences, and technology development. Collaboration skills involve not only the ability to get along with others but also co-creating solutions in ways that synthesize or build on ideas offered by other team members. 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.

**Justification:** Educational measurement is a collaborative endeavor that requires people with varied skill sets to work together to design, develop, administer, and evaluate instruments that satisfy specific uses. As new methods are integrated into the field, collaboration with experts in other fields becomes increasingly important. Productive collaboration requires effective communication. Communication is also essential for supporting valid interpretation and use of educational measurements by end-users.

**Illustrative Examples:** Foundational competencies in communication and collaboration enable learners to present findings from psychometric analyses or validation efforts to both technical and general audiences. These competencies support learners as they build consensus about how to define constructs, measure constructs, and report scores. Communication competencies can also improve clarity and precision in item writing and task design. These competencies can enable collaborations with graphic artists, marketing team members, and web designers to produce an interactive report that supports valid interpretations and uses of test score information. These competencies are the foundation of pedagogical and presentational competencies that enable teachers and presenters to engage audiences at all levels of understanding.

## Domain 2: Technical, Statistical, & Computational Competencies

**Description:** Statistics is the science of describing and modeling physical and social phenomena using data to improve prediction and understanding. Educational measurement requires competency in a variety of statistical and research methods, including sampling theory and methods, exploratory data analysis, computational approaches to parameter estimation, multilevel modeling, Bayesian methods, and experimental and quasi-experimental methods for causal inference. Technical skills include the ability to use statistical software to manage and transform data, design and conduct simulations, generate reports and preregister and test hypotheses. Computational competencies include the ability to write software code and programs, and the ability to understand the logic and purpose of algorithms.

**Justification:** Measurement typically results in numeric values and associated estimates of uncertainty. These estimates of uncertainty are formalized using probabilistic models. As many measurement endeavors occur at a large scale, practical application of statistics and measurement requires fluency within one or more computing and statistical software environments (e.g., R, Python, Stata, SAS). With advancing technology, increasing access to data, and advances in Artificial Intelligence and Natural Language Processing, computational competencies are becoming more important for educational measurement professionals to develop. Because many educational tests are administered digitally, educational measurement experts may develop assessment environments that monitor, record, and score examinee-item interactions.

**Examples:** A competent educational measurement professional can develop, adapt, and evaluate statistical models and computational algorithms for educational measurement applications. Foundational competencies in this domain would enable a professional to understand the purpose and use of automatic content generation using computational language models while identifying possible sources of bias. Foundational competencies may also support a professional in gathering cognitive process data from digital environments to understand how examinees may arrive at answers, to design experiments to test whether engagement with certain item types leads to different educational outcomes, and to evaluate the results of such experiments using appropriate statistical methods.

### **Domain 3: Educational Measurement Competencies**

Foundational competencies in educational measurement include five subdomains: A) an overarching subdomain related to the social, cultural, historical, and political context of measurement; B) an undergirding subdomain related to validity, validation, and fairness; C) theory and instrumentation, D) precision and generalization, and E) psychometric modeling. Subdomain A is "overarching" because these contexts frame and suffuse measurement processes, analyses, and reporting. Subdomain B is "undergirding" because it is the basis for motivating, evaluating, and improving measurement activities. Together, competencies in these subdomains support common educational measurement efforts, including designing and developing measurement instruments, scoring responses, estimating and reporting score precision, establishing performance standards, and ensuring that scores are comparable through the use of scaling and equating methods. Educational measurement professionals use these competencies to evaluate and improve the validity, reliability, and fairness of scores for their intended and enacted purposes.

#### ***Subdomain A: Social, Cultural, Historical, Political, and Learning Context***

**Description:** Context competencies support and frame learning and activities not only in educational measurement but also in Domains 1 and 2. Placing them as an overarching subdomain for educational measurement emphasizes the responsibility of educational measurement learners and professionals to develop and advance these competencies both within this field and beyond it. Context competencies for educational measurement include the ability to identify social, cultural, historical, and political factors that influence and intersect with the measurement process and may affect the definition of constructs, respondents' interactions with measurement instruments, the interpretation of responses and scores, and the appropriate interpretation, use, and communication of results and findings. Learners with competencies in this subdomain can account for these factors to improve the likelihood of valid interpretations and intended effects and minimize the likelihood of unintended negative effects.

- **Social Context:** The social structure in which a respondent is situated influences their opportunities, expectations, and norms in ways that affect their interaction with measurement instruments. Relevant structures include schools, classrooms, families, professions, and neighborhoods. Relevant examples include social structures based on race, gender, class, culture, language, and disability.
- **Cultural Context:** The cultures in which respondents live and learn influence their ways of knowing, communicating, and interacting, as well as their beliefs, values, and world views. These in turn influence how respondents interact with measurement instruments and how users interpret scores.

- *Historical Context:* A respondent's experience with a measurement or beliefs about a construct can affect their subsequent interaction with measurement instruments. A social group's history with measurement can also affect respondent engagement. The history of educational measurement, which includes the misuse of intelligence tests to justify racist policies and practices, is essential context for the design, deployment, and reporting of educational measurement procedures.
- *Political Context:* Educational measurements can serve multiple political goals at different levels of educational systems. This political context can influence, positively and negatively, the development, use, and resulting properties of measurement instruments.

**Justification:** Sociocultural theories of learning emphasize the importance of the context in which educational measurement occurs. Educational measurement also serves increasingly varied purposes for increasingly diverse populations, requiring respondents with diverse social positions, cultures, and histories to interact in an engaged manner with an instrument. Designing engaging instruments and administration conditions, interpreting and rating responses, and communicating results requires educational measurement specialists to be responsive to and inclusive of the diverse social, cultural, and historical influences respondents bring to their interactions with instruments. Instruments and testing programs must also respond to the political needs that motivated their development. These contextual factors can influence test scores and must therefore inform test score interpretation and use.

**Examples:** Learners who are competent in this overarching subdomain can identify and recognize the importance of identifying some of the relevant social, historical, and political factors in common testing applications including accountability testing, admissions testing, certification exams, or classroom assessment. They understand the importance of developing bias, sensitivity, and accessibility guidelines that are responsive to the social, cultural, and historical contexts of the intended respondents. This includes understanding the importance of designing and administering items that maximize construct-relevant engagement and minimize construct-irrelevant bias. Competent learners also understand how experiences with poorly designed tests and score reports can themselves be harmful by reinforcing negative perceptions and stereotypes, among respondents about themselves, or among other score users about respondents and respondent subgroups.

### ***Subdomain B: Validity, Validation, and Fairness***

**Description:** This competency relates to learners' abilities to state intended interpretations and uses of test scores, and to produce and evaluate theory and evidence supporting these interpretations and uses. A competent learner can identify and use different sources of evidence to construct or evaluate a validity argument. A competent learner can evaluate evidence about the fairness of interpretations and uses of tests and test scores in education.

**Justification:** Validity is the foundational consideration underlying the interpretation and use of test scores. A principal effort of educational measurement scholars and practitioners is to produce and evaluate validity evidence, an activity known as validation. Validation also requires evaluating whether uses and interpretations of educational test scores are fair for individuals and subgroups and supported by evidence and theory.

**Examples:** A competent learner can explain how and why multiple sources of evidence are useful to support valid and fair uses of educational test scores for different purposes. For example, a competent learner can explain whether, when, and why a) content alignment is important for a test for educational accountability, b) the correlation between test scores and college grades is important evidence to support the use of an admissions test, or c) internal consistency is important for the use of a diagnostic screening test. The competent learner would also be able to identify additional sources of evidence that would further improve the argument for the validity of score interpretations and uses, including evidence that the scores and score uses are fair for different subgroups.

### ***Subdomain C: Theory and Instrumentation***

**Description:** Developing a measure of a construct in education requires a theory of how learners learn within the relevant subject area, content, or professional domain. These theories guide instrument development and the design of validation studies. Such studies should include evidence that instruments are sensitive to variation in the levels of the construct and can support intended interpretations and uses. Learners with this competency understand theories about learning within these domains and/or understand the importance of collaborating with and including those who possess this experience and expertise.

Learners with this competency have experience applying principled approaches to test design and development. This may include experience developing test specifications and blueprints; defining performance level descriptions; and authoring, generating, or evaluating items, tasks, and scoring rubrics. Learners may also have experience assembling items and tasks in fixed-format or dynamic environments; adhering to evolving content, bias, sensitivity, security, and accessibility guidelines; and authoring manuals for administration and technical documentation that summarize the evidence relevant to intended score interpretations and uses.

**Justification:** Theories of learning should guide instrument development in education and the collection of evidence that the instruments are distinguishing among levels of the construct as developers intend. Sound instrument design and development is a critical component of validity evidence. Knowledge of instrument design and development processes, procedures, and principles is foundational for modern approaches to assessment design and test development. Digital and computational approaches open new possibilities for instrumentation and item generation, and this may lead to deeper interactions with computational competencies. Just as advances in learning theories may lead to new approaches and methods of instrumentation, advances in instrumentation may lead to novel insights that lead to changes in learning theories.

**Examples:** A competent learner can develop or collaborate with others to articulate a theory of learning to guide task development toward identifying variation in levels of a construct. A competent learner understands the importance of working with content experts to develop a construct definition based on this theory that guides validation efforts. A competent learner may have experience creating items and tasks that align with theoretical models of learning and cognition. This may include using, developing, and critiquing test specifications, content guidelines, scoring approaches, and procedures for evaluating bias, sensitivity, and accessibility. Competent learners may have experience seeking and using theory to guide computational methods for item generation, test scoring, and validation.

### ***Subdomain D: Precision and Generalization***

**Description:** A competent learner in this subdomain can state the intended extent of generalization of test scores and can estimate and interpret corresponding indices of precision. A competent learner can identify common targets of generalization in educational measurement, including generalization to other items, raters, occasions, and aggregate scores, and provide examples of their corresponding reliability coefficients and error estimates. A competent learner can also identify the evidentiary limits of generalization and design studies to expand the evidence base for generalization in support of desired uses. A competent learner also understands how scale score properties and combination procedures interact with precision and generalization, including how transforming, averaging, or differencing scale score units can affect score precision and interpretation.

**Justification:** Valid score interpretations and uses require an understanding of the degree of score precision and the extent to which a score can support a generalized inference about the construct of measurement. Educational measurement scholars and practitioners distinguish themselves by their experience and expertise in explaining the nature of measurement error, estimating error variance, and anticipating its consequences. Educational measurement scholars and practitioners also understand and know how to minimize threats to score comparability and interpretation through appropriate use of scaling and equating methods.

**Examples:** A learner who has foundational competency in this subdomain understands how to distinguish the concepts of reliability and measurement error from the statistical models that are used to estimate them. Practically, this entails the ability to estimate more than one type of “reliability coefficient,” and explain how each has different limits of generalization. Such a learner can also estimate and interpret different standard errors corresponding to each desired generalization. Foundational competencies enable measurement professionals to estimate and communicate error in complex situations, including degrees of precision that are heterogeneous across groups and conditional on proficiency levels.

### ***Subdomain E: Psychometric Modeling***

**Description:** Psychometric models play a critical role in instrument design and development and in formalizing and evaluating concepts such as precision, uncertainty, reliability, generalizability, invariance, and comparability. Psychometric models are important tools for investigating hypotheses about relationships among the measured construct, item characteristics, and external variables. Learners with this foundational competency can select, fit, evaluate, and interpret results from multiple well-known statistical and psychometric models. Students with this foundational competency should be able to explain similarities and differences among classical test theory, item response theory, and factor analytic models; understand the assumptions these latent variable models make; and understand whether and how to evaluate assumptions underlying models and methods. These psychometric models may complement or overlap with methods from other domains, including statistical models like mixed effects models and computational methods using Artificial Intelligence and Natural Language Processing.

**Justification:** Large-scale educational measures use modern psychometric models well-suited for intended score interpretations. Selecting from among these psychometric models and using relevant model diagnostics and parameter estimates to improve score interpretation and use for educational purposes is a distinguishing competency of educational measurement professionals.

**Examples:** Someone with foundational competency in this area should be able to, for example, identify the relative strengths of and interrelationships between classical test theory and item response theory models, suggest a set of statistical or psychometric models to evaluate items or score examinees, and suggest how to assess and monitor whether the recommended model is serving its intended purpose. The competent learner may also understand how psychometric models intersect and interact with methods from other domains, including statistical models like mixed effects models and computational methods using Artificial Intelligence and Natural Language Processing.

## **Part 2: Foundational Competencies in Educational Measurement Careers**

Part 2 of this report demonstrates how measurement professionals use foundational competencies and continue to develop them in educational measurement careers. The field of educational measurement encompasses many diverse career roles, and some foundational competencies are more relevant in certain roles than others. This section defines careers in educational measurement, outlines possible career pathways for professionals trained in educational measurement, characterizes the work that requires foundational competencies, and provides examples of competencies in industry and the academy.

### **What defines a career in educational measurement?**

Educational measurement involves measurement of knowledge, skills, dispositions, and abilities for some educational purpose, such as supporting learning, certifying learning, or identifying policies and practices that improve learning. A career in educational measurement is conceived broadly as work that supports the design, use, and evaluation of measures of cognitive, affective, and other psychological constructs developed in educational, training, and learning environments.

Career pathways for those trained in educational measurement vary and offer a wide range of opportunities and experiences that require and build on foundational competencies. This includes work in K-12 assessment, higher education, licensure and certification, government, research, consulting, advocacy, education technology, philanthropy, and international organizations. Many professionals are self-employed. Established professionals also serve on governing boards and advisory committees for measurement efforts. A doctorate is typically required for academic careers and for many leadership positions in non-academic organizations. Graduates with Master's degrees can work in supporting roles as data analysts and researchers and occasionally manage programs that use test scores. Some educational measurement professionals work on teams with similarly trained and competent colleagues, whereas others are the sole or lead measurement expert responsible for educational measurement activities. Educational measurement professionals can play a critical role by advocating for measurement perspectives and principles that others on their team or in their organization may not have.

### **How do foundational competencies manifest in educational measurement careers?**

Foundational competencies cover a wide range of essential knowledge and skills. A recent graduate with either a Master's degree or a Ph.D. would not be expected to have developed full expertise in all of these areas. Rather, graduate school training is a starting point. Expertise in these foundational areas is developed, often in collaboration with others, while on the job—through opportunities to offer training and professional development, by consulting on educational measurement projects and funding proposals, and by conducting research. These applied work opportunities enable educational measurement professionals to develop expertise in foundational and other job-related competencies over time.

Applied measurement rarely conforms to theoretical models and idealized assumptions. On-the-job application of these foundational competencies often takes place within a set of operational constraints, political and social contexts, and financial uncertainties. There may be no easy or obvious solution nor predictable effects of measurement decisions and approaches. Thus, workplace applications of educational measurement often require measurement professionals to integrate different

competencies to consider alternatives and constraints. Educational measurement programs can facilitate this transition by incorporating real-world examples into coursework and requiring students to make recommendations while balancing one or more operational constraints.

### **How do competencies continue to develop throughout educational measurement careers?**

Measurement professionals should expect to grow in each of these domains and subdomains throughout their entire career. Advances in society, measurement organizations, and the broader measurement field require measurement professionals to learn new norms and practices. The foundational competencies in this report should support this learning and evolve over time to support new uses and contexts for educational measurement.

Recent years have illustrated the importance of foundational competencies and how likely it is that they must continue to adapt. For example, communication and collaboration competencies have become more salient as remote work policies rise in popularity (Domain 1). Digital assessment environments increasingly provide rich data that require increasing computational competency to understand and analyze (Domain 2). Sociocultural models of learning and critical social theories such as Critical Race Theory and Intersectionality Theory can have important implications for context and fairness (Subdomains A and B). Society, context, and scholarship will continue to interact to demand both reconceived and new competencies in educational measurement careers.

Mentorship, professional organizations, and scholarship are three ways to continue developing competencies in educational measurement careers. Like good instructors, good mentors take the time to explain the “why” underlying measurement decisions. They can identify connections among foundational competencies and explain how competencies interrelate to inform measurement decisions. Mentors also can provide career guidance, identify opportunities for research, and help professionals to make connections within and outside of their organization to expand their network and knowledge. Professional organizations similarly connect measurement professionals to ongoing scholarship and current practices. Active participation and consumption of scholarship also requires measurement professionals to continue to develop from foundational competencies in educational measurement.

### **How can different educational measurement careers require foundational competencies?**

Although the specific job requirements in these careers can vary, the job requirements of educational measurement careers have many overlapping characteristics. A focus on communication, data analysis and computing, and measurement are core threads running through most measurement jobs. The following lists describe duties in the form of illustrative job descriptions across K-12 assessment (Example 1), educational technology (Example 2), and licensure and certification (Example 3), as well as research careers in non-profit organizations (Example 4) and universities (Example 5). These job descriptions are adaptations and amalgams of those that some measurement organizations post on NCME listservs.

While core measurement activities are often similar across organizations, the focus of that work is likely to vary. This focus will likely be driven by the clients that each organization serves. For instance, K-12 assessment work will focus heavily on adherence to state and federal requirements and guidelines. Educational technology work will focus on supporting product development. Licensure and certification will focus on accreditation and/or industry needs. The nature of the work can vary substantially,

including the size and scope of the measurement work (sample sizes, instrument types and times), and the degree of measurement representation and expertise in the organization.

Most domains and subdomains described earlier are well-represented across these job duties. The job descriptions themselves illustrate how the domains are related and overlapping in the work, so it can be difficult to map some duties cleanly to a single domain. Communication and collaboration (Domain 1) comprise a large portion of the job duties, with a strong emphasis on communicating effectively to a variety of internal and external audiences with an emphasis on the ability to communicate research findings. Educational measurement (Domain 3) comprises a large portion of job duties for psychometric jobs. The statistical and technical domain (Domain 2) appears in job descriptions in ways that can be difficult to disentangle from measurement competencies. Because of the variety of positions in the industry, these descriptions represent an illustrative not exhaustive list.

### ***Example 1: Illustrative Duties of Psychometricians in K-12 Assessment Companies***

#### **Domain 1: Communication and Collaboration Competencies**

- Produce technical documentation related to item, test, and program performance
- Interact with clients and technical committees for standard setting, measurement explanations and other appropriate topics
- Support the writing of research proposals or proposals for funded work
- Present and publish theoretical and/or application papers in conferences and journals
- Coordinate and collaborate with members of content, technology, and program management groups to plan and implement work
- Participate in internal training and individual development of technical skills

#### **Domain 2: Technical, Statistical, and Computational Competencies**

- Design and implement research projects
- Plan, coordinate, and perform statistical designs and reviews of designs
- Apply advanced knowledge of statistical procedures and their applications
- Use statistical programming tools
- Manipulate and validate specifications, data files and reports

#### **Domain 3: Educational Measurement Competencies**

- Develop and implement a plan that ensures that examination services comply with industry standards for security, validity, reliability, fairness, and transparency

### ***Example 2: Illustrative Duties of Psychometricians/ Researchers in K-12 Educational Technology Companies***

#### **Domain 1: Communication and Collaboration Competencies**

- Collaborate with internal teams to understand their goals
- Prioritize among the needs of psychometric, research, and data support teams
- Develop and improve data visualizations for different audiences
- Collaborate with product research teams to improve understanding of data sources and their relevance for score interpretations

#### **Domain 2: Technical, Statistical, and Computational Competencies**

- Apply knowledge of data visualization, statistical procedures, psychometric methods, and statistical programming

- Develop quality control protocols for deliverables
- Evaluate tradeoffs among statistical analyses and make recommendations
- Use statistical and computational programming skills to solve emerging problems
- Improve file organization for collaborative access that adhere to data privacy guidelines
- Research and engage in data science activities

Domain 3: Educational Measurement Competencies

- Perform psychometric analyses
- Provide technical information about items to assessment content development teams
- Create visualizations for test scores for educator dashboards
- Review and revise technical documentation
- Develop new measurement models and methods that advance scholarship and practice
- Demonstrate new applications of existing measurement models for practical problems

***Example 3: Illustrative Duties of Psychometricians in Licensure and Certification Organizations***

Domain 1: Communication and Collaboration Competencies

- Develop short and long-term schedules for psychometric activities
- Work with the information technology and quality assurance departments to identify psychometric principles for developing online testing environments
- Represent measurement interests on cross-departmental teams
- Facilitate standard setting and key validation meetings
- Represent the organization on assessment and psychometric and research issues with internal and external constituents
- Conduct presentations at psychometric and measurement-related conferences

Domain 2: Technical, Statistical, and Computational Competencies

- Assist in accreditation processes by supplying relevant technical information

Domain 3: Educational Measurement Competencies

- Identify skills required by jobs and roles; design tasks that require these skills
- Conduct and evaluate psychometric analyses on examination programs including IRT calibrations, scaling and equating projects, item analysis and key validation, standard setting, item bank analyses, and test form construction
- Conduct psychometric quality assurance processes on exam related data queries and reports
- Assist in efforts to improve validation procedures according to new standards in the field
- Read and advocate for adoption of recent relevant research
- Develop new measurement models and methods that advance scholarship and practice
- Demonstrate new applications of existing measurement models for practical problems

***Example 4: Illustrative Duties of Research Associates in Educational Research and Consulting Companies***

Domain 1: Communication and Collaboration Competencies

- Participate in a team that values diverse voices and ideas, prioritizes codesign of deliverables, and emphasizes team building, learning, and professional growth
- Write reports that inform the decisions of educational policymakers and practitioners
- Contribute to the creation of proposals and outreach efforts that advance organizational goals

- Help organize and design presentations, trainings, and other dissemination materials or activities
  - Write and format summary tables, graphs, presentations, and reports
  - Mentor and develop junior staff
- Domain 2: Technical, Statistical, and Computational Competencies
- Design and implement research and analysis plans
  - Analyze data using a variety of statistical software
  - Clean complex data sets and conduct descriptive analyses
- Domain 3: Educational Measurement Competencies
- Conduct reviews of academic literature or internal documentation
  - Develop or evaluate assessments, including cognitive and non-cognitive measures
  - Develop new measurement models and methods that advance scholarship and practice
  - Demonstrate new applications of existing measurement models for practical problems
  - Conduct classical and IRT analyses, test scoring and equating, and standard setting
  - Synthesize information for preliminary and final deliverables

***Example 5: Illustrative Duties of University Faculty in Educational Measurement***

- Domain 1: Communication and Collaboration Competencies
- Teach and mentor students
  - Write, publish, and present research findings at conferences and colloquia
  - Collaborate on research projects and disseminate results in co-authored publications, such as articles
  - Advise measurement practitioners in testing organizations or district, state, and federal educational agencies
  - Serve in membership and leadership roles in academic and professional organizations
- Domain 2: Technical, Statistical, and Computational Competencies
- Apply and improve statistical methods for research and practice
  - Develop and improve software, code, or other computational procedures to enable others to replicate research or apply new methods
- Domain 3: Educational Measurement Competencies
- Develop and evaluate educational measurement instruments
  - Develop new measurement models and methods that advance scholarship and practice
  - Demonstrate new applications of existing measurement models for practical problems

**How can specific educational measurement career scenarios require foundational competencies?**

The previous examples review duties and requisite competencies in educational measurement organizations across a range of contexts. The following examples illustrate the competencies as they may appear in professional settings in four scenarios. Two scenarios focus on academic contexts, with an emphasis on mentoring graduate students and participating in the broader educational measurement community. The other two scenarios describe hypothetical situations in operational measurement and illustrate how competencies support productive responses to these situations.

These scenarios illustrate that the competencies appear in many day-to-day educational measurement tasks. These also provide informative detail, particularly for those considering educational measurement

careers, about the kinds of issues measurement professionals encounter as part of their work. This is a limited list, but it begins to illustrate the breadth of tasks that educational measurement professionals can encounter.

### ***Example 1: Mentoring Graduate Students to Become Educational Measurement Professionals***

Graduate students in educational measurement benefit from introduction into and connection with the broader measurement community (Domain 1). They can supplement their coursework by interacting and networking with researchers and practitioners in the measurement field. For example, they could do the following:

- Become familiar with and submit research papers to journals within the measurement community
- Become familiar with measurement resources provided by NCME, such as ITEMS, Formative Assessment Modules, and the Software Database (Domains 2 and 3)
- Attend and present at regional, national, or international measurement conferences.

Faculty can support the development of Domain 1 skills when they:

- Incorporate articles from measurement journals into course requirements and introduce students to measurement journals in the field
- Encourage students to become members of regional and national measurement organizations
- Fund students to attend annual meetings of measurement organizations
- Co-author measurement research articles that can be presented at annual meetings, having the student make the presentations
- Provide feedback on papers in courses, encourage students to turn conference papers into journal submissions, and introduce the student to the publication process
- Include students in grant writing projects from submitting a proposal to conducting the required research to submitting final reports

### ***Example 2: Senior Faculty Contributions to Measurement***

Senior faculty in educational measurement need to play an active role in contributing to the foundational competencies and shaping future directions, standards, and policy for the measurement field. This includes the need to continue their momentum in research, teaching, and service (Domain 1), and continue to have a meaningful impact within the measurement community, contributing to both measurement theory and practice (Domains 1, 2, and 3). Some of the ways they accomplish these goals include the following:

- Actively mentor students by guiding them through their dissertation research (Domain 1)
- Actively pursue external funding from government or private funding agencies to support their research efforts and the efforts of their students and colleagues (Domain 1)
- Assume an active role in serving as a reviewer or editor of measurement journals (Domains 1, 2, and 3)
- Contribute to ITEMS, NCME Formative Assessment Modules and Software Database (Domains 2 and 3)
- Support colleagues who may have less educational measurement expertise
- Serve in leadership roles in professional organizations
- Serve on Technical Advisory Committees for licensure, certification or measurement/research companies or state educational testing programs (Domain 1)

The next two scenarios are based on the experiences of task force members working in industry.

***Example 3: Helping to Solve an Operational Test Design Issue***

Measurement specialists enter work with a testing program at various stages, sometimes after the test is initially developed. In this example, an English Speaking test includes 12 items, each of which belongs to one of the following 4 types:

- 3 × Read a text aloud,
- 3 × Describe a photo,
- 3 × Listen to an audio clip and answer questions shown on the screen, and
- 3 × Share your view on a topic.

Test administrators observed that test users found the 3rd item type to be too difficult and wanted to drop one of the three items of that type. They asked the measurement professional for their opinions about the consequences and reasonableness of the move. What analyses should the measurement professional perform, and how should they respond to this request?

In this scenario, the measurement professional would be expected to:

- Discuss with subject matter experts whether the change in content raises questions about the validity of score interpretations in the context of intended score uses (Domain 1, 3A, 3B, & 3C)
- Discuss estimation of reliability coefficients using past data for 11-item tests that one would obtain after omitting one item of Type 3 (Domain 3D)
- Suggest inclusion of one or two items of other types based on item statistics and estimate reliability of the hypothetical new test, possibly using simulations from an IRT model (Domain 3E)
- Explore what may be causing increased difficulty with the 3rd type of item and consider a modification to make that part of the domain more accessible (Domain 3C)
- Think about how to equate test scores from the old to the new form after the change (Domain 3E)
- Communicate the methodology and findings to the program managers in ways that support decision-making (Domain 1)

***Example 4: Helping to Solve an Operational Item Scoring Issue***

Sometimes, measurement specialists engage in very specific tasks, such as evaluating the quality of constructed-response item scoring. In this example, an automated scoring engine was used alongside human scoring in a testing program to score short constructed response items. The engine and the human scorers show low levels of agreement on one type of item across grades. What analyses should the measurement professional perform, and how should they respond?

In this scenario, the measurement professional would be expected to:

- Examine human rater performance to the expert ratings used to monitor raters (Domain 3B)
- Identify potential sources for or patterns of differences in rater performance in context, particularly related to the item type with low agreement (Domain 1, Domain 3A & 3C)
- Examine the samples used to train the scoring models versus those observed in operational testing to determine whether and how they differ (Domain 2, Domain 3D)
- Determine whether retraining the engine could help to ameliorate the issue, using different predictive features or an expanded training and validation dataset (Domain 2, 3E)
- Share findings with stakeholders and propose next steps (Domain 1)

### **Part 3: Foundational Competencies in Educational Measurement Courses, Programs, and Activities**

Educational measurement programs and faculty develop students' foundational competencies by designing course sequences, selecting and sequencing course topics, and developing through-course, end-of-course, and comprehensive assessments. Programs and faculty also develop competencies through co-curricular structures and supports, including mentoring, research assistantships, teaching assistantships, internships, colloquia, and engagement with professional organizations. The Task Force chose to meet its charge to, "illustrate one or more curricular models for a graduate program in educational measurement," by providing three illustrations: A) how a program's curricular and co-curricular structures can develop each of the foundational competencies, B) one possible design for a two-semester sequence in educational measurement, and C) how curricular activities can develop foundational competencies.

Programs in educational measurement vary in size, focus, and mission. The Task Force proposes the following curricular structures, sequences, and activities as illustrative, not prescriptive. Educational measurement programs can develop foundational competencies in a variety of ways.

#### **Where in a curriculum can programs develop students' foundational competencies?**

**Domain 1. Communication and Collaboration:** Although some elective courses may focus on specific skills like communicating test scores to various audiences, programs develop general Domain 1 competencies by giving students experience with and feedback on presentations and collaboration in courses and through co-curricular activities like colloquia and internships. To develop this foundational competency, course instructors include final projects or presentations and partnered work in their courses. Instructors provide students with explicit guidance and feedback to help students improve the effectiveness of their written and oral communication and collaboration. Faculty mentors also introduce their advisees into professional networks to improve their opportunities for productive collaboration and communication.

**Domain 2. Technical, Statistical, and Computational Competencies:** Developing these foundational competencies typically requires two to four courses in applied statistics. Applied coursework generally requires statistical programming. Foundational coursework also prepares students for measurement in adaptive digital environments. Advanced and elective coursework further develops necessary competencies for digital measurement work, including machine learning, Artificial Intelligence, and multimodal analytics.

**Domain 3, Subdomain A. Social, Cultural, Historical, and Political Context:** Developing learner understanding of the important interactions between context and measurement requires instructors to intentionally situate measurement methods within these contexts. Although programs can support this competency indirectly through coursework in respective disciplines or a standalone course, integrating examples of social, cultural, historical, and political contextualization into foundational educational measurement coursework is necessary to develop this competency as it applies to educational measurement.

**Domain 3, Subdomain B. Validity, Validation, and Fairness:** 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. In contrast, developing validity, validation, and fairness as an undergirding foundational

competency requires elevating these concepts such that they are visible in all educational measurement activities throughout the curriculum. This subdomain motivates a range of additional methods and techniques related to fairness, including equating and setting performance standards.

**Domain 3, Subdomain C. Theory and Instrumentation:** Practical experience with this foundational competency subdomain in a first-year measurement sequence helps to emphasize the importance of construct definition, motivate the application of measurement models, and demystify the educational measurement process. Further engaging with the design and development of a new measure and validation agenda late in a first-year sequence, in more advanced coursework, and in co-curricular activities can help learners to orient all foundational competencies coherently in support of a common goal.

**Domain 3, Subdomain D. Precision and Generalization:** Foundational competency in this subdomain typically begins early in a first-year measurement course and continues to advance in concert with developing competencies in the subdomain of psychometric modeling. Foundational conceptions of reliability associated with Classical Test Theory and Cronbach's alpha are a common early topic in a first-semester course. Contrasting reliability coefficients with different assumptions and intended targets of generalization should be covered in a first-year sequence, as should related conceptions of precision, such as information from Item Response Theory. Instruction in advanced and elective topics like Generalizability Theory, scaling, and equating can continue to develop this competency over time.

**Domain 3, Subdomain E. Psychometric Modeling:** Learners can begin to develop foundational competencies in psychometric modeling early in a first-year measurement course. Early instruction supports additional development in more advanced and elective courses. A first-year sequence typically introduces Classical Test Theory, Factor Analysis, and Item Response Theory, including opportunities to establish relationships among these approaches, fit models, and interpret results. Psychometric modeling also supports additional measurement efforts beyond what a first-year sequence may cover, including going into greater depth on topics such as differential item or test functioning, equating, and setting performance standards. Advanced and elective courses in psychometric modeling can include diagnostic classification models, hierarchical models, multidimensional models, and other generalized latent variable and mixed effects models.

### **What are examples of sequences of course topics in first-year educational measurement courses?**

The following two-course sequence is an example of one that could serve as a foundation for entry into the field of educational measurement. A full educational measurement program would include many other courses and co-curricular activities. The two courses described below develop many of the foundational competencies outlined in Part 1 of this document.

Although the two courses listed below assume a 13-week semester, each course sequence could be supplemented or reduced to accommodate shorter or longer semesters. Each course also assumes intermediate statistical competency, particularly the second course.

The first course focuses on breadth over depth of coverage and includes subdomains such as theory and instrumentation, precision and generalization, psychometric modeling, and validity, validation, and

fairness. Its structure is premised on a semester-long activity that involves the development and analysis of a test or survey instrument. In some graduate programs, it may not be possible to teach a course narrowly focused on educational measurement. In such contexts, it may be necessary to situate testing within the broader framework of the sorts of instrumentation typical in psychology, sociology, or other disciplines.

The specific topics in the second course are flexible and intended to focus on depth over breadth of coverage. This course can be centered around a specific technique, model, or theory, such as Item Response Theory, Generalizability Theory, diagnostic measurement models, or validity theory. The second course illustrated here expands on the content introduced in the first course by focusing more narrowly on precision and generalization, psychometric modeling within IRT, and applications of IRT. Understanding these topics is reinforced with class activities and assignments of the sort in the following section.

These tables include topics listed by week. Although the topics listed focus primarily on Domain 3 competencies, instructors can design course activities and assignments to promote the development of competencies in Domains 1 and 2, as the next section illustrates.

***Course 1: An Introduction to Educational Measurement or Measurement in Survey Research***

Week	Topics
1	<ul style="list-style-type: none"> <li>● Introduction to Tests and Survey Instruments</li> <li>● Social, cultural, historical, and political context of test and survey instruments</li> <li>● Validity and Reliability</li> </ul>
2	<ul style="list-style-type: none"> <li>● Big Picture Issues <ul style="list-style-type: none"> <li>○ Considering the Context: Why administer a test or survey?</li> <li>○ What are Constructs? What is Measurement?</li> </ul> </li> <li>● Validity, Validation and Fairness (Part 1) <ul style="list-style-type: none"> <li>○ Consensus Definitions in The Standards</li> <li>○ Historical Overview</li> <li>○ Basic structure of validity arguments</li> </ul> </li> </ul>
3	<ul style="list-style-type: none"> <li>● Sampling <ul style="list-style-type: none"> <li>○ Defining the Population of Interest</li> <li>○ Developing a Sampling Frame</li> <li>○ Probability Samples, Pilot Samples, &amp; Convenience Samples</li> <li>○ Sampling Weights</li> </ul> </li> <li>● Review of chance error, sampling distributions, standard error of a mean</li> <li>● Nonresponse Bias</li> </ul>
4-7	<ul style="list-style-type: none"> <li>● Instrument development: <ul style="list-style-type: none"> <li>○ The role of theory</li> <li>○ Designing items for cognitive constructs</li> <li>○ Designing items for affective constructs</li> <li>○ Fairness, diversity, and equity in item design</li> <li>○ Pilot testing, Cognitive Interviews, and Item Review Panels</li> </ul> </li> </ul>

8	<ul style="list-style-type: none"> <li>Item analysis <ul style="list-style-type: none"> <li>Frequency Distributions and Descriptive Statistics</li> <li>Item difficulty and discrimination</li> </ul> </li> <li>Introduction to Classical Test Theory (CTT) <ul style="list-style-type: none"> <li>The concept of measurement error</li> <li>CTT as a model</li> <li>Reliability coefficients</li> </ul> </li> </ul>
	<ul style="list-style-type: none"> <li>Estimating Reliability and Quantifying Measurement Error <ul style="list-style-type: none"> <li>Internal consistency (vs. stability), Cronbach's Alpha</li> <li>Test-Retest</li> <li>The Standard Error of Measurement</li> <li>The Limits of CTT (GT as an expansion)</li> </ul> </li> </ul>
10	<ul style="list-style-type: none"> <li>Introduction to Item Response Theory (IRT): <ul style="list-style-type: none"> <li>Foundational principles and conceptual overview of IRT</li> <li>Models for dichotomous item responses</li> <li>Application with software and data</li> </ul> </li> </ul>
11	<ul style="list-style-type: none"> <li>Estimating and Evaluating IRT Models <ul style="list-style-type: none"> <li>Conceptual overview of IRT estimation procedures</li> <li>Basics of model-data fit</li> <li>Application with software and data</li> </ul> </li> </ul>
12	<ul style="list-style-type: none"> <li>Introduction to Exploratory Factor Analysis (EFA) <ul style="list-style-type: none"> <li>Conceptual overview of EFA</li> <li>Scree Plots and Parallel Analysis</li> <li>Interpreting EFA results</li> </ul> </li> </ul>
13	<ul style="list-style-type: none"> <li>Validity, Validation and Fairness (Part 2) <ul style="list-style-type: none"> <li>Different perspectives on validity, the role of consequences</li> <li>Examples/illustrations of validation studies in practice</li> <li>Evolving views on fairness</li> </ul> </li> </ul>

## ***Course 2: Item Response Theory and its Applications***

Week	Topics
1	Introduction and overview of psychometric modeling. Establish relationship between modeling, theory and instrumentation, and precision and generalization.
2	Review of Classical test theory (CTT)
3	Item Response Theory (IRT): Models for dichotomous data
4-5	Item Response Theory: Models for polytomous items
6	Item Response Theory: Evaluation of model fit, item fit, and person fit
7	Item Response Theory: Parameter Estimation
8	Classical and IRT Methods of Assessing Differential Item Functioning (DIF)
9-10	Equating Using CTT and IRT methods
11	Computer Adaptive Testing and Automated Scoring
12	Vertical Scaling and Growth
13	Multidimensional IRT Models and IRT Applications

## How can curricular activities develop foundational competencies?

This section provides additional specificity about the foundational competencies by illustrating five activities that educational measurement instructors can lead in a first-year course sequence. We locate each of these activities in the foundational curriculum from Part B, which can act as a launching pad for entry into the field of educational measurement.

### ***Curricular Activity 1: The Meaning of Measurement***

The purpose of this activity is to have students engage with the definition and meaning of measurement, to appreciate that there are different perspectives about this meaning, and to understand that different perspectives invoke different commitments related to the philosophy of science. Students should be able to develop their own answers about the sense in which one can claim that test and survey instruments produce measures. The activity should enhance students' appreciation of what makes the "Educational Measurement Competency" domain distinct from other fields.

This activity can span part of the first two weeks of a standard measurement course. In class, the instructor can ask students to spend 10-15 minutes individually writing an answer to the following prompt:

*If you were asked to define the word measurement in a sentence (without consulting any resources from the internet), what would that sentence be?*

*Before arriving at your sentence, please take a few minutes to think about the following: First, what, in your view, distinguishes an activity that results in a measurement from one that does not? Second, how does measurement in educational and psychological contexts compare to measurement in the physical sciences?*

The instructor then pairs students to discuss their definitions. After this discussion, the instructor has each pair report out and assembles all the different definitions. This activity sets the stage for students to engage with selected readings that the instructor can pair with a reading guide and discussion questions. The readings approach the philosophy of measurement from different perspectives and should be written at an entry level for beginning students. The assignment for the next class is to evaluate their own definition of measurement that they provided in the first week from the standpoint of the readings they have just done. This becomes the basis for an additional think-pair-share discussion.

This activity develops foundational competencies by improving student appreciation of the nature of Educational Measurement Competencies (Domain 3). Conversation and collaboration with classmates also improve Communication and Collaboration Competencies (Domain 1) by encouraging perspective-taking about alternative conceptions of measurement. The activity also offers opportunities to discuss Subdomain A relating to historical perspectives and conceptualizations of measurement and Subdomain B by asking students to consider how different conceptualizations of measurement may lead to priorities in validation.

### ***Curricular Activity 2: Sampling Foundations of Classical Test Theory***

Measurement students may have heard of Cronbach's alpha and standard errors of measurement, but few understand its conceptual foundations beyond "reliability is good" and "error is bad." Students with

strong statistical foundations (Domain 2) may understand sampling and standard errors from a statistical perspective but not a measurement perspective. To build intuition about true scores and the Spearman-Brown relationship between errors and test length (Subdomains D and E), instructors can conduct the following in-class activity when introducing Classical Test Theory.

The instructor begins by distributing a 6-sided die to each student along with a standard heads-and-tails coin. The instructor tells students that their first roll is very important: It will be their “true score.” The students roll the die, and the instructor records their “true scores” in a spreadsheet. Then, the instructor tells students that they will interact with a series of “items,” but their responses will contain “error” in the range of  $[-6,6]$ . They simulate this error by flipping a coin and rolling the die again, where the coin indicates the sign of the error (heads=positive, tails=negative), and the die value indicates the error magnitude. The instructor records each student’s error in the spreadsheet along with their implied 1-item observed score following the classical test theory model:  $X = T + E$ . Students then repeat this multiple times, simulating test scores that are the averages of these item scores for test lengths 2, 4, and 8 items long. To further engage students, the instructor can give two prizes, one for the student with the highest observed score (greatest positive bias), and one for the student with the least bias. In large classes, instructors can simulate these scores, but the die-and-coin combination tends to be more memorable and engaging.

The activity improves student intuition around many implications of Classical Test Theory, including:

- Single-item test scores can be extremely imprecise and feel unfair for high-stakes uses (Subdomains C and E).
- Averaging over random errors leads to predictable increases in precision. This intersects with statistical foundations (Domain 2).
- In contrast with introductory statistics, measurement is multilevel, with an emphasis on interindividual differences and true score variance (Domains 2 vs. 3).
- The variance of observed scores is greater than the variance of true scores in a population (Subdomain C).
- This ratio of true to observed score variance, or reliability, improves with test length (Subdomains C and D).
- The correlations between sets of observed scores (split-half reliabilities) likewise improve predictably with test length (Subdomain D).
- Second-order equity: Imprecise tests benefit low-true-score examinees compared to high-true-score examinees (Subdomains C and E).
- Students can discuss what true score and error dice rolls represent and how instrumentation can elicit them (Subdomains A and B).
- Because true and error distributions are discrete and uniform (or uniform not including 0), theoretical standard errors and reliabilities are derivable (Domain 2).
- The instructor can also connect this activity to important conceptual questions in measurement, including where true score variance might come from, and what might cause an error.

### ***Curricular Activity 3: Contrasting Classical Test Theory and Item Response Theory***

To distinguish between classical test theory and item response theory, instructors can engage in a range of activities after presentation of classical test theory models early in measurement sequence. These primarily develop competencies in the psychometric modeling subdomain.

A graphical approach can begin with empirical item characteristic curves using a large dataset with plots of percent-correct on total score for each item. Instructors can emphasize that these are clearly nonlinear and seem difficult to model. Instructors can then show how a log-odds transformation of the y-axis (percent-correct scale) can partially linearize the relationship, and a hypothetical scale transformation of the x-axis (from the total score scale to a latent score scale) might further enable accurate linear predictions and, if theory holds, conditional independence. Allowing items to have different slopes leads to item characteristic curves with additional parameters. This can help to develop insight and competencies in psychometric modeling (Subdomain E).

From these item characteristic curves, instructors can introduce scale anchoring and item maps, to illustrate how theory predicts the order of persons and items on a unidimensional continuum and aligns them assuming a given response probability. This can be the basis of a lab activity using real data where students write brief reports including tables and figures with item parameter estimates and fitted item characteristic curves (Domain 1). Instructors can also provide particular social, cultural, historical, or political context for the data and for the intended interpretations and uses of scores. Students can then discuss the relevance of this context in small groups to help them understand how measurement concepts interact with context (Subdomain A).

In later modules within the course sequence, instructors can further develop student competencies by emphasizing the usefulness of the IRT conditional independence assumption in test design and development. The additive properties of item information and the uses for which developers can deploy particular items can motivate a range of applications, including criterion-referenced testing, adaptive testing, and equating.

#### ***Curricular Activity 4: Theory and Instrumentation***

The purpose of this activity is for students to integrate the concepts they are learning throughout the course and to think about how these interact. In the activity, students develop an instrument to measure an educational or psychological construct of interest to them. This activity is typically completed during a first course in measurement, although it can be repurposed for a second course by focusing on more advanced and/or technical measurement competencies. The activity primarily develops competencies in Subdomain 2, although it also interacts with other domains and subdomains. The activity runs throughout the semester, with students turning in different parts as they learn the associated concepts and analyses. Students receive feedback after completing each part and ensuing class discussions focus on the differences and similarities among students' analyses and how each part of the activity sets the stage for the next.

1. Define the construct to be measured, the theory or process model that explains why people have different levels of the construct, how they progress from low to high levels, and whether the construct operates similarly across different social and cultural contexts. (Subdomains A and C)
2. Why does a new instrument need to be developed? What are similar measures from the past, and why do they not meet existing needs? (Subdomains A and C)
3. Explain how the resulting scale would be used and in what context. In the explanation, be specific about the intended uses and interpretations of the resulting scores. (Subdomain A and Competency B)
4. Discuss and defend the choice of item types and formats. (Subdomain C)

5. Develop items, administer them to a relevant group of examinees, and collect response data. Discuss what resources were used to generate items, why these are appropriate for the construct and population being measured, and possible sources of construct-irrelevant variance. (Subdomains A and C)
6. Conduct an item analysis appropriate to the type of item, for example:
  - a. Item review panels, cognitive interviews, focus groups (Subdomain C)
  - b. Descriptive and classical test statistics (Domain 2; Subdomain C)
  - c. Information from Item Response Theory or Factor Analysis (Subdomain E)
7. Reliability evidence (Subdomain D)
  - a. Discuss the types of reliability evidence that are appropriate for the scale.
  - b. Obtain relevant reliability coefficients and assess the appropriateness of the magnitude.
8. Validity evidence (Subdomain B)
  - a. Outline a simple validity argument for which one can collect evidence. The argument should match the intended use and interpretations of the scale and scores.
  - b. Identify any assumptions underlying the argument and discuss whether these have been met.
  - c. Design and, if possible, conduct the appropriate analyses and discuss the degree to which these support the argument. If results are not as expected, discuss possible reasons for this.
  - d. Discuss other types of validity evidence that would support the argument and how this evidence could be collected.
  - e. Discuss whether there are any issues related to fair use of test scores.

### ***Curricular Activity 5: Psychometric Properties of Gain Scores***

This activity introduces gain scores within a realistic context of working with a local school district developing Student Learning Outcomes to assess student growth and evaluate teachers. The activity consists of two parts. The first component is more applied and revolves around conducting a gain score analysis, communicating results to stakeholders, and considering validity and fairness issues related to using student assessment scores for teacher evaluation. The second component is more conceptual and requires students to build on their knowledge of Classical Test Theory, think critically about the benefits and limitations of gain score analyses, and consider the interplay between instrument design and interpreting results. Instructors can introduce this activity in a first course after covering Classical Test Theory or in a second course to provide the foundation for other statistical and psychometric options for modeling growth like Student Growth Percentiles, regression residuals, Latent Growth Models, and Diagnostic Classification Models.

For Part 1, the instructor provides students with an item response data set with deidentified teacher, school, and student indicator variables and demographic codes for students. The prompt describes the two-wave assessment administration at the beginning and end of the course and asks students to analyze the data; summarize growth at the student, teacher, and school levels; provide measures of uncertainty and reliability around the individual and aggregate growth estimates; provide easy-to-read data visualizations; and provide summaries of subgroup disparities in growth. The instructor should require the ability to program in R or another computational language to analyze the data, thereby

building computational competencies (Domain 2) and psychometric modeling competencies (Subdomain 5). By writing and presenting results to the district stakeholders in groups, students are building capacity in communication and collaboration (Domain 1). By discussing how race or gender gaps in gain scores can cause deficit-based interpretations, or how item-level data can support asset-based score interpretations, students can build capacity in establishing context (Subdomain A). By discussing validity and fairness issues around using student assessments for teacher evaluation, students also build capacity in evaluating validity and fairness (Subdomain B).

For Part 2, after summarizing the benefits and limitations of gain scores, the instructor provides students with an R script that simulates data under classical assumptions with specified arguments, including the number of items, the number of testing occasions, item properties, sample properties, and the correlation between time points. The instructor asks students to manipulate assessment conditions to observe how these conditions impact the reliability of the gain scores. In observing how test characteristics can affect gain score reliability, students improve their understanding of the relationship between instrument design and score precision (Subdomains B and C). More advanced activities can require students to contrast these classical analyses with those using Item Response Theory.

### **Afterword: A Concluding Reflection from Task Force Chair Andrew Ho**

This consensus report represents over two years of discussion and debate among task force members, informed by formal and informal comments from NCME membership. The resulting framework and career and curricular models nonetheless feel to me like a beginning rather than an end. I expect our task force will continue fulfilling our charge by publishing and soliciting commentaries in the months and years to come. I know we also wholeheartedly support continued efforts toward broader consensus about foundational competencies among NCME membership, a consensus that must be built through thoughtful and sustained discussion and debate, in person and in our publications of record.

The cognitive bias known as “the curse of knowledge” refers to experts’ tendency to overestimate the knowledge of novices and misperceive how novices understand the world. This suggests that this framework, shaped as it is by task force members ranging from “established” to “senior,” may be less useful for those who may have limited relevant knowledge and experience with the terms, justifications, and examples in this report. I hope it will nonetheless be relevant to NCME members responsible for designing curricula and shaping work environments in ways that enable learners to acquire these competencies. And to the extent that learners see this framework early in their careers, one measure of the success of the framework is that they will increasingly appreciate the domains, subdomains, and intersections as their expertise develops.

We release this report in a period of ongoing turmoil in education and in educational measurement. Critics and proponents of measurement in education are debating with and past each other, sometimes presuming different purposes and goals, and rarely invoking the competencies we present in this report. Historical and empirical research is improving our understanding of the scope of past and present misuses of test scores in psychology, statistics, and education, as well as the ways in which such misuses can reverberate in educational practice and public consciousness. Evolving and diversifying digital and physical learning environments increasingly complicate the question of how we can generalize an educational measurement inference from one context to another. And artificial intelligence and natural language models seem poised to transform instrumentation and perhaps even suggest new constructs or measurement conditions.

Set within this context, I hope our report offers a foundation for common ground, both within the NCME community and, perhaps through this community, for others beyond NCME. We should aspire to be communicators and collaborators. We should aspire to technical, statistical, and computational fluency. We should understand and account for context. We should aspire to gather validity evidence and to use measurements fairly. We should design instruments creatively, guided by theory. We should be precise about our limits of generalization. And we should use psychometric models deftly to achieve all the above aims.

I know our task force members are committed to continued effort toward consensus about foundational competencies in educational measurement. Discussion and debate over foundational competencies can improve coherence in the field and the NCME professional community. As our report stated in our 6 principles at its outset, we expect foundational competencies to shift as society and science advance. We welcome continued engagement and periodic revisitation and advancement of this document by others.

Respectfully submitted,



Andrew Ho

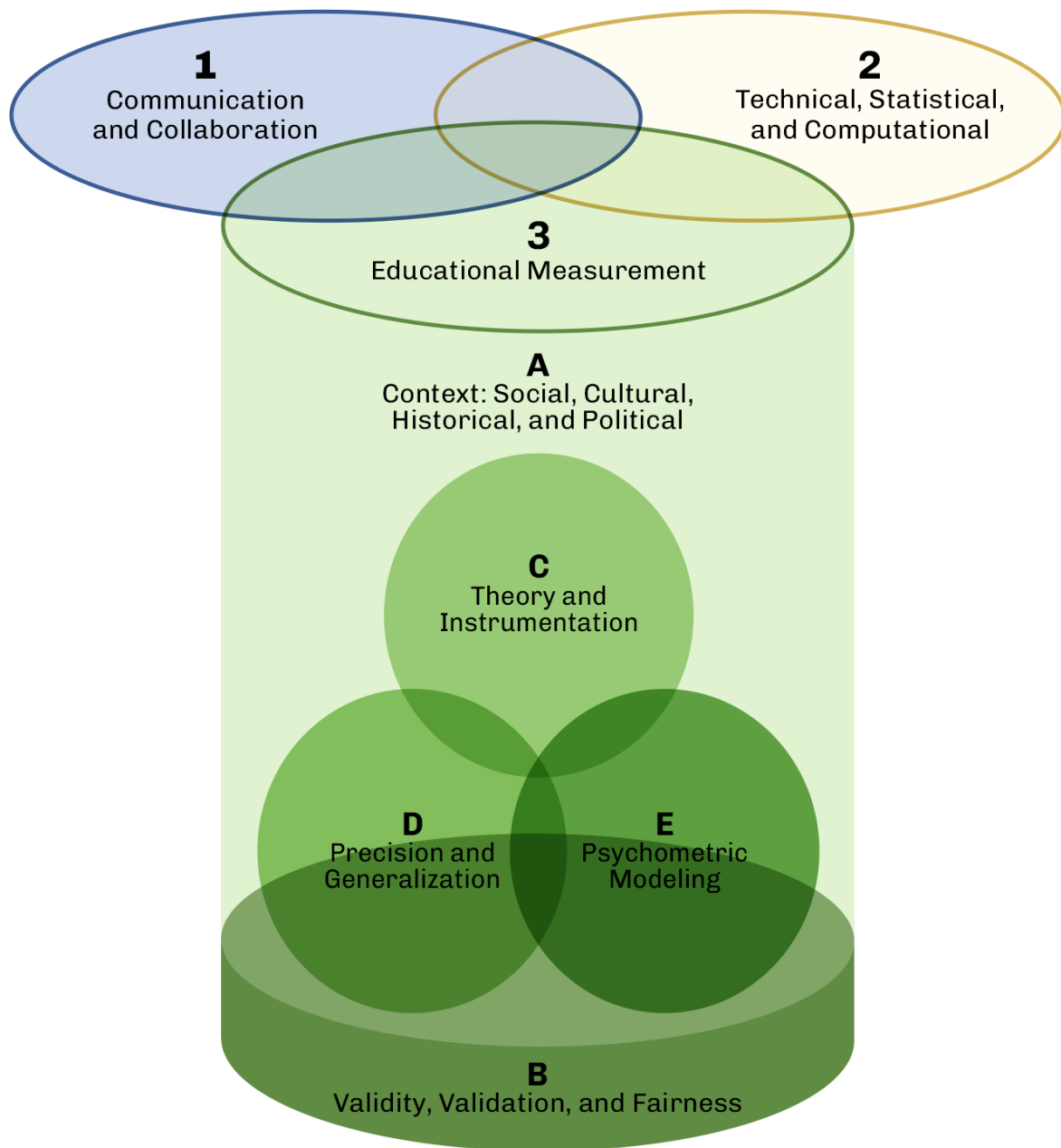
Chair, Task Force on Foundational Competencies in Educational Measurement

Members of the 2021-2023 Task Force on Foundational Competencies in Educational Measurement:

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*

**Appendix A.** Alternative representation of task force consensus domains (1-3) and subdomains (3A-3E) in educational measurement.

### Foundational Competencies in Educational Measurement



*Note:* This figure illustrates the task force conception of foundational competency domains in educational measurement (1: Communication and Collaboration Competencies, 2: Technical, Statistical, and Computational Competencies, and 3: Educational Measurement Competencies) and educational measurement subdomains (3A: Social, Cultural, Historical, and Political Context, 3B: Validity, Validation, and Fairness, 3C: Theory and Instrumentation, 3D: Precision and Generalization, and 3E: Psychometric Modeling). The figure emphasizes how domains and subdomains intersect and interact. The figure also captures how subdomain 3A is *overarches* other measurement competencies and subdomain 3B *undergirds* other measurement competencies.