FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT

A Presidential Task Force Draft Report

Released for NCME member comment on September 16, 2022
Comments requested by October 31, 2022, to ncme@talley.com
FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT

Background, Membership, and Charge

Process

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

TASK FORCE PROVISIONAL DEFINITIONS OF TERMS

PART 1: A FRAMEWORK FOR FOUNDATIONAL COMPETENCIES

Domain 1: Communication & Collaboration

Domain 2: Technical, Statistical, & Computational Competencies

Domain 3: Educational Measurement Competencies
  Subdomain A: Social, Cultural, Historical, and Political Context
  Subdomain B: Validity, Validation, and Fairness
  Subdomain C: Instrument Design and Development
  Subdomain D: Precision and Generalization
  Subdomain E: Psychometric Modeling

PART 2: FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT CAREERS

What defines a career in educational measurement?

How do foundational competencies manifest in educational measurement careers?

How do competencies continue to develop throughout educational measurement careers?

How can different educational measurement careers require foundational competencies?
  Example 1: Common Duties of Psychometricians in K-12 Assessment Companies
  Example 2: Common Duties of Psychometricians/Researchers in K-12 Educational Technology Companies
  Example 3: Common Duties of Psychometricians in Licensure and Certification Organizations
  Example 4: Common Duties of Research Associates in Non-Profit Educational Research and Consulting Companies
  Example 5: Common Duties of University Faculty in Educational Measurement

How can specific educational measurement career scenarios require foundational competencies?
  Example 1: Mentoring Graduate Students to Become Educational Measurement Professionals
  Example 2: Senior Faculty Contributions to Measurement
  Example 3: Helping to Solve an Operational Test Design Issue
  Example 4: Helping to Solve an Operational Item Scoring Issue

PART 3: FOUNDATIONAL COMPETENCIES IN EDUCATIONAL MEASUREMENT COURSES, PROGRAMS, AND ACTIVITIES
Where in a curriculum can programs develop students’ foundational competencies?

What are examples of course topic sequences in a first-year educational measurement sequence?
- Course 1: An Introduction to Educational Measurement or Measurement in Survey Research
- Course 2: Item Response Theory and its Applications

How can curricular activities develop foundational competencies?
- Curricular Activity 1: The Meaning of Measurement
- Curricular Activity 2: Sampling Foundations of Classical Test Theory
- Curricular Activity 3: Contrasting Classical Test Theory and Item Response Theory
- Curricular Activity 4: Instrument Design and Development
- Curricular Activity 5: Psychometric Properties of Gain Scores

TASK FORCE DRAFT REPORT CONCLUSION AND NEXT STEPS
Foundational Competencies in Educational Measurement

Background, Membership, and Charge
On October 20, 2021, NCME President Derek Briggs convened the 12 members of the NCME Task Force on Foundational Competencies in Educational Measurement: Terry Ackerman, Deborah Bandalos, Derek Briggs (ex officio), Howard Everson, Andrew Ho (chair), Sue Lottridge, Matthew Madison, Michael Rodriguez, Michael Russell, Sandip Sinharay, Alina von Davier, and Stefanie Wind.

The charge to the Task Force is threefold:
1. To develop and maintain 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 consensus framework for foundational competencies, discuss curricular models, and consider how different career paths require foundational competencies. This September 2022 report summarizes Task Force consensus to date.

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 five 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. Foundational competencies are not a discrete list; they are defined 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 develop in formal schooling, training, and other learning environments.

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 supporting inferences about an attribute or property of an object, event, or phenomenon.\(^1\)

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 respondents to demonstrate their capabilities with respect to the construct of measurement.

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.

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 full-time in an educational measurement career.

\(^1\) 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 Irrribara, 2021).
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.
- 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: Instrument Design and Development
- Subdomain D: Precision and Generalization
- Subdomain E: Psychometric Modeling

Figure 1 illustrates these unordered and overlapping competency domains and subdomains.
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 in education, psychology, computational sciences, and technology development. Collaboration skills involve more than the ability to get along with others. Collaboration entails 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, 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, and generate reports. 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 statistical methods and 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 more powerful algorithms, computational competencies are becoming more important for educational measurement professionals to develop. Because many educational tests are administered digitally, educational measurement experts may work with engineers to develop software for assessment environments to 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 competences 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) instrument design and development, D) precision and generalization, and E) psychometric modeling. Together, competencies in these subdomains support common educational measurement efforts, including designing and developing measurement instruments, scoring responses, estimating and reporting score precision, 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 purposes.

Subdomain A: Social, Cultural, Historical, and Political Context

Description: Context competencies include the ability to identify social, cultural, historical, and political factors that intersect with the measurement process and may affect respondents’ interactions with measurement instruments, the interpretation of responses and scores, and the appropriate 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.

- **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 examples include social structures based on race, gender, class, culture, and ableness.

- **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 the measurement field, including the misuse of measurement science to justify racist policies and practices, is essential context when designing, enacting, and reporting the results of an educational measurement procedure.
• **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.

**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 items and scoring procedures that adhere to bias, sensitivity, and accessibility guidelines to maximize test-taker engagement and minimize potential construct-irrelevant bias. Competent learners also understand the importance of test administration conditions that appropriately engage respondents with an instrument. They can understand the importance of political context by, for example, anticipating how test-based accountability policies for teachers and schools may inflate aggregate trends over time.

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

**Description:** This competency relates to the 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 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 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 why content alignment is important for a test for educational accountability, why the correlation between test scores and college grades is important evidence to support the use of an admissions test, and why 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.
Subdomain C: Instrument Design and Development

Description: Instrument design and development requires understanding the processes and procedures employed throughout the life cycle of an instrument to ensure that the instrument is sensitive to variation in levels of the intended construct of measurement. To do this successfully requires understanding theories about learning within subject area content domains, and/or understanding the importance of collaborating with those who possess this expertise. Learners with this competency have experience developing test specifications and blueprints; defining performance level descriptions; and authoring items, tasks, and scoring guidelines that adhere to content specifications and to bias, sensitivity, and accessibility guidelines. Learners may also have experience assembling items and tasks into test forms; designing and selecting delivery interfaces that are usable and accessible; developing item security and item replacement protocols and procedures; and authoring manuals for administration and technical documentation.

Justification: Measurement instruments are a primary tool through which evidence about a targeted construct is collected. 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 sound educational measurement.

Examples: A competent learner understands the importance of working with content experts to develop a definition of a targeted construct, and to develop a test blueprint that adequately represents the construct. A competent learner has had some experience authoring items and tasks that adhere to test specifications; content guidelines; and bias, sensitivity, and accessibility guidelines. A competent learner can develop scoring rubrics that focus rater attention on core elements of the targeted construct.

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, and occasions, 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 scaling procedures interact with precision and generalization, including how the choice of scale score units can affect score precision.

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 as tools for investigating hypotheses about the relationship 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 models make; and understand whether and how the assumptions can be evaluated. Competent learners should also understand the purpose and context for generalizability theory, structural equation modeling, latent class analysis, and hierarchical linear modeling.

**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.
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.

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. starting a career in industry or government, for example, would not be expected to have developed 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 the foundational competencies and develop other job-related competencies.

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. These foundational competencies should themselves 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 example below represent duties in job descriptions across K-12 assessment (Figure 1), educational technology (Figure 2), and licensure and certification (Figure 3), as well as research careers in non-profit organizations (Figure 4) and universities (Figure 5). The common duties align with foundational competency domains to illustrate the distribution of the responsibilities across domains.

The domains and subdomains described earlier are well-represented across all of these job duties, and the job descriptions themselves illustrate the inter-relatedness and overlapping nature of the work. It can be difficult to map some duties cleanly to a single domain. Importantly, communication and collaboration (Domain 1) comprise a large portion of the job duties, with a strong emphasis on communicating effectively, both internally and externally, to a variety of audiences with an emphasis on the ability to communicate research findings. Not surprisingly, educational measurement (Domain 3) comprises a large portion of job duties for the psychometrician jobs, and often the statistical and technical domain (Domain 2) appears in job descriptions in ways that can be difficult to disentangle from measurement competencies.
Example 1: Common Duties of Psychometricians in K-12 Assessment Companies

Domain 1: Communication and Collaboration Competencies
- Produce technical documentation related to item, test, and program performance
- Present and publish theoretical and/or application papers in conferences and journals
- Interact with clients for job analysis, standard setting, measurement explanations and other appropriate topics
- Work with members of testing services teams, account managers, technology teams, etc.
- Participate in internal training and individual development of technical skills.
- Develop research proposals.

Domain 2: Technical, Statistical, and Computational Competencies
- Manipulate and validate data files.
- Design and implement research projects.

Domain 3: Educational Measurement Competencies
- Identify and/or develop the way each test is scored.
- Analyze test and item-level results using appropriate methods.
- Equating, linking & scaling of tests.
- Work with content development to ensure that item pools meet the needs of the program.
- Ensure accuracy of tests and test analyses.
- Develop a plan that ensures that examination services comply with industry standards for security, validity, reliability, fairness, and defensibility.
- Develop an appropriate test design including mode of administration, test specifications, and pretest strategy.

Example 1: Common Duties of Psychometricians/Researchers in K-12 Educational Technology Companies

Domain 1: Communication and Collaboration Competencies
- Collaborate with internal teams to understand their goals and jointly scope psychometric/research/data support to meet their needs
- Organize and collaborate on data for visualizations
- Collaborate with Product Research team to understand data sources/structures/meaning

Domain 2: Technical, Statistical, and Computational Competencies
- Apply sufficient knowledge of statistical procedures, psychometric methods, and statistical programming to work independently on most aspects of statistical analysis work, while receiving instruction and guidance in other aspects
- Ensure quality control of deliverables
- Be accountable for basic statistical design and analysis decisions
- Evidence excellent computer programming skills
- 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 expertise to assessment content development teams
- Create visualization of assessment metrics and concepts in educator dashboards
- Review and revise technical documentation
**Example 2: Common Duties of Psychometricians in Licensure and Certification Organizations**

Domain 1: Communication and Collaboration Competencies
- Provide input, as needed, to the development of short and long-term schedule of psychometric activities.
- Work with the IT and Quality Assurance departments to explain psychometric principles to aid in developing software and conducting quality assurance processes.
- Represent the organization on assessment and psychometric and research issues with internal and external constituents.
- Conduct presentations at psychometric and exam related conferences
- Facilitate standard setting, key validation, and other exam related meetings
- Serve as an active member on cross departmental teams as needed.

Domain 2: Technical, Statistical, and Computational Competencies
- Assist in the accreditation process by supplying data as required.

Domain 3: Educational Measurement Competencies
- Conduct routine psychometric analyses on all examination programs including IRT calibrations, scaling and equating projects, item analysis and key validation, standard settings, item bank analyses, test form construction
- Evaluate the analyses and documentation completed by internal psychometric staff.
- Conduct psychometric quality assurance processes on exam related data queries and reports.
- Assist senior psychometric staff in efforts to improve exam products through innovative assessment methodology
- Keep abreast on latest trends and relevant research in certification and licensure testing.

**Example 3: Common Duties of Research Associates in Non-Profit Educational Research and Consulting Companies**

Domain 1: Communication and Collaboration Competencies
- Participate in a team that values diverse voices and ideas, prioritizes co-designing practices and deliverables, and emphasizes team building, learning, and professional growth.
- Assist in the writing of reports intended to impact the decisions of educational policymakers and practitioners
- Contribute to the creation of proposals and outreach efforts that advance agency goals
- Help organize and design presentations, trainings, and other dissemination materials or activities
- Write and format summary tables, graphs, presentations, and reports.

Domain 2: Technical, Statistical, and Computational Competencies
- Carry out both simple and complex data analyses using statistical software
- Clean complex data sets and conduct simple descriptive analyses
- Design data collection protocols
- Collect, code, and analyze data

Domain 3: Educational Measurement Competencies
- Conducting research reviews, including literature reviews and document reviews
- Synthesize information for preliminary and final deliverables
Example 4: Common Duties of University Faculty in Educational Measurement

Domain 1: Communication and Collaboration Competencies
- Teaching and mentorship
- Writing and presenting research findings at conferences and colloquia
- Collaborating on research projects and coauthoring articles
- Advising external testing organizations or district, state, and federal educational agencies
- Service and leadership for academic and professional organizations

Domain 2: Technical, Statistical, and Computational Competencies
- Applying or improving statistical methods for research and practice
- Programming to enable others to replicate research or apply new methods

Domain 3: Educational Measurement Competencies
- Developing or evaluating educational measurement instruments
- Developing new measurement models that advance scholarship and practice
- Demonstrating new applications of existing measurement models for problems of practice

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 next 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, on the kinds of issues measurement professionals encounter as part of their work. They begin 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. To do so, 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 measurement journal articles into their 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 ITMSCS, 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 examples are hypothetical scenarios that an applicant to testing organization might be given as part of a job interview.

**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,
• 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 psychometricians’ opinions about the consequences and reasonableness of the move. What analyses should the psychometricians perform and how should they respond?

In this scenario, the applicant would be expected to:

• Discuss with subject matter experts whether the change in content raises questions about the validity of score interpretations (Domain 1, Domain 3B & C)
• 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)

**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 a type of item across grades. What analyses should the psychometricians perform and how should they respond?

In this scenario, the applicant would be expected to:
• Discuss the need to examine human rater performance on the validity papers used to monitor raters (Domain 3B)
• Discuss the issue with the rater group to identify any potential sources or patterns of differences, particularly related to the item type with low agreement (Domain 1, Domain 3C)
• Discuss the need to examine the characteristics samples used to train and validate the scoring models versus those observed in live testing to determine whether and how they differ (Domain 2, Domain 3D)
• Absent detectable rater issues, determine whether retraining the engine may solve the issue, using different predictive features or an expanded training and validation dataset (Domain 2, 3E)
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 presentation 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 partner 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. Instrument Design and Development:** 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 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 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 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 course topic sequences in a first-year educational measurement sequence?**

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 instrument design and development, 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**

<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
</tr>
</thead>
</table>
| 1    | ● Introduction to Tests and Survey Instruments  
    | ● Social, cultural, historical, and political context of test and survey instruments  
    | ● Validity and Reliability |
| 2    | ● Big Picture Issues  
    | o Considering the Context: Why administer a test or survey?  
    | o What are Constructs? What is Measurement?  
    | ● Validity, Validation and Fairness (Part 1)  
    | o Consensus Definitions in The Standards  
    | o Historical Overview  
    | o Basic structure of validity arguments |
| 3    | ● Sampling  
    | o Defining the Population of Interest  
    | o Developing a Sampling Frame  
    | o Probability Samples, Pilot Samples, & Convenience Samples  
    | o Sampling Weights  
    | ● Review of chance error, sampling distributions, standard error of a mean  
    | ● Nonresponse Bias |
| 4-7  | ● Instrument development:  
    | o The role of theory  
    | o Designing items for cognitive constructs  
    | o Designing items for affective constructs  
    | o Fairness, diversity, and equity in item design  
<pre><code>| o Pilot testing, Cognitive Interviews, and Item Review Panels |
</code></pre>
<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction and overview of psychometric modeling. Establish relationship between modeling, instrument design and development, and precision and generalization.</td>
</tr>
<tr>
<td>2</td>
<td>Review of Classical test theory (CTT)</td>
</tr>
<tr>
<td>3</td>
<td>Item Response Theory (IRT): Models for dichotomous data</td>
</tr>
<tr>
<td>4-5</td>
<td>Item Response Theory: Models for polytomous items</td>
</tr>
<tr>
<td>6</td>
<td>Item Response Theory: Evaluation of model fit, item fit, and person fit</td>
</tr>
<tr>
<td>7</td>
<td>Item Response Theory: Parameter Estimation</td>
</tr>
<tr>
<td>8</td>
<td>Classical and IRT Methods of Assessing Differential Item Functioning (DIF)</td>
</tr>
<tr>
<td>9-10</td>
<td>Equating Using CTT and IRT methods</td>
</tr>
<tr>
<td>11</td>
<td>Computer Adaptive Testing and Automated Scoring</td>
</tr>
<tr>
<td>12</td>
<td>Vertical Scaling and Growth</td>
</tr>
<tr>
<td>13</td>
<td>Multidimensional IRT Models</td>
</tr>
</tbody>
</table>

Course 2: Item Response Theory and its 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: Instrument Design and Development**

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.

**Task Force Draft Report Conclusion and Next Steps**

With the release of this draft report, the Task Force has made initial progress towards its charge:

1. To develop and maintain 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.

The Task Force will engage with the NCME community through the comment period in the fall of 2022 and respond with a final report in the winter. There is the possibility of further engagement with NCME membership at the NCME Annual Meeting in Chicago in April of 2023.

The Task Force supports continued effort toward consensus about foundational competencies in educational measurement. It also believes that discussion and debate over foundational competencies can improve coherence in the field of educational measurement and the professional community of NCME. It must also be acknowledged that foundational competencies may also shift as society and science advance. As such, it will be important to revisit and update this document periodically.