

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 26 Number 2, February 2021

ISSN 1531-7714

Graduate Training in Educational Measurement and Psychometrics: A Curriculum Review of Graduate Programs in the U.S.

Jennifer Randall, *University of Massachusetts*
Joseph A. Rios, *University of Minnesota*
Hyun Joo Jung, *University of Massachusetts*

This mixed-methods study included a curriculum review of 118 graduate (masters & doctoral) programs in educational measurement, assessment, evaluation, psychometrics, and/or quantitative psychology in the United States to examine both the content and skills prioritized in graduate training. In addition to required content, programs/program curricula were coded with respect to intellectual home (psychology v. education departments), level of program (M.A. v. EdD or PhD), total credits, and number and rank of faculty. Patterns with respect to content variation are presented. To supplement these data, interviews were conducted with measurement professionals – working in industry, government, and the academy- to determine what skills and content knowledge they believe to be critical for success in the field and to evaluate any disconnects between content knowledge thought to be important to practitioners and the actual content knowledge taught.

Introduction

The National Council on Measurement in Education (NCME) was founded in 1938,¹ and has grown to include over 2,000 members from countries all over the world making it the largest professional measurement organization in the world. As the professional organization for measurement specialists has grown and expanded, so has the number of graduate programs in educational measurement/psychometrics (see Russell, Ludlow, & O'Dwyer, 2019 for a more detailed history of the growth of measurement programs). With this growth, measurement specialists have suggested the need to discuss/consider the best approach, or curricular roadmap, to graduate training (e.g., Russell et al., 2019).

One question arises, should a common set of competencies exist that can serve as a standardized foundation of graduate training across the United States? The development of common core curricula has been advocated by such individuals as Bill Gates, due to the ability to ensure breadth of training and a common identity within a given field (Long, 2013). Of course, arguments against a common core have been likewise well articulated most notably within [or with respect to] K12 education – with opponents citing unwanted curricula constraints/restraints and the oppression of the intellectual freedom/flexibility needed for innovation and the continuous development advancement of the profession as a whole (Gerwitz, 2012). To be sure, in the history of educational measurement, there have been no

¹ In 1942 the name was changed to the National Association of Teachers of Educational Measurements. In 1961 the name was changed once more to the National Council on Measurement in Education.

recorded efforts to standardize graduate education and training. Although more recently, Russell et al. (2019) have argued for educational measurement programs to develop a new integrated curricular scope and sequence that prepares doctoral students for the next generation of educational measurement. However, in order to do so, first requires that we understand the current curricula taught within the field to address two questions posed by Gilbert (2004): “(1) To what extent does the research training provided in various doctoral programs achieve the stated goals of these programs and (2) To what extent do the goals and content of doctoral programs lead to research training which meets the needs of students, interested parties and the community as a whole in a context of social, cultural, economic and technological change?” In doing so, the priorities of doctoral/measurement programs and, by extension, the measurement field at large are documented.

Examining the existing curriculum

Although there has been a dearth of published research on doctoral curricula in educational measurement and psychometrics specifically, the field of psychology, more broadly has periodically reviewed both the curriculum development processes and curricular choices of its graduate programs with respect to measurement offerings. Merenda (1996) estimated the degree to which 52 programs across 46 universities in the United States were offering training in psychological measurement. This was done by using a 1995 booklet initiated by the Psychometric Society and published by ACT that included the graduate curriculum in psychological measurement. He found that, out of a total of 736 courses, 10.7% ($n=79$) included topics in test theory, test construction, and scaling with considerable variability across programs (from 1 to 7 courses, median = 3). Moreover, 38.5% of programs ($n=20$) offered no courses on these topics at all. Aiken et al. (1990) reviewed the doctoral curriculum of psychology programs with an eye towards evaluating the quality/quantity of students’ training in statistics, measurement, and methodology. Based on their review, the authors concluded that: (a) doctoral students received traditional training in methodology and statistics that focused on laboratory rather than field research; (b) that the measurement curriculum had declined considerably with many students lacking knowledge of the most basic/classic concepts of

psychological measurement; (c) and that students were not receiving adequate training in new techniques and methodologies.

Despite the authors’ explicit concern about the lack of measurement content included in the psychology curriculum in 1990, Aiken et al.’s (2008) review nearly twenty years later revealed a similar concern. Although more programs began to offer some curriculum related measurement, they found that the median number of weeks for measurement requirements was only 4.5; and 54% of programs judged that their graduates could not assess the reliability of their own measures. Childs and Eyde (2002), using program descriptive materials and course syllabi from 84 programs (46% response rate), investigated the extent to which clinical psychology doctoral programs accredited by the APA provide training in psychological assessment. The authors found that requiring students to take courses in personality and intelligence assessment as well as supervised training in the administration and interpretations of assessments was common among programs. However, similar to Aiken et al. (1990, 2008), they found that coverage of psychometric topics varied widely, and were covered as part of another course (not dedicated to psychometrics) in one or two class meetings. In fact, over a third of programs failed to cover validity and reliability.

To date the examination/review of course curriculum in educational measurement programs has been more limited. Over twenty years ago, Guo and Nitko (1996) examined the content/training that students enrolled in graduate educational measurement programs received. The authors examined program and practicum requirements as well as the characteristics of the students enrolled in 54 graduate training programs. Forty-eight of the 54 programs offered a doctoral degree; and the authors found that doctoral programs fit into one of three major orientations: (a) 19 focused specifically on educational measurement; (b) 16 offered a degree in research methodology (a broader field); and (c) 13 offered doctoral degrees in other areas. Doctoral degrees in research methodology programs (84 credits) were found on average to require more credit hours than educational measurement (78 credits) or “other” (67 credits) doctoral programs. Overall, students enrolled in doctoral programs were required to take an average of 15.4 credit hours in statistical methods, 12.4 hours

in educational or psychological measurement, 5.3 hours in research design, 2.9 hours in psychology/educational psychology, 1.6 hours in evaluation, and less than one hour (0.4) in curriculum development. When comparing educational measurement and research methodology-oriented programs, the authors found that students enrolled in methodology-oriented programs typically took more course credit hours in statistical methods courses (24 v. 21), measurement courses (15 v. 12), and research design courses (9 v. 6). Furthermore, across degree types, nearly 81% of doctoral programs required students to engage in some form of supervised research/internship.

More recently Russell et al. (2019) evaluated the curricula content of 17 educational measurement doctoral programs in the U.S. classifying courses into one of nine categories (e.g., core statistics, qualitative research, measurement theory/instrument development). The authors found, as one might expect, that all programs required at least one course, and most required two or more courses in measurement (e.g., one program required seven courses). The authors also reported that several programs required at least one qualitative research methods and/or evaluation course beyond the advance courses in research, statistics, and measurement. Russell et al. noted a pattern, within the measurement theory courses, suggesting that programs have over time simply added courses to their required curricula to address new methods/techniques.

Significance/Purpose of Work

The intended outcomes of a doctoral program are often vaguely articulated, if at all; but one can infer what is valued through an examination of the curricular content of such programs. Yet, to date, there have been only two reviews of the curricula of educational measurement programs in the U.S. (see Guo & Nitko, 1996; Russell et al., 2019); and these reviews have reflected only a proportion of programs that train educational measurement specialists. Indeed, there has been no comprehensive review of all educational measurement/psychometric program curricula. Furthermore, there has been no attempt to evaluate whether the current course content taught in graduate programs reflects established professionals' perceptions of requisite skills and knowledge to be successful in practice. This study addresses these gaps

by providing a comprehensive curricula review of educational measurement/psychometric programs. Our purpose is twofold: (1) to identify what graduate programs are teaching and any systematic commonalities and/or variations in their curricula across department types (e.g., education vs. psychology); and (2) to describe what working professionals (across multiple sectors) believe to be important/critical and/or missing in graduate level curricula.

The findings from this study have the potential to instigate meaningful conversations about (a) the educational measurement training doctoral students receive; (b) the key skills and content necessary to be a successful professional in the field; and (c) the overlap/gaps doctoral programs should consider when making curricula decisions moving forward.

Method

We employed a mixed-method design in which both quantitative and qualitative approaches were used in order to gain a better understanding of the research questions we posed. Specifically, we used quantitative methods to identify the current course curriculum of graduate programs in educational measurement and a qualitative approach to determine if the current curricula content is perceived as important to working measurement professionals across multiple sectors. To that end, we used a convergent parallel design, which allowed us to collect quantitative (content analysis) and qualitative (semi-structured interview) data simultaneously (DeVos, Strydom, Fouche, & Delpont, 2001). Both data were analyzed separately and interpreted concurrently. Content analysis is a non-invasive research approach to determine the presence of specific words, themes, and/or concepts with qualitative data (i.e., doctoral programs of study), so that researchers can quantify and analyze the meanings and relationships among these themes and concepts. Semi-structured interviews, on the other hand, require the full engagement of participants and allows the researcher to address very specific themes and concepts, which in this case, are designed to elucidate data gathered through the content analysis. Below we separately describe the quantitative and qualitative research methods employed.

Quantitative Study: Content Analysis

Sample Identification

Graduate programs in educational measurement were identified using the NCME (2019) and the American Psychological Association Division 5 (2019) lists of current measurement programs. The two lists shared 52% of programs in common; 9% of the programs in our sample were not found on either list, but were identified by the coder during the review process and confirmed by the principal investigator as educational measurement programs. To be included in this study, programs had to be located within the United States; offer at least one graduate degree (Master's, Specialist, EdD, or PhD) in the area of educational measurement, psychometrics, quantitative methods, or related field; and provide students with at least one measurement course within the recommended program of study. Twenty-two programs were excluded from the final coding process as they did not meet these criteria. The final coding sample thus consisted of 118 programs, including multiple programs (e.g.-master's and doctoral) located within the same university, from across 76 unique institutions.

Curriculum Data Coding

The curricula requirements of all programs were obtained in one of two ways: (a) a search of program websites for program requirements and electives; (b) if this information was not made readily available on program websites (n=4), a request about the program's curricula requirements was sent to the program coordinator or contact person. None of the four contacted programs responded to our email requests. Data were coded by program and course characteristics. In regard to the former, the following variables were coded: (a) intellectual home of program (e.g., in psychology department or education department); (b) level of program (masters, doctoral); and (c) number and rank of program faculty. For course characteristics, we encoded information related to: (a) whether the course was required or an elective; and (b) the course type, which we describe further in the following paragraphs. Course categorizations were double-coded by one of the authors and a doctoral student. When coders were unsure of or disagreed about an appropriate classification, a discussion was held until consensus was reached.

Course Categories/Operational Definitions

Twelve categories of course types were encoded (educational measurement, psychological theory, pedagogical theory, statistics, research methods, sociology & anthropology, history, educational policy, program evaluation, professional development, diversity & inclusion, miscellaneous). In the following text we provide a brief description of the operational definition for each course type. The reader may refer to Appendix A for the more detailed description: (1) *Educational Measurement* included course topics that describe procedures on the development and administration of assessments to ensure reliable and valid measures such as courses in test development, Classical Test Theory, and Item Response Theory; (2) *Pedagogical Theory* concerns itself with how topics ought to be taught, and how students learn best including topics such as the philosophy of education and student learning and instructional technology; (3) *Psychological Theory* is the study of human behavior and the function of the mind at different life stages. Course topics in this area deal with motivation, emotion, cognition, and memory; (4) Courses in *Statistics* describe methodologies for evaluating numerical data collected from educational and psychological measures to make descriptive and inferential claims including multivariate statistics, structural equation modeling, and nonparametric statistics; (5) *Research Methods* describe underlying philosophy/theory, strategies, techniques, procedures used in data collection; (6) *Program Evaluation* courses cover topics related to how systems are evaluated, with the end goal of judging or informing a program/policy, or making program decisions; (7) *Sociology and Anthropology* includes classes with a focus on sociology and anthropology; (8) *History* includes coursework covering the past, with a focus on educational issues were coded with this designation; (9) *Educational Policy* includes coursework covering the principles of action adopted or proposed to govern educational systems and practices associated with improving learning, building teachers' capacity, and engaging stakeholders; (10) *Professional Development* courses are dedicated to improving and increasing capabilities of students to become effective researchers and practitioners (includes course credit based on internships, consulting, supervised research, writing, and communication); (11) *Diversity and Inclusion* courses focus on any special population for which equity measures (i.e., extra measures such as tutoring,

mentoring, or separate classes) may be taken including English learners, special education students. Courses that critically analyzed race and gender issues were also included; and (12) *Miscellaneous* included *all* courses that did not meet the definitions of course types listed above such as dissertation credits and directed studies.

Data Analysis

Given our intended purpose, the data were summarized to include both the frequency counts and percentages of courses across course categories and intellectual home (i.e., psychology or education department). In addition, to get a sense of the course-offering patterns within each course category, we examined more closely the programs that accounted for the top one-third of course offerings with each designation².

Qualitative Study: Content Analysis

Sample Selection & Procedures

Purposeful sampling was used to determine interview participants. Specifically, we employed maximal variation sampling in an effort to obtain multiple perspectives. Maximal variation allows the researcher to sample individuals that differ on some characteristic or trait. In the case of our study, we sought to find working professionals who varied with respect to sector (government, academy, industry) and years of experience. We also sought to identify participants who earned their graduate degrees from a variety of doctoral programs and who were diverse with respect to sex (2 male; 7 female) and race/ethnicity (2 Hispanic, 1 Black, 1 Person of Color, 1 Asian, & 4 White). In total, nine participants were interviewed working in industry (n=3), state and federal government (n=2), and the academy (n=4). Potential participants were emailed by the first author and asked to participate in a 30 – 60-minute interview. Participants were provided with a copy of the interview protocol (see Appendix B at the time of the invitation. All contacted participants agreed to speak via Zoom

and each interview was recorded and transcribed to facilitate data analysis. In addition, all participants

completed a brief survey (included in the interview protocol) in which they indicated their perceptions about the importance/criticality of specific measurement-related content for professional success.

Data Analysis

Our interview data analysis process was an iterative and simultaneous process as outlined/described by Creswell (2012) in Figure 1. Transcribed audio files were thoroughly reviewed to establish initial codes, and then similar codes were aggregated together to identify the themes that emerged from participant's responses.

Results

Description of Graduate Programs

In this study, we present the findings from the 69 doctoral programs across 65 unique institutions. On average these programs required 59.5 credit hours for graduation, with programs housed in education (59.3) and psychology (59.1) requiring nearly an identical number of credit hours. The average number of faculty across all institutions (note some institutions have multiple programs) was seven (4.5 tenure holding; 1.7 tenure earning). On average 67.1% of affiliated faculty were tenure holding (associate or full); and 23.3% were tenure earning (assistant professors). Non-tenured faculty (i.e., adjunct and clinical faculty) comprised 5% of the sample; and approximately 4.6% of program faculty held emeritus status.

In total 2,903 courses were reviewed and classified into one of twelve (e.g., statistics, evaluation) categories. Of those courses, 1,768 were included in doctoral programs of study. We found considerable overlap within institutions with respect to shared course offerings across multiple programs. Tables 1 and 2 show the frequency/percentage of doctoral courses and descriptive statistics by course designation, respectively, disaggregated by intellectual home. Within programs housed in psychology departments, the largest percentage of the courses offered in doctoral programs were classified as statistics (37.1%)

² Note. Because so few courses were offered in evaluation (n=72), history (n=18), diversity & inclusion (n=29), educational policy (n=47), educational/pedagogical theory (n=95), and sociology/anthropology (n=18), we examined course patterns across the entire population of courses to get a better sense of course patterns.

or psychological theory (24.4%) followed by courses in educational measurement (12.7%) and professional development (9.5%). Although, courses in statistics (27.1%) also composed the greatest percentage of offerings within education departments, these departments boasted much smaller percentages of courses in psychological theory (8.9%) and professional development (5.6%). Courses in research methods (20.0%), educational measurement (13.6%), and educational theory (6.5%) composed a large

percentage of offerings in education departments when compared to psychology programs.

In the following section, we will present the results juxtaposing our findings from the curriculum audit (quantitative data collection) with those from interviews with working professionals (qualitative data collection). Using the convergent parallel design for data collection and analysis, three primary themes emerged with respect to doctoral training: (1) technical skills; (2) substantive/theoretical knowledge; and (3) professional development.

Figure 1. The Qualitative Process of Data Analysis (From Creswell 2012)

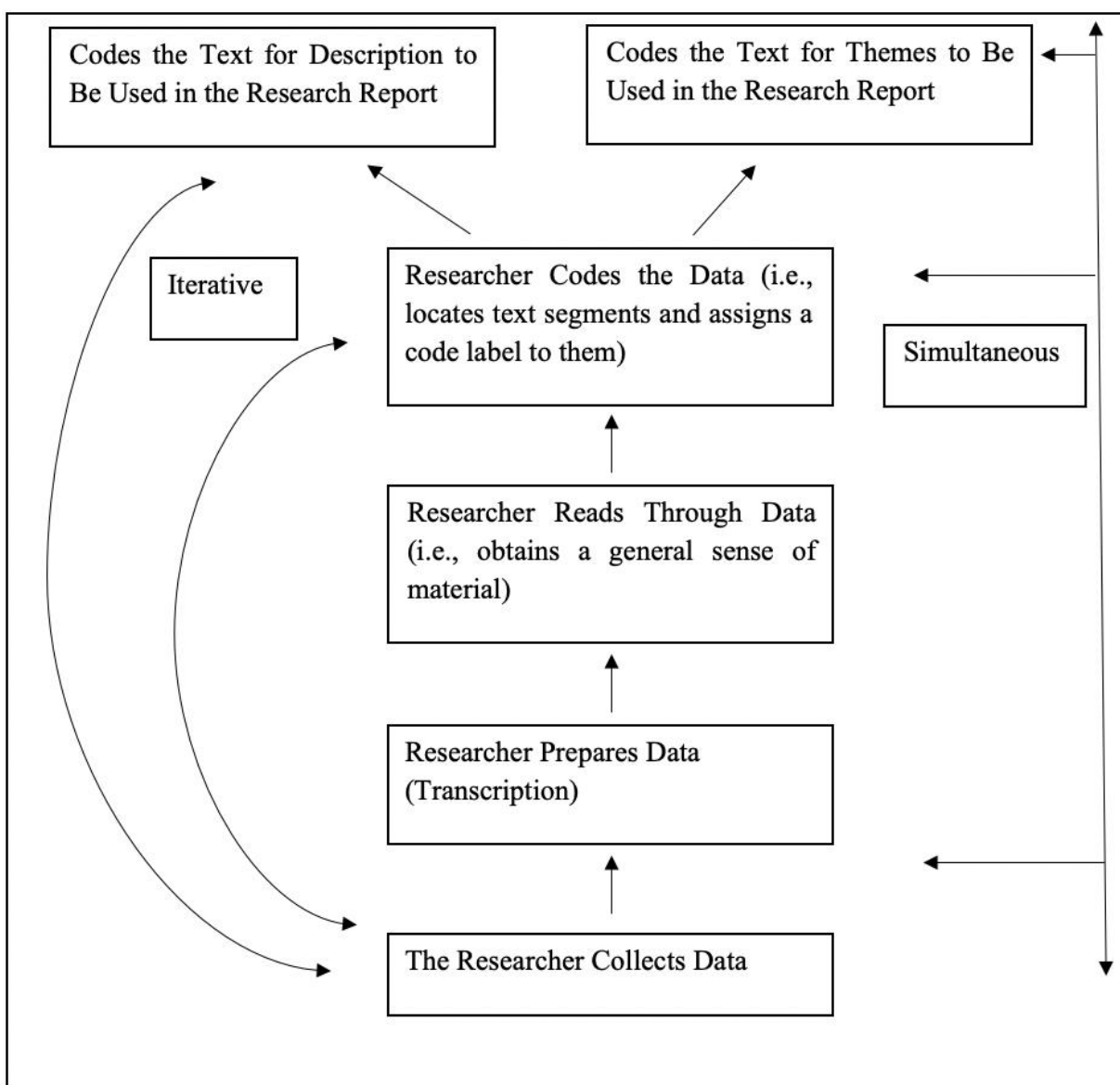


Table 1. Course Frequencies for Doctoral Programs Disaggregated by Intellectual Home (Percentages in Parenthesis)

Row Labels	STAT	RM	EM	PSYCH	ET	EVAL	EP	DI	HIST	SOAN	MISC	PD	Grand Total
EDUC	398 (27.1)	294 (20.0)	200 (13.6)	131 (8.9)	95 (6.5)	72 (4.9)	45 (3.1)	28 (1.9)	17 (1.2)	15 (1.0)	94 (6.4)	82 (5.6)	1471
PSYCH	102 (37.1)	24 (8.7)	35 (12.7)	67 (24.4)	--	--	2 (0.7)	--	1 (0.4)	3 (1.1)	15 (5.5)	26 (9.5)	275
OTHER	13 (59.1)	2 (9.1)	3 (13.6)	1 (4.5)	--	--	--	1 (4.5)	--	--	--	2 (9.1)	22
Grand Total	513 (29.0)	320 (18.1)	238 (13.5)	199 (11.3)	95 (5.4)	72 (4.1)	47 (2.7)	29 (1.6)	18 (1.0)	18 (1.0)	109 (6.2)	110 (6.2)	1768

STAT= Statistics; RM=Research Methods; EM=Educational Measurement; PSYCH=Psychological Theory; ET = Educational Theory; EVAL = Evaluation; EP = Educational Policy; DI=Diversity/Inclusion; HIST = History of Education; SOAN= Sociology/Anthropology; MISC= Miscellaneous; PD = Professional Development; EDUC = Education; PSYCH = Psychology

Note. Values in Parenthesis represent row percentages. For example, 27.1% of the 1471 courses in departments/colleges of education are statistics courses.

Table 2. Number of Courses Offered for Doctoral Programs Disaggregated by Intellectual Home

Dept		STAT	RM	EM	PSYCH	ET	EVAL	EP	DI	HIST	SOAN	MISC	PD	Grand Total
Overall	Mean	7.3	4.6	3.4	2.8	1.4	1.0	0.7	0.4	0.3	0.3	1.6	1.6	25.3
	Median	7.0	3.5	3.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	22.5
	Mode	7.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0
	SD	4.1	3.4	2.4	5.0	2.4	1.1	1.7	1.1	0.8	0.8	1.6	2.0	11.7
Education	Mean	6.9	5.1	3.4	2.3	1.6	1.2	0.8	0.5	0.3	0.3	1.6	1.4	25.4
	Median	6.5	4.0	3.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	22.5
	Mode	7.0	2.0	3.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	16.0
	SD	4.2	3.5	2.5	3.9	2.5	1.1	1.8	1.2	0.9	0.9	1.7	1.8	12.4
Psych	Mean	9.3	2.2	3.2	6.1	0.0	0.0	0.2	0.0	0.1	0.3	1.4	2.4	25.0
	Median	9.0	3.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	24.0
	Mode	10.0	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0
	SD	2.5	1.6	2.4	8.0	0.0	0.0	0.6	0.0	0.3	0.6	1.4	2.5	7.2
OTHER	Mean	13.0	2.0	3.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	22.0
	Median	13.0	2.0	3.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	22.0
	Mode	-	-	-	-	-	-	-	-	-	-	-	-	-
	SD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

STAT= Statistics; RM=Research Methods; EM=Educational Measurement; PSYCH=Psychological Theory; ET = Educational Theory; EVAL = Evaluation; EP = Educational Policy; DI=Diversity/Inclusion; HIST = History of Education; SOAN= Sociology/Anthropology; MISC= Miscellaneous; PD = Professional Development; EDUC = Education; PSYCH = Psychology

Technical Skills

Nearly 61% (n=1071) of all courses in doctoral programs focused on technical skills in educational measurement (mean = 3.4), statistics (mean = 7.3), and/or research methods (mean = 4.6). In an effort to ascertain course patterns within this large domain of technical skills as well as the extent to which actual course patterns reflected the insights of our interviewed participants, we examined the course offerings across the top-third of programs. In doing so, we found considerable overlap and congruence. For example, all interviewees identified course content focused on Classical Test Theory, Item Response Theory, test development, and validity theory as critical for professional readiness; and we found most of these courses to be among the most frequently offered in the top programs. Courses in assessment/test construction and Item Response Theory (n=21) were frequently taught, with nearly all programs offering at least one and many offering multiple courses on these topics specifically. Moreover, nearly every program listed at least one survey course that we described as general psychometrics (n=21). These courses provided an overview of multiple measurement theories and topics including Item Response Theory and Classical Test Theory. Add to that, programs, to a far lesser extent, also provided more dedicated courses in scaling (n=4), validity theory (n=2), equating (n=3), Classical Test Theory (n=4), G-Theory (n=3), and computer-based testing (n=1). With respect to technical skills related to educational measurement specifically, the pattern of course offerings seem to provide students with the content/skill development considered important for success in the field by working professionals. Indeed, even in programs (n=13) that offered only a single course in educational measurement, that solo course, in most cases, surveyed topics in Classical Test Theory and Item Response Theory (n=10).

In reflecting on both their own statistical training and current needs as working professionals, all interviewees indicated that coursework in regression, ANOVA, and computing/programming are critical; and our audit of doctoral program curricula reflected this perceived importance. Among the top programs, multiple courses on these topics were made available to students (regression, n=22; ANOVA, n= 11, & computing/programming, n=14). We also found,

however, that top programs tended to make even more technical courses in statistics available to students including multi-level and/or longitudinal data analysis (n=19), structural equation modeling (n=18), Bayesian methods (n=9), and multivariate statistics (n=15). In fact, on average, doctoral programs offered 6.7 classes in statistics (median & mode = 7.0) which represents more courses than any other domain of courses we examined.

Approximately 18.1% (n=320) of all audited doctoral courses focused on research methods, with an average of 4.6 offered across doctoral programs. All interviews indicated the criticality of a course in quantitative research design; and, to be sure, among top programs these courses were readily available for students including courses in general research methods (n=17) and survey design (n=11). Every program also offered at least two courses (n=30) in general qualitative methodology, which is interesting given that only three of our interviewees identified content in qualitative methodology as “critical.” In fact, some programs offered a wide array of qualitative methodology courses including ethnography (n=2), discourse analysis (n=4), video/visual analysis (n=2), and action research (n=3); none of which were identified by our interviewees as critical or semi-critical. Although being well-versed in these qualitative methodologies may not be considered critical to working professionals in educational measurement, the doctoral programs audited in this study frequently made them available to students. In fact, approximately 26% of research methods/design courses focused solely on qualitative methodologies; and one-fifth of courses included a survey of both quantitative and qualitative methods.

Substantive and/or Theoretical Knowledge

In our interviews with working professionals, the value of having a strong foundation in substantive and/or theoretical knowledge was a primary theme. Indeed, the importance of – as noted by one interviewee – “getting a very well-rounded view of your field and adjacent fields” was echoed in every interview. Multiple participants- particularly those working in the government and industry sectors- discussed the need for doctoral programs to provide students with more than an understanding of measurement and statistical theory. Specifically, they called for students to take courses that provided them

with the larger context in which assessment/measurement happens, so that they could be more critical/thoughtful in their work. As one participant who works in industry noted, “I feel like we are missing more on theory of action. For example, moral consequences of testing, more on the history of...more about the social angle of measurement to testing...I mean some of the decisions that we are using today were made in a completely different context and we just adopted them and said ‘okay, that’s the new way to go.’” Yet, courses that focused on the history of education (mean = 0.3), sociology or anthropology of education (mean = .03), or issues related to diversity and inclusion in education (mean = .4) were rarely identified in doctoral program curricula representing only 3.6% of all audited courses.

Moreover, we found that of all 69 doctoral programs, many failed to include even one course in psychological (n=33) or educational (n=38) theory. Nonetheless, interviewed participants indicated the benefit of having this background knowledge when working in the field. For example, one participant who has worked in both government and industry sectors noted receiving a doctoral education that focused on assessment, but “I have no knowledge about curriculum...So it would have been helpful for me to know a little bit about curriculum and about teaching and educational psychology.” Still, among programs that did offer courses in educational/pedagogical (note that no psychology departments offered courses in this domain) and/or psychological theory, most programs provided at least one course in theories of learning (n=28), while courses in curriculum development/instructional design (n=10) and educational psychology (n=16) were also offered. Within psychological theory, courses in developmental (n=37) and cognitive (n=27) psychology were the most frequently listed in programs of study overall and, as might be expected, were far more likely to be identified in psychology departments than education departments. Courses in social psychology (n=20) and behavioral psychology (n=11) were also frequently made available.

Finally, with respect to substantive knowledge, the value of having an understanding of educational policy was unanimously considered important for professional success. Still, on average, doctoral programs offered fewer than one class (0.6) in

education policy (median & mode = 0). Indeed, we identified only 47 educational policy courses across all doctoral programs. As might be expected only one psychology program made courses in education policy available. Of programs that offered any course in educational policy, most offered at least one general education policy/introduction to policy course. Other institutions listed courses in policy and law, administration/leadership, and/or economics/finance. This lack of focus on educational policy is not surprising and was reflected in one participant’s comments surrounding their own training: “I would have liked to get some more exposure to policy when I was in graduate school. You know educational policy...and those are things I learned on the job, and if I had more I think to understand the various differences too as a graduate student also helps you with your career later on.”

Professional Development

Courses in professional development (e.g., credit-based internships, supervised research, writing, & communication) accounted for 6.2% of all audited courses (n=110); yet the need for more practical experience/focus on practical applications (including training in communication skills) in measurement was the most salient theme in interviews with working professionals. Multiple participants noted that their own measurement training lacked this critical experience. Each participant interviewed, stressed, at some point, the importance of providing students with practical experience (e.g., communicating results to stakeholders, applying technical skills in a meaningful way). Some suggested that this training could be provided via internships/assistantships, and, indeed, internships/practicums were the most common type of professional development (38%) offered by programs. Still, others also noted the importance of having it integrated into the curriculum in such a way that every student (even those who do not participate in internships) would have access to the experience. To that end, we did find several independent studies/supervised research courses (n=15) and courses in writing (n=14) intended to provide students with this critical professional development. Several participants noted that their programs focused primarily on course work (i.e., measurement/assessment in theory), but had fewer opportunities to apply what they learned in operational conditions. One

participant said, when referring to the need for more practical experience as a doctoral student, “I would have liked just more hands on experience.” Still, although all participants emphasized the importance of opportunities to engage in applied/practical work, not all believed their programs failed to provide that experience. In fact, one participant noted: “I think we did have a lot of chances to work on practical problems kind of outside of the classroom, which was helpful.”

Moreover, participants discussed the need to be able to explain test score results to lay people to improve interpretability; and the importance of measurement programs providing that experience within their curricula. One participant noted “Communication of technical information [is critical], because so much of my job now is with communicating with people who are not psychometricians and I think that is really challenging. I know there are courses on educational policy...I think there’s probably many fewer courses on communication of psychometric information. But I think that is a need because there’s a real dearth in our field of people who are both technically competent but also can communicate with people who are not technical, and not that interested in technical information.” Another participant echoed this sentiment “...communication is key. And I think the more that can be emphasized, the better both in giving students opportunities to participate in more real-life experiences through assistantships or through internships and actually being in an environment where they’re seeing what conversation looks like.” During the interview, one participant remarked that a course in which students practiced communicating measurement/assessment data/results to non-psychometricians was a critically helpful and important part of their doctoral experience. Still, in our review of doctoral curricula we identified few courses ($n=3$) that focused on these skills specifically. In fact, 26 programs offered no courses in professional development at all, with psychology departments (mean = 2.4) offering slightly more, on average than education departments (mean =1.4).

Discussion

In this study we provided a comprehensive description of the course content currently offered in

programs that produce educational measurement professionals as well as insights from working professionals with respect to what they believe to be critical for success in the field. Our findings revealed several key points. First, in terms of technical content-based coursework, our quantitative and qualitative findings were well aligned. For example, interviewees noted that courses in regression, ANOVA, and computing (specifically R) were critical; and this belief was reflected in programs’ course offerings. Indeed, across programs offering the most statistics courses, we found 22 that focused on regression, 14 on computing/programming, and 11 on ANOVA. Moreover, when we looked more closely at programs offering merely one or two courses in statistics, we found that these courses primarily focused on ANOVA and regression topics. We found a similar pattern with respect to measurement courses. Interviewees believed that providing students with course content focused on Classical Test Theory, Item Response Theory, test development, and validity theory were critical for success in the field; and program course offerings mostly reflected this need. Among the programs offering the most courses in educational measurement ($n=13$), there were 21 courses offered that focused on Item Response Theory. Although only three programs offered a course that focused solely on Classical Test Theory, nearly every program offered a general psychometrics course that addressed Classical Test Theory. Courses in assessment (which includes test construction/development) were also among the most frequently offered ($n=21$). Although we only found two programs that offered a course explicitly focused on validity theory; most general psychometrics and assessment course descriptions listed validity as a primary topic within the course. These findings are also well-aligned with previous research. In a 1996 study of psychology programs, Merenda found that approximately 10.7% of course offerings focused on [psychological] measurement with a median of three courses being offered in each program. We found similar percentages with 12.7% of courses offered in psychology departments covering content such as test construction, Item Response Theory, or scale development. Of the 1,768 doctoral courses included in this study, 320 (18.1%) were courses in research methods/design – content identified as critical by interviewees. With respect to non-critical technical

course content, our quantitative course review mostly mirrored the opinions of interviewees with courses focused exclusively on multidimensional scaling and non-parametric statistics offered rarely. We did find (as did Russell et al., 2019), however, that nearly all programs offered at least one qualitative methodology course (26% of all methods courses focused solely on qualitative methods) and most offered at least two qualitative methods courses despite only three of our experts indicating that such content was critical to work in the field. Our course audit suggests that programs spend considerable resources (i.e., faculty teaching time) focused on developing students' technical skills; and less time on professional development and/or substantive/theoretical knowledge development.

Second, although programs, regardless of intellectual home (education or psychology department), offered more courses in statistics (7.3) on average than any other content area, the similarities with respect to extensive course availability do not extend beyond this course category. Doctoral programs housed within education departments, on average, offered more courses in research methods (mean = 5.1) than programs housed in psychology departments (mean = 2.2); whereas the availability of courses in psychological theory was far more extensive in psychology (mean = 6.1) than education departments (mean = 2.3). Although all programs offered at least one course in educational measurement (a criteria to be included in the study), programs within education departments offered slightly more of these courses on average (3.4) than programs within psychology departments (2.2).

Third, most programs did not offer a single course in educational policy, that focused on issues of diversity/inclusion/equity, or provided a history of education (mode = 0). Given the fact that measurement does not happen in a vacuum and context is critical, the lack of available coursework to provide students with that context is discomfoting. The U.S. census estimates nearly a third of the population is currently Black (13.4%) or Hispanic (18.3%) and that these subpopulations are expected to grow. With this shift in the U.S. population of students/test-takers as well as increasing calls for instructional, assessment and measurement practices that are more culturally responsive (see Gordon, 1995; Lee, 1998; Aguirre et

al., 2012; Gonzalez et al., 2005; Eglash et al., 2006), doctoral students – more than ever- should be required to take courses that provide them with both a historical and contemporary lens of education especially for historically marginalized groups. Measurement programs must not assume that students will garner this knowledge/understanding elsewhere; but rather should provide or, better, require students to situate their measurement expertise within the larger context of an ever-shifting society. Indeed, interviews with working professionals echoed the importance of students understanding the context in which the work happens. Moreover, according to Randall et al. (2020), approximately 30% of graduates from U.S. doctoral measurement programs from 1997 – 2016 were foreign nationals, suggesting that much of the field's professional supply has limited personal experience with the politics of American education and its history of marginalization.

Finally, only 6.2% of doctoral courses offered focused on supporting students' professional development (PD). Although psychology-based departments offered an average of two professional development courses, most education-based departments offered no courses in PD. This finding is particularly surprising given that every participant interviewed noted the importance of providing students with (a) practical experiences to develop their skills and (b) opportunities to improve their communication skills. Still, 58% of programs did offer at least one course in evaluation, and evaluation work could arguably be considered the integration of technical skills (e.g., statistics, research designs), substantive knowledge (e.g., theory of change/action), and professional development (e.g., communicating results to stakeholders). Courses like program evaluation (n=29) and applied evaluation (n=30) and others could provide students with the opportunity to practice these critical professional development/communication skills.

Limitations

Although this study provides the most comprehensive curricula review of educational measurement/psychometric programs in the United States to date, it is not without limitations. First, the results from this study assume that the curricula information provided via websites accurately reflects the programs of studies with respect to (a) courses

offered; (b) courses required; and (c) course descriptions. Indeed, descriptions of curricula are necessarily snapshots in time, given that programs are expected to change with respect to course offerings. Although it is possible that several program websites are not up-to-date or do not accurately reflect the current curricula options and/or expectations for students, these data (at worse) provide a general overview of curricula patterns in U.S. measurement programs. In addition, in many programs students may be encouraged to take [elective] courses outside of their departments and these courses would not be explicitly listed on the official programs of study. Moreover, our categorization of courses relied solely on the current course descriptions provided on program websites. As the content addressed in courses is expected to change over time – due to new developments in research, changes in instructors, or changes in faculty/student interests and needs – the publicly available course descriptions may or may not be updated to reflect these changes. To that end, some course designations may reflect the content previously taught in the course and not current content. In addition, our data collection method did not account for the hidden curriculum of many measurement programs. For example, content or skills categorized as professional development may not be explicitly listed within the articulated curriculum; students may have these experiences via mandatory work on research teams/projects, internships, and ad-hoc seminars/workshops. Finally, our data included all courses listed in programs of study or in course catalogs; but it is difficult to know how often these courses are actually taught. It is certainly reasonable to expect that some courses are offered so infrequently that students have little opportunity to take them at all. Nonetheless, the explicit publicly available curricula options and course descriptions are what perspective students see and may very well impact their decision to apply for/enroll in a program or particular courses. Indeed, the fact that programs continue to include these courses in their official programs of study (as core or elective courses) implies that they continue to be considered, to some extent, valuable.

Recommendations and Implications for Future Research

Our review, though comprehensive, focused solely on measurement/psychometric programs in the United States and, consequently, did not address the

doctoral training professionals educated abroad typically receive. To be sure, the field of educational measurement is practiced and includes researchers/practitioners from all over the world who are trained in non-U.S. doctoral programs. Additional studies that compare doctoral curricula across the globe would provide even more meaningful insight into what is valued in all measurement programs.

Another area of future research not addressed in this study is the need to better understand the graduate student perspective. To date, previous studies have focused on (a) current curricula practices within graduate programs and/or (b) feedback from faculty and/or working professionals about their curricula choices and needs. To be sure these stakeholders (faculty and working professionals) have extensive experience and a more comprehensive view of the field and its needs than incoming and current students. Still, having a greater understanding of what students believe to be valuable as they navigate their graduate educations seems prudent. Indeed, one might reasonably assume that students consider these values when making decisions about which doctoral programs to apply to and attend. To that end, interviews with students to determine what they believe are critical skills and content could also inform higher-level decisions about program curricula.

Finally, as we noted earlier, currently the field lacks a set of basic competencies/content knowledge for practicing measurement professionals. Indeed, individuals can declare themselves measurement professionals virtually at will. Likewise, programs can list themselves as measurement programs with little to no evidence or oversight. Given the importance of the work and the vulnerability of some of the populations (i.e., historically marginalized youth) the field often serves, the lack of even a rudimentary set of standards is problematic. Consequently, we recommend that NCME (the largest measurement organization in the world) begin the process of facilitating discussions across and between measurement programs and the industry about what it means to be a measurement professional in an effort to establish some consensus. It is our hope that the findings from this study will assist in initiating these conversations.

References

- Aguirre, J., Turner, E., Bartell, T., Kalinec-Craig, C., Foote, M., McDuffie, A., & Drake, C. (2012). Making connections in practice: How prospective elementary teachers connect to children's mathematical thinking and community funds of knowledge in mathematics instruction. *Journal of Teacher Education, 64*(2), 178-192
- Aiken, L., West, S., Sechrest, L. & Reno, R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A Survey of PhD programs in North America. *American Psychologist, 45* (5), 721-734.
- Aiken, L., West, S., & Millsap, R. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's 1990 Survey of PhD programs in North America. *American Psychologist, 63* (1), 32-50.
- Childs, R. & Eyde, L. (2002). Assessment training in clinical psychology doctoral programs; What should we teach? What do we teach? *Journal of Personality Assessment, 78*(1), 130-144.
- Creswell, J. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. 4th Edition, Pearson: Boston.
- DeVos, A.S., Strydom, H., Fouche, C.B., & Delport, C.S.L. (2011). *Research at grass roots. For the social sciences and human service professions*. 4th edition. Pretoria: Van Schaik.
- Eglash, R., Bennett, A., O'Donnell, C., Jennings, S., & Cintonino, M. (2006). Culturally situated design tools: Ethnocomputing from field site classroom. *American Anthropological Association, 108*(2), 347-362.
- Gates, B. (2014). Bill Gates: Commend Common Core. USA Today, Published Feb. 12, 2014 at 6:47 am. Retrieved August 17, 2020 from <https://www.usatoday.com/story/opinion/2014/02/11/bill-melinda-gates-common-core-education-column/5404469/>
- Gerwitz, C. (2012). Advocates worry implementation could derail common core. *Education Week, 31* (29), s6-s9. Retrieved August 17, 2020 from <https://www.edweek.org/ew/articles/2012/04/25/29cs-overview.h31.html>
- Gilbert, R. (2004). A framework for evaluation the doctoral curriculum. *Assessment and Evaluation in Higher Education, 29* (3), 299-309.
- Gonzalez, N., Andrade, R., Civil, M., & Moll, L. (2005). Funds of Distributed Knowledge (Ch. 15). *Funds of Knowledge: Theorizing Practices in Households, Communities, and Classrooms*, p.257-270.
- Gordon, E. (1995). Toward an Equitable System of Educational Assessment. *The Journal of Negro Education, (3)*, 360. <https://doi.org/10.2307/2967215>
- Guo, F., & Nitko, A. (1996). Graduate programs that prepare educational measurement specialists. *Educational Measurement: Issues and Practice*, 28-31.
- Lee, C. (1998). Culturally Responsive Pedagogy and Performance-Based Assessment. *The Journal of Negro Education, 67*(3), 268. <https://doi.org/10.2307/2668195>
- Merenda, P. (1996). Note on the continuing decline of doctoral training in measurement. *Psychological Reports, 78*, 321-322.
- Randall, J., Rios, J., & Jung, H. (2020). A Longitudinal analysis of doctoral graduate supply in the educational measurement field. *Educational Measurement: Issues and Practice*. Retrieved September 16, 2020 from <https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12395>.

Citation:

Randall, J., Rios, J.A., & Jung, H. (2021). Graduate Training in Educational Measurement and Psychometrics: A Curriculum Review of Graduate Programs in the U.S. *Practical Assessment, Research & Evaluation, 26*(2). Available online: <https://scholarworks.umass.edu/pare/vol26/iss1/2/>

Corresponding Author

Jennifer Randall
College of Education, Department of Educational Policy, Research, and Administration
University of Massachusetts Amherst
Amherst, MA, 01003

Email: jrandall [at] umass.edu

Appendix A

Table A-1. Operational Definitions

Course Category/ Designation	Operational Definition
Educational Measurement	Included in this category are course topics that: (a) describe procedures on the development and administration of assessments to ensure reliable and valid measures, and (b) are concerned with the development and application of statistical techniques to assign numerical estimates to performance on educational and psychological measures. Example courses include test development, survey development, validity theory, classroom assessment, Classical Test Theory, Item Response Theory, Generalizability Theory, and diagnostic classification models.
Pedagogical/ Educational Theory	Pedagogical theory concerns itself with how topics ought to be taught and how students learn best. Course topics include the philosophy of education and student learning (philosophical learning theory), educational psychology, classroom management, instructional design, instructional technology.
Psychological Theory	Psychology is the study of human behavior and the function of the mind at different life stages. Course topics in this area deal with motivation, emotion, cognition, and memory. Topics on how people learn (mechanistic learning theory) as well as courses in biology, neurobiology, physiology, neurophysiology, psychopharmacology, health planning and counseling, human development, and behavioral science fall under this category. Courses in psychometrics are excluded.
Statistics	Course topics describe methodologies for evaluating numerical data collected from educational and psychological measures to make descriptive and inferential claims. Example courses include multivariate statistics, structural equation modeling, longitudinal modeling, hierarchical linear modeling, matrix algebra, nonparametric statistics, and factor analysis; as well as computer courses such as programming and data mining. This category excludes courses that fall under psychometrics, and research methods courses that cover some statistical analysis.
Evaluation	Program evaluation courses cover topics related to how systems are evaluated, with the end goal of judging or informing a program, or making program decisions. Course topics may focus on data collection or analysis for the purpose of evaluation, either generally or for specific types of programs. Human evaluation (how humans perceive the world) is not included here.

Sociology & Anthropology	Coursework includes classes with a focus on sociology and anthropology. Additionally, while most courses in this category will be labeled with either of these titles, courses that focus on the general study of humans and their relationship to their surroundings are included (excluding health and psychology courses).
History	Coursework covering the past, with a focus on educational issues were coded with this designation. Courses explore the history of educational systems at every level, including the institutions, theories, and other topics about the evolution of various aspects of education. This excluded courses that provide history as a means to introduce the substantive area of interest.
Educational Policy	Included in this category is coursework covering the principles of action adopted or proposed to govern educational systems and practices associated with improving learning, building teachers' capacity, and engaging stakeholders. Example courses include social policy and educational law.
Professional Development	This category includes coursework that is dedicated to improving and increasing capabilities of students to become effective researchers and practitioners (includes course credit based on internships, consulting, supervised research, writing, and communication). These capabilities go beyond content knowledge in the areas of measurement, statistics, and research methods by focusing on the competencies necessary for the day-to-day activities of communicating results (oral and written) and workplace professionalism. Included are apprenticeships, courses in ethics and standards, labs, and practicums.
Diversity & Inclusion	Courses were given this designation if they focused on any special population for which equity measures (i.e., extra measures such as tutoring, or separate classes) may be taken. This includes English learners, special education students, and gifted and talented students. Courses may also critically analyze race and gender issues. Example courses include <i>Race and Racism in Education and Society</i> , <i>Families and Disabilities</i> , and <i>Gender and Education</i> . It should be noted that this category is often used in conjunction with another category (e.g., pedagogical theory).
Miscellaneous	Miscellaneous is a catch-all term utilized to categorize courses that did not meet the definitions of course types listed above. Examples include special topic courses, dissertation credits, and directed studies.

Appendix B

Curriculum Review – Qualitative Interview Protocol

1. Talk about your current position – your role and responsibilities.
 - a. Follow up Question: Previous jobs [in field] and responsibilities
2. How did you learn about/come to be in the measurement profession?
3. Where did you earn your measurement degree? When?
 - a. Where did you earn your BA degree? When? In What?
 - b. Did you also earn a master's degree? When? Where? In What?
4. Describe your work experiences (if any) before entering the field of measurement.
5. What are the five most beneficial courses that you took as a measurement student for your current role?
 - a. Follow up Question: Previous role?
6. What content/knowledge do you wish you had/was missing from your measurement program?
7. What are the skills/content that you believe are available/attainable through work experience only (i.e., not via course work)?
 - a. Follow Up: Did your measurement program provide you with those experiences? Or facilitate your attainment of those skills/content?
8. If you were asked to provide one or two pieces of advice to measurement programs in terms of their preparation of measurement professionals, what advice would that be?
 - a. Note: Make sure advice includes at least one DO and one DO NOT

For Supervisors Only:

1. What are the skills you are looking for in new hires? Non-negotiables? What skills are nice, but not critical to have?
 - a. Psychometricians?
 - b. Research scientists?
 - c. Data Analysts?
 - d. Other

All:

Please review the list of courses below and rate them as (a) nice, but not critical, (b) critical, or (c) limited/no value:

Research Designs

- Experimental & Quasi-Experimental Designs
- Non-Experimental Designs

Evaluation

Ethical Research Practices

Analysis of Variance

- Factorial, ANCOVA, repeated measures

Regression

- Linear, multiple linear, logistic

Meta-analysis

Multi-Level Modeling

Multivariate Statistics

Nonparametric Statistics

Bayesian Statistics

Validity Theory

Classical Test Theory

Item Response Theory

Equating (CTT)

Equating (IRT)

Multidimensional IRT

Generalizability Theory

Structural Equating Modeling

Multidimensional Scaling

Computer Programming

- Python
- C++
- R
- SPSS
- STATA
- SAS
- MPlus
- Winsteps/Facets
- Other (please specify):

Test Development

Scale Development

History of Measurement or Assessment

Classroom Assessment

Assessment/Measurement Policy

Qualitative Research Methods

Other (Please Specify):