DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

# Foundational Competencies in Educational Measurement

## Citation

Ackerman, T.A., Bandalos, D.L., Briggs, D.C., Everson, H.T., Ho, A.D., Lottridge, S.M., Madison, M.J., Sinharay, S., Rodriguez, M.C., Russell, M., von Davier, A.A. and Wind, S.A. (2024), Foundational Competencies in Educational Measurement. Educational Measurement: Issues and Practice. https://doi.org/10.1111/emip.12581

## Published version

https://doi.org/10.1111/emip.12581

## Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37379196

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

## Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

## Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

# Foundational Competencies in Educational Measurement

Terry A. Ackerman, *University of Iowa*
Deborah L. Bandalos, *James Madison University*
Derek C. Briggs, *University of Colorado Boulder*
Howard T. Everson, *SRI International and CUNY*
Andrew D. Ho,[1] *Harvard Graduate School of Education*
Susan M. Lottridge, *Cambium Assessment, Inc.*
Matthew J. Madison, *University of Georgia*
Sandip Sinharay, *Educational Testing Service*
Michael C. Rodriguez, *University of Minnesota*
Michael Russell, *Boston College*
Alina A. von Davier, *Duolingo*
Stefanie A. Wind, *University of Alabama*

## Abstract

This article presents the consensus of an NCME Presidential Task Force on Foundational Competencies in Educational Measurement. Foundational competencies are those that support future development of additional professional and disciplinary competencies. The authors develop a framework for foundational competencies in educational measurement, illustrate how educational measurement programs can help learners develop expertise with these competencies, and demonstrate how foundational competencies continue to develop in educational measurement professions. The framework introduces three foundational competency domains: Communication and Collaboration Competencies; Technical, Statistical, and Computational Competencies; and Educational Measurement Competencies. Within the Educational Measurement Competency domain, the authors identify five subdomains: Social, Cultural, Historical, and Political Context; Validity, Validation, and Fairness; Theory and Instrumentation; Precision and Generalization; and Psychometric Modeling.

[1] Corresponding author and Task Force Chair.

What are foundational competencies in educational measurement, and can consensus about these competencies improve the training of educational measurement professionals? In his presidential term for the National Council on Measurement in Education (NCME) in 2021–22, Derek Briggs was motivated by these questions to create an NCME Presidential Task Force to

(1) develop a set of foundational competencies for the field of educational measurement,

(2) illustrate one or more curricular models for a graduate program in educational measurement, and

(3) engage NCME membership and the field with Task Force findings through conference presentations and published journal articles.

President Briggs solicited nominations for Task Force membership from the NCME Board of Directors and other interest groups and convened 12 members, listed here as authors, including academic researchers and professional psychometricians from a range of institutions and organizations. The Task Force released a draft report and solicited commentary from the NCME community. Some NCME members provided comments in open webinars, and 17 NCME members provided comments via email. Through its discussions, and informed by NCME member feedback, the Task Force came to consensus about the nature of its charge under six broad principles:

Foundational competencies in the field of educational measurement:

1. 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. 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. overlap and interact with each other. They are not a discrete list; many are characterized by intersections and interactions.
4. are both descriptive of the profession and discipline and aspirational about the future of the profession and discipline.
5. 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. 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.

The Task Force also developed provisional definitions of terms shown in Figure 1. These definitions helped to focus Task Force discussion and resulted in the consensus framework for Foundational Competencies shown in Figure 2. The consensus framework includes 3 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, as communication, collaboration, technical, statistical, and computational competencies are valued and relevant in a range of academic disciplines and professional fields (Deming, 2017; Levy & Murnane, 2004). However, developing a foundation of understanding for the competencies in these two domains that intersect with the domain of educational measurement requires special training and

experience. The third domain includes five competency subdomains that are more specific to educational measurement. The first two subdomains are:

- Subdomain A is an *overarching* competency subdomain related to the context—including 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 theoretical concepts and technical skills within educational measurement:

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

Figure 2 illustrates these unordered and overlapping competency domains (1–3) and subdomains (3A–3E). Appendix A includes an alternative visualization that emphasizes overlap and intersections.

Insert Figures 1 and 2 here

## Foundational Competencies in Educational Measurement: Descriptions and Examples

The following subsections provide descriptions, justifications, and examples for each domain and subdomain in turn. Although the exposition focuses on competencies representative of each domain and subdomain separately, we emphasize the six broad principles from our introduction. In particular, learning and applying competencies in educational measurement

rarely happens in any individual domain or subdomain without intersecting with other domains and subdomains.

**Domain 1: Communication & Collaboration**

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. Domain 1 competencies 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. They also include the ability to understand a variety of perspectives, manage priorities of all participants, and meet expectations as a member of a team.

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.

A starting point for developing expertise in communication and collaboration competencies is fostered through experience. As such, learners should be given opportunities to present findings from psychometric analyses or validation efforts to both technical and general audiences, and to collaborate with graphic artists and web designers to produce an interactive report that supports valid interpretations and uses of test score information.

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

Technical, statistical, and computational competencies include 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 skills include the ability to write software code and programs, and the ability to understand the logic and purpose of algorithms.

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 need skills that allow them to design assessment environments where examinee-item interactions are monitored, recorded, and scored automatically.

Educational measurement professionals are often involved in work in which they are expected to develop, adapt, and evaluate statistical models and computational algorithms for educational applications. A starting point for expertise in these activities is successfully engaging

with coursework in statistics, computer science, and cognitive science, and with practical experiences that involve coding. A learner who has done this will be able to further their expertise by consulting the research literature and more experienced professionals when, for example, working on a project involving automatic content generation using computational language models while identifying possible sources of bias. A learner at this starting point will also be able to work on projects that gather and analyze cognitive process data from digital environments, and develop the skills needed to think critically about the design and analysis of experiments that would test whether engagement with certain item types leads to different educational outcomes.

**Domain 3: Educational Measurement Competencies**

Educational measurement competencies include five subdomains: A) an overarching subdomain related to the context—including social, cultural, historical, and political contexts—in which measurement occurs; 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. These efforts include 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: Context: Social, Cultural, Historical, and Political*

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 their expertise with 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 expertise in this subdomain can account for these factors to improve the likelihood of valid interpretations.

- *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 those stratified by race, ethnicity, gender, class, language, and disability, such as schools, classrooms, families, professions, and neighborhoods.

- *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 important 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.

Sociocultural theories of learning emphasize the importance of the context in which educational measurement occurs (Mislevy, 2018; National Academies of Sciences, Engineering, and Medicine, 2018). 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.

Learners with expertise in this overarching subdomain are able to 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. Competent learners understand the importance of developing bias, sensitivity, and accessibility guidelines that are responsive to the social, cultural, and historical contexts of the intended respondents. Competent learners also understand the importance of designing and administering items that maximize construct-relevant engagement and minimize construct-irrelevant bias. A starting point for learning in this subdomain can involve exposure to literature and news reports documenting how experiences with poorly designed tests and score reports can themselves be harmful by reinforcing negative perceptions and stereotypes, among respondents about themselves (e.g., Randall, 2021), or among other score users about respondents and respondent subgroups (e.g., Quinn, 2020; Quinn et al., 2019).

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

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. Validity is the fundamental consideration underlying the interpretation and use of test scores (AERA/APA/NCME, 2014). 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.

A learner with expertise in this competency 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 relevant evidence when using test scores in educational accountability policies, b) correlations between test scores and college grades is relevant evidence when using test scores in college admissions decisions, or c) internal consistency is relevant evidence when using test scores in diagnostic screening decisions. A learner with expertise would also be able to identify additional sources of evidence that 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.

A starting point for the development of expertise on this competency involves exposure to the Validity and Fairness chapters in the *Standards for Educational and Psychological Testing* as well as relevant chapters published in the edited volumes of *Educational Measurement* (e.g., Brennan, 2006) and other peer-reviewed journals. This can be further enhanced by giving novice learners the experience of identifying and using different sources of evidence to construct or evaluate a validity argument, and within this context, identifying and evaluating ways that the fairness of a test can be threatened or enhanced though a priori design and post hoc analysis.

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

Developing a measure of a construct in education requires a theory of the construct and how learners learn this 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. 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 (e.g., those reviewed by Ferrara et al., 2016). 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.

Learners with expertise in 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 expertise have extensive 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. They can develop or collaborate with others to articulate a theory of learning to guide task development toward identifying variation in levels of a construct. They understand the importance of working with content experts to develop a construct definition based on this theory and of creating items and tasks that align with theoretical models of learning and cognition. This may also include the use of theory to guide computational methods for item generation, test scoring, and validation.

A starting point for learners as they develop expertise is an awareness of the importance of applying principled approaches to test design and development. This will generally require some introductory exposure to relevant literature found in, for example, the chapters of

*Educational Measurement* (e.g., Mislevy, 2006), the NCME Book Series (https://www.ncme.org/resources-publications/books/book-series), consensus reports (e.g., National Research Council, 2001), and contemporary scholarly research. Expertise can be fostered by experience developing test specifications and blueprints; defining performance level descriptions; and authoring, generating, or evaluating items, tasks, and scoring rubrics. It can also be fostered by activities in which learners are asked to evaluate and critique existing test specifications, content guidelines, scoring approaches, and procedures for evaluating bias, sensitivity, and accessibility.

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

A learner with expertise in the competency in this subdomain can state the intended extent of generalization of test scores and can estimate and interpret corresponding indices of precision. They 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. Such a learner can also identify the evidentiary limits of generalization and design studies to expand the evidence base for generalization in support of desired uses. They understand how score properties and combination procedures interact with precision and generalization. They also understand how transforming, averaging, or differencing score units can depend upon the properties of the scale (e.g., whether the scale is assumed to have interval or ordinal properties) and can therefore have an effect on score interpretation and estimates of precision.

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.

A starting point for a learner developing expertise in this subdomain is to be able 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. These foundations support learners as they develop the knowledge and skills needed to model and communicate information about measurement error in complex situations (e.g., estimates of precision that are heterogeneous across groups and/or conditional on proficiency levels).

### *Subdomain E: Psychometric Modeling*

Psychometric models play a critical role in instrument design and development. They are also valuable for 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 expertise in psychometric modeling competencies can select, fit, evaluate, and interpret results from multiple well-known statistical and psychometric models. They can 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. Large-scale educational measures use 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. 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.

A starting point for the development of expertise in this subdomain is introductory coursework in psychometric modeling. Upon completion of introductory coursework, a learner should be able to 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 appropriate for its intended purpose.

## Foundational Competencies in Educational Measurement Careers

One test of the coherence of the preceding framework is to evaluate whether professionals use these 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 experts 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, skills, abilities and behaviors. A recent graduate with either a master's or doctoral degree 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 outlined 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 foundational competencies in educational measurement.

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

As part of its work, the Task Force reviewed job descriptions related to educational measurement and psychometrics on listservs and search engines. The Task Force found that job requirements of educational measurement careers vary but have many overlapping characteristics. A focus on communication, data analysis and computing, and measurement are core threads running through most measurement jobs. These jobs can include those in K-12 assessment, educational technology, licensure and certification testing, research careers in non-profit organizations, and positions in universities.

While core measurement activities are often similar across organizations, the Task Force found that the focus of that work is driven by the clients that each organization serves. For example, K-12 assessment work focused heavily on adherence to state and federal requirements and guidelines. Educational technology work focused on supporting product development. Licensure and certification focused more on accreditation and/or industry needs. Facets of work that varied included the size and scope of the measurement work as well as instrument types and times. The scope of work also varied and tended to be larger as the degree of measurement representation and expertise in the organization was smaller.

The Task Force found that foundational competency domains and subdomains are generally well-represented across job descriptions and responsibilities. Job descriptions often illustrate how the domains are related and overlapping in the work, consistent with the overlap in the framework in Figure 2. Communication and collaboration (Domain 1) comprised a large portion of the job duties, where job descriptions emphasized communicating effectively to a variety of internal and external audiences with an emphasis on the ability to communicate research findings. For example, desired competencies like “conduct quantitative analyses and

communicate the results clearly and succinctly to non-technical audiences," were common. Educational measurement (Domain 3) comprised a large portion of job duties for psychometric jobs. The statistical, computational, and technical domain (Domain 2) often appeared in job descriptions in a manner that overlaps with educational measurement competencies (Domain 3). For example, the following desired competency was representative: "demonstrated ability to design and conduct psychometric analyses using statistical analysis and programming tools such as R, SAS, or Python." The specification and integration of competencies is consistent with the identification and overlap of domains in Figure 2.

## Foundational Competencies in Educational Measurement Courses and Programs

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 discussing how a program's curricular and co-curricular structures can develop each of the foundational competencies and illustrating one possible design for a first-semester course in educational measurement that can develop these foundational competencies.

Programs in educational measurement vary in size, focus, and mission. The Task Force intends this discussion to be illustrative, not prescriptive. Educational measurement programs

can develop foundational competencies in a variety of ways. As a result, this discussion focuses more on content than pedagogy.

**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 adaptive testing, multimodal analytics, and Artificial Intelligence including machine learning.

**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?**

Figure 3 provides an example of a course that could serve as a foundation for subsequent courses in the field of educational measurement. A full educational measurement program would include many other courses and co-curricular activities. The course assumes a 13-week semester and basic statistical competency. This introductory 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.

Insert Figure 3 about here

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. Subsequent coursework can focus on depth over breadth of coverage. These courses can focus on a specific technique, model, or theory, such as Item Response Theory,

Generalizability Theory, diagnostic measurement models, or validity theory.

## Discussion

This consensus report represents over two years of discussion and debate among task force members, informed by formal and informal comments from NCME membership. What does it mean to be learning educational measurement? Is there consensus about the foundational competencies that NCME members have and hope to develop? And as technology and policy continue to interact with educational measurement, how might the foundations of our field change?

There is both immense opportunity and ongoing turmoil in education and in educational measurement. Critics and proponents of measurement in education risk debating with and past each other. They may presume different purposes and goals. They may benefit from acquiring and applying the competencies in this article. Meantime, 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, this article offers a foundation for common ground, both within the NCME community and, perhaps through this community, for others beyond NCME. The article encourages educational measurement professionals to

- aspire to be communicators and collaborators

- aspire to technical, statistical, and computational fluency
- understand and account for context
- gather validity evidence and use measurements fairly
- design instruments creatively, guided by theory
- be precise about the limits of generalization
- use psychometric models deftly to achieve all the above aims.

As this article stated in the six principles at its outset, Task Force members expect foundational competencies to shift as society and science advance. Therefore, the Task Force welcomes continued engagement and periodic revisitation and advancement of this framework by others.

## References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (Ed.) (2006). Educational Measurement, 4th edition. Westport , CT : Praeger Publishers

Briggs, D. C. (2021) *Historical and Conceptual Foundations of Measurement in the Human Sciences: Credos and Controversies*. Routledge

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics, 132,* 1593-1640.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications,* 41-74.

Levy, F. & Murnane, R. J. (2004). *The New Division of Labor: How Computers are Creating the Next Job Market*. Princeton and Oxford: Princeton University Press.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd edition*. Cham: Springer.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press Cambridge, England.

Mislevy, R. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational measurement*, 4th edition (pp. 257-305). Santa Barbara: Greenwood Publishing Group.

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. Philadelphia, PA: Routledge.

National Academies of Sciences, Engineering, and Medicine. (2014). *How People Learn II: Learners, Contexts, and Cultures*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24783.

National Research Council. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: The National Academies Press.

Quinn, D. M. (2020). Experimental effects of "achievement gap" news reporting on viewers' racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher, 49*(7), 482–492.

Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). "Achievement Gap" language affects teachers' issue prioritization. *Educational Researcher, 48*(7), 484-487.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82-90.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Torres Irribarra, D. (2021). *A pragmatic perspective of measurement*. Springer Brief in Theoretical Advances in Psychology. https://doi.org/10.1007/978-3-030-74025-2

**Figure 1. 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.

**Competencies** refer to a constellation of knowledge, skills, abilities, and behaviors that are necessary for professionals involved in research and/or practice in the field of educational measurement**.** Professionals vary in the extent of their expertise with specific knowledge, skills, abilities, and behaviors that define these competencies.

**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[2].

**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.

[2] The proper definition and conceptualization of measurement has been a matter of active debate for more than a century. Contrast, for example "the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity" (Michell, 1999) with "the assignment of numerals to objects or events according to rules" (Stevens, 1946) or "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" (Torres Irribara, 2021). For a historical and conceptual exploration of these debates in the human sciences, see Briggs (2021). For an interdisciplinary perspective, see Mari, Wilson, and Maul (2023).

**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.

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.

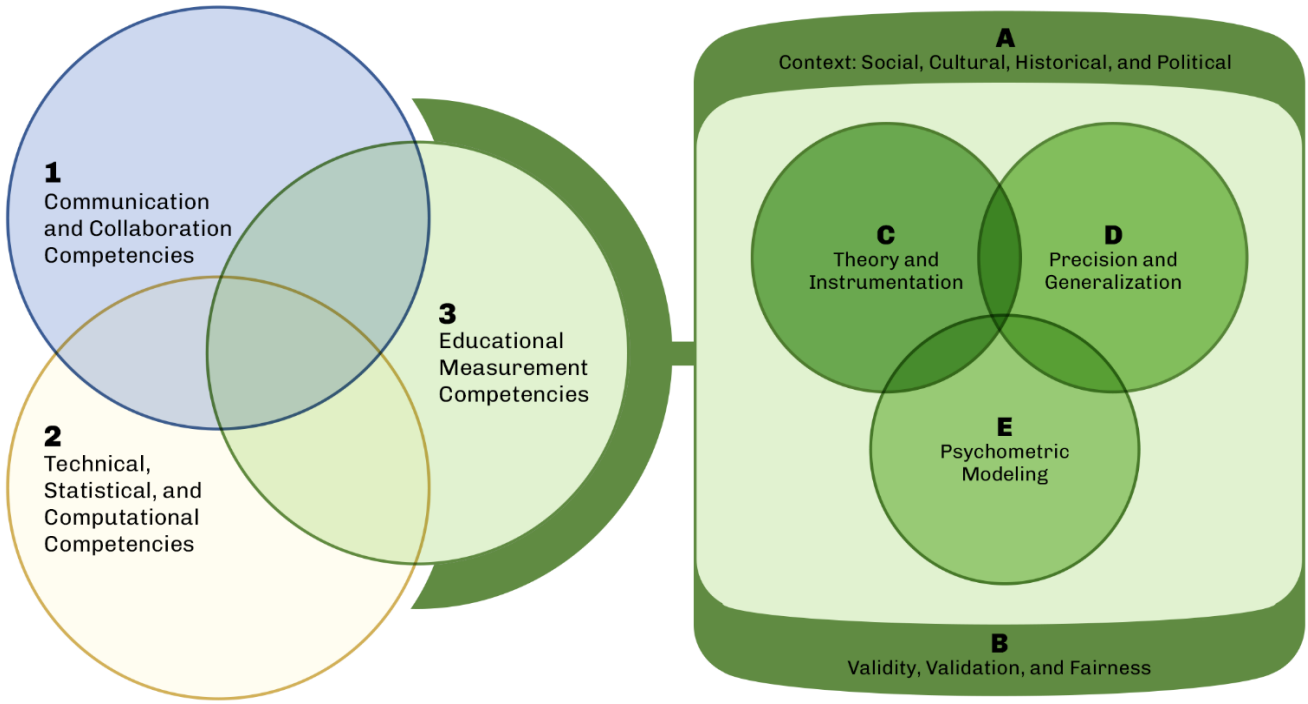



*Figure 2*. A framework for foundational competencies in educational measurement

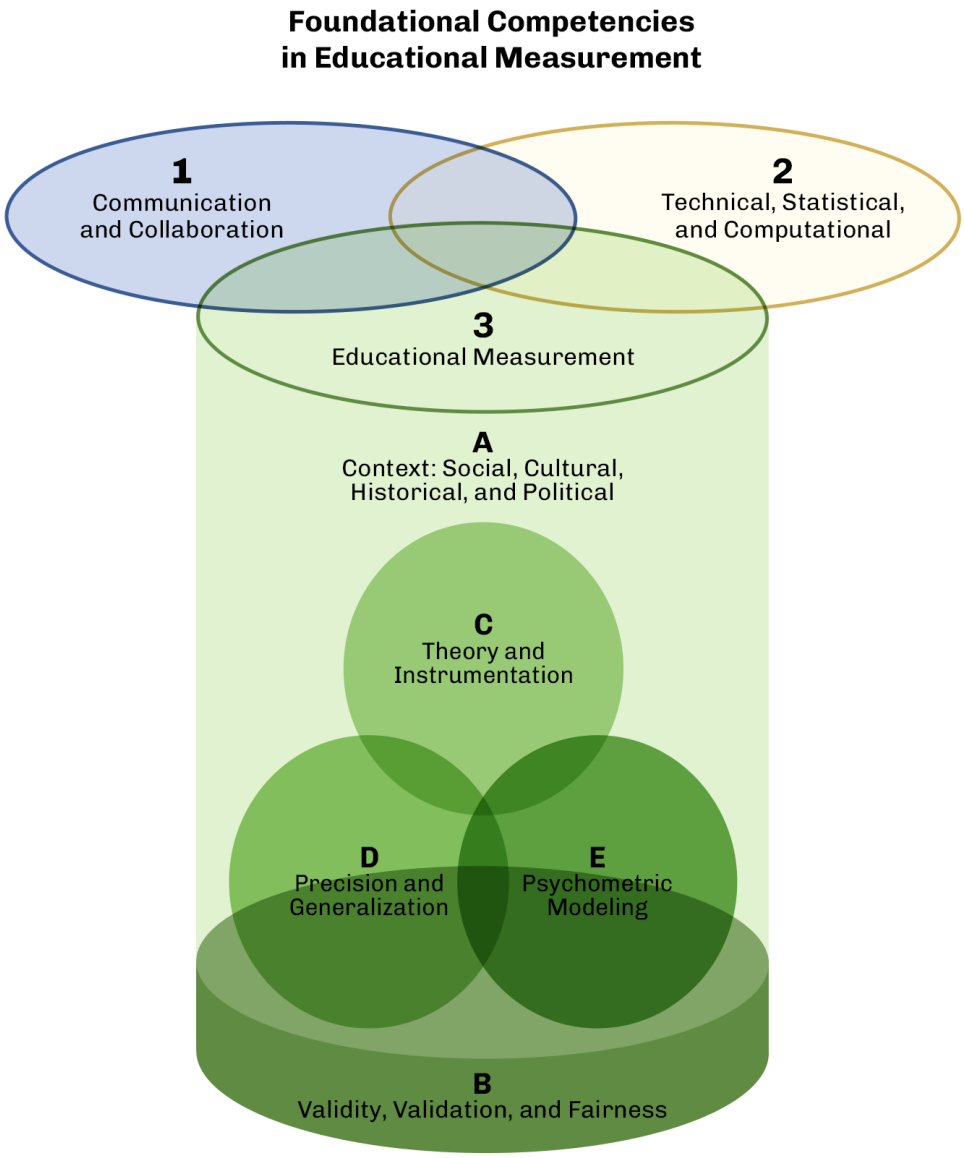
***Figure 3. Illustrating a First Course in Educational Measurement***

| Week | Topics |
|---|---|
| 1 | ● Introduction to Tests and Survey Instruments<br>● Social, cultural, historical, and political context of test and survey instruments<br>● Validity and Reliability |
| 2 | ● Big Picture Issues<br>  o Considering the Context: Why administer a test or survey?<br>  o What are Constructs? What is Measurement?<br>● Validity, Validation and Fairness (Part 1)<br>  o Consensus Definitions in The Standards<br>  o Historical Overview<br>  o Basic structure of validity arguments |
| 3 | ● Sampling<br>  o Defining the Population of Interest<br>  o Developing a Sampling Frame<br>  o Probability Samples, Pilot Samples, & Convenience Samples<br>  o Sampling Weights<br>● Review of chance error, sampling distributions, standard error of a mean<br>● Nonresponse Bias |
| 4-7 | ● Instrument development:<br>  o The role of theory<br>  o Designing items for cognitive constructs<br>  o Designing items for affective constructs<br>  o Fairness, diversity, and equity in item design<br>  o Pilot testing, Cognitive Interviews, and Item Review Panels |
| 8 | ● Item analysis<br>  o Frequency Distributions and Descriptive Statistics<br>  o Item difficulty and discrimination<br>● Introduction to Classical Test Theory (CTT)<br>  o The concept of measurement error<br>  o CTT as a model<br>  o Reliability coefficients |
| 9 | ● Estimating Reliability and Quantifying Measurement Error<br>  ● Internal consistency (vs. stability), Cronbach's Alpha<br>  ● Test-Retest<br>  ● The Standard Error of Measurement<br>  ● The Limits of CTT (G Theory as an expansion) |
| 10 | ● Introduction to Item Response Theory (IRT):<br>  o Foundational principles and conceptual overview of IRT<br>  o Models for dichotomous item responses<br>  o Application with software and data |
| 11 | ● Estimating and Evaluating IRT Models<br>  o Conceptual overview of IRT estimation procedures<br>  o Basics of model-data fit<br>  o Application with software and data |
| 12 | ● Introduction to Exploratory Factor Analysis (EFA) |

| | |
|---|---|
| | ○ Conceptual overview of EFA<br>○ Scree Plots and Parallel Analysis<br>○ Interpreting EFA results |
| 13 | ● Validity, Validation and Fairness (Part 2)<br>○ Different perspectives on validity, the role of consequences<br>○ Examples/illustrations of validation studies in practice<br>○ Evolving views on fairness |

**Note:** Figure 3 presents topics listed by week. Although the topics reference Domain 3 competencies, instructors can design course activities and assignments to promote the development of competencies in Domains 1 and 2.

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