APSY-GE 2524

PSYCHOLOGICAL MEASUREMENT

Steinhardt School of Culture, Education, and Human Development
New York University

Tuesdays, 11am-1:30pm
246 Greene St., Kimball Hall, Room 506W

Instructor:
Alejandro J. Ganimian
Assistant Professor of Applied Psychology and Economics
246 Greene St., Kimball Hall, Room 406W
+1 212 998-4029
alejandro.ganimian@nyu.edu

Office hours:
https://calendly.com/alejandro-ganimian/office-hours

1. Objectives

This course seeks to introduce students to key considerations in the design, administration, and analysis of instruments for psychological research. It focuses on three overarching questions: (a) how can we design instruments to measure our construct(s) of interest?; (b) how can we administer instruments to maximize the amount of useful information we will obtain (and conversely, minimize error)?; and (c) how can we analyze individuals' responses to accurately represent the measurement procedure? It offers an overview of approaches applicable to a wide array of instruments used in psychology, but an important part of the course focuses on psychological and educational measurements in schools (e.g., scales of social-emotional skills, achievement tests, and classroom observations). It draws on examples of quantitative research from psychology and economics.

The components of the course aim to achieve different, but complementary, objectives:

- The readings, to be completed before each lecture, will introduce students to a problem in measurement (e.g., measurement error), the conceptual frameworks that can be used to think about this problem (e.g., classical and generalizability theory), and the analytical strategies employed to address the problem (e.g., Cronbach's alpha and G-studies).
- The lectures will briefly review the problem introduced in the readings, discuss its implications in greater detail, and compare different approaches to solve the problem, drawing extensively on examples from applied research.
- The problem sets, which can be completed in pairs, but must be written-up individually, will provide students with opportunities to practice implementing the approaches discussed in lectures on their own using a statistical package.
- The final take-home exam (for master's students) or project (for doctoral students), which must be completed individually, will assess students' ability to apply the material covered in the course independently.

The sequencing of these components (i.e., the fact that students will first complete the readings, then come to lecture, complete problem sets in pairs, and finally apply what they learn independently) aims to provide students with the necessary scaffolding to become critical consumers of research in psychological measurement. By the end of the course, students will be expected to understand the concepts, methods, and analytical strategies on their own.

This course draws on many other related classes, including: Statistical and Psychometric Methods for Educational Measurement (taught by Daniel Koretz and Andrew Ho), Introduction to Test Theory (taught by Ben Domingue), Survey Design and Analysis (taught by Morgan Polikoff), Measurement in Survey Research (taught by Benjamin Shear), and Survey Research Methods (taught by Daphna Harel). The instructor thanks instructors who shared their materials.

2. Pre-requisites

Students are expected to have taken APSTA-GE 2001 ("Statistics for the Behavioral and Social Sciences") or APSTA-GE 2003 ("Intermediate Quantitative Methods: The General Linear Model") or equivalent courses before taking this course. They are also expected to review the chapter on "Statistical concepts for test theory" posted on the syllabus for Lecture #1. For additional support, the instructor has posted the syllabus and slides for a previous introductory statistics course that he has taught. Students are encouraged to use those slides for review. Students who are unsure as to whether they meet these pre-requisites should make an appointment to see the instructor during office hours (see link on the first page of the syllabus).

3. Auditing

This course may be taken for a letter-grade only, not on a satisfactory/no credit basis. Auditors are not allowed for two reasons. First, students are unlikely to master the material in the course if they do not complete all requirements (i.e., attend class regularly, participate, and complete the problem sets and exam or project). If a student plans to complete these requirements, they should receive credit. Second, the instructor works hard to support registered students throughout the semester. Auditors place additional demands on the instructor, which invariably limit his capacity to provide this support.

4. Readings

There is no textbook for this course. Instead, the instructor will post scanned versions of the readings assigned each week on the Contents tab of the course site: https://brightspace.nyu.edu/d2l/home/351466.

Many of the assigned readings will draw on the following texts:

- Crocker, L. & Algina, J. (2008). *Introduction to classical & modern test theory*. Cengage Learning.
- Groves, R. M. et al. (2009). *Survey methodology (2nd edition).* Wiley.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Harvard University Press.

- Raykov, T. & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Taylor & Francis.
- Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
- Thorndike, R. M. & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education (8th edition)*. Pearson.
- Wilson, M. (2005). *Constructing measures: An item-response modeling approach.* Psychology Press.

The readings will often introduce new concepts with which students may be unfamiliar and use mathematical notation that students may not have seen in a while. Students are not expected to understand the readings in detail before each lecture, but they must have completed them and made a good-faith effort to develop an initial understanding.

5. Grading

Each student's grade in the course will be determined as follows:
a) attendance (5%);
b) class participation (15%);
c) three problem sets (50%); and
d) final take-home exam or project (30%).

Attendance and punctuality: Students are expected to attend all class meetings, arriving *before* the start of each meeting to allow the class to start on time. In accordance to school and department policies, students will be allowed up to two excused absences during the semester. An "excused" absence is one in which the student has notified the instructor either 24 hours before or 48 hours after the day of the absence. If the absence is due to illness, no supporting documentation is required and students are not expected to disclose private health information. If the absence is not due to illness, such documentation is required when notifying the instructor. An "unexcused" absence is one in which the student has not notified the instructor within the time specified above. If a student has more than two absences (excused or unexcused), the instructor is required to alert their graduate advisor.

In accordance to NYU's calendar policy on religious holidays, students who let the instructor know of their absences due to religious holidays ahead of time will not incur any penalty. However, they are still expected to post the "prepared" questions before the classes they miss and reaction memos afterwards (relying on class recordings). If students require extensions on prepared questions or reaction memos due to religious holidays, they should reach out to the instructor early so that such extensions may be extended to all other students.

Each student's attendance score will be calculated as follows. The student will receive a score of 1 for attending each meeting before the official start time, a score of 0.5 for arriving after the official start time, and a score of 0 for an unexcused absence. The student's total attendance score will be the sum of all the individual scores over the total number of meetings, multiplied by 100. For example, if a student attended 12 of 14 meetings, their score will be (12/14)*100 or 86. The maximum attendance score is 100.

For reference, the mean attendance scores for previous iterations of this course were: 94 (fall 2017) and 97 (spring 2020).

Class participation: During each lecture, students may answer questions from the instructor about the readings and/or ask clarifying questions themselves. All three of these types of interventions will be considered in the 15% of the unadjusted course grade assigned to class participation.

Each student's class participation score will be calculated as follows. On each lecture, a student will receive a score of 1 for making a good-faith effort to participate (even if they do so incorrectly) or a score of 0 for attending class but refraining from participating. The student's total participation score will be the sum of the scores for all lectures over 10, multiplied by 100. Based on this scheme, to obtain a perfect class-participation score by the end of the semester, a student must have participated on 10 instances (out of a total of 14 lectures). For example, if a student participated in 9 of 14 lectures, his/her/their score will be (9/10)*100 or 90. The maximum participation score is 100.

For reference, the mean participation scores for previous iterations of this course were: 79 (fall 2017) and 95 (spring 2020).

Problem sets: Students are expected to complete three problem sets throughout the semester. As stated in the course objectives, these problem sets are meant to provide students with opportunities to practice the material covered in lecture. Students can complete problem sets in groups (ideally, pairs), but they must write up their results individually. Instructions on how to format and submit problem sets will be included at the beginning of each assignment. The problem sets from previous iterations of the course are posted on the "Resources" tab of the course site. These are meant to provide students with general guidance on the expected level of detail of their answers and the instructor's approach to grading. Yet, the content and types of questions in problem sets vary from one semester to the next as the course continues to evolve.

Each student's problem-sets score will be calculated as follows. The student will receive a score of 0 to 100 on each problem set, based on the proportion of questions they answered correctly. Partial credit will be awarded for partially correct answers, so students are encouraged to show their work. The student's overall problem set score will be the average of the two highest problem-set scores (i.e., the lowest score will not count). This provision is meant to account for the fact that some students may find some of the problem sets more difficult than others, and to prevent one low problem set score from playing a large role in determining students' overall grade. It is also meant to allow students to "drop" (i.e., choose not to complete) one problem set during the semester (e.g., if they cannot complete the problem set on time due to unforeseen circumstances). For example, if a student obtained scores of 50, 80, and 100, his/her score will be (80+100)/2 or 90. The maximum problem set score is 100.

For reference, the mean problem-sets scores for previous iterations of this course were: 84 (fall 2017) and 97 (spring 2020).

Final take-home exam or project: Master's students are expected to complete one final take-home exam. As stated in the course objectives, the exam aims to assess students' ability to apply

the material covered in lecture independently. Students must complete the exam individually. Doctoral students are expected to complete one final project. Ideally, these projects will help students make progress towards required submissions for their respective doctoral programs (e.g., qualifying papers or dissertations).

Note that these are simply the "default" options for master's and doctoral students. If students wish to switch (e.g., a master student wants to complete a final project instead of the take-home exam), they can do so by simply notifying the instructor over e-mail. However, all students who are scheduled to complete a final project must comply with all the milestones outlined in the course calendar below, regardless of whether they were assigned to a project by default or whether they made the switch during the semester.

The final exams and projects from previous iterations of the course are posted on the "Resources" tab of the course site. The exams are meant to provide students with general guidance on the expected level of detail of their answers and the instructor's approach to grading. Yet, the content and types of questions vary from one semester to the next as the course continues to evolve. The projects are meant to illustrate the types of questions students examine, as well as the expected length, structure, and format of the final projects. These vary based on the intended purpose of the project (e.g., if a student plans to use the project a dissertation appendix, his/her write-up will differ from that of a peer who will use it a second-year paper).

Each student's final-exam or project score will be calculated as follows. The student will receive a score of 0 to 100 on the exam or project, based on criteria to be specified before/after each assignment (in the case of the exam, the instructor will post an answer key after grading all exams; in the case of the projects, the instructor will post instructions for each milestone). In the exam, partial credit will be awarded for partially correct answers, so students are encouraged to show their work. For example, if a student obtained a score of 90, that will be his/her score.

For reference, the mean final-exam scores for previous iterations of this course were: 82 (fall 2017) and 88 (spring 2020). The mean final-project score in the spring of 2020 was 95 (there was no option to complete a final project in the fall of 2017).

Overall course grade: The overall numeric score for each student will be calculated as the weighted average of his/her attendance, class participation, problem sets, and final exam or project. The weights correspond to the percentages allotted to each score above. For example, if a student obtained an 86 for his/her attendance, a 93 for his/her class participation, a 90 for his/her problem sets, and a 90 for his/her final exam or project, his/her overall numeric score will be (86*0.05)+(93*0.15)+(90*0.5)+(90*0.3) or 90.

The overall letter grades will be determined based on the distribution of numeric scores for all students in the course. This is meant to account for the fact that some student cohorts may find the material more or less difficult than others. Letter grades will be assigned as follows:

If a student has a numeric score that is…	…they will earn a/an…
…0.5 standard deviation (SD) above the mean…	…A
…above the mean by less than 0.5 SD…	…A-

…below the mean by less than 0.5 SD…	…B+
…between 0.5 and 1 SD below the mean …	…B
…between 1 and 1.5 SD below the mean…	…B-
…between 1.5 and 2 SD below the mean…	…C+ or lower

Students interested in understanding their relative standing in the course at any point during the semester should make an office-hours appointment. This mechanism is not meant to raise the costs of finding out your grade, but rather to use your grade as a starting point for a broader conversation on your performance and what you need to do to succeed in the course.

The cutoff scores have varied across semesters as follows:

			Spring 2020		Spring 2022
Criterion	Letter grade	Fall 2017	MA students	PhD students	PhD students
>=0.5 SDs above the mean	…A	88	96	97	92
<0.5 SDs above the mean	…A-	83	94	90	89
<0.5 SDs below the mean	…B+	78	92	83	87
>=0.5 and <1 SDs below the mean	…B	72	90	76	85
>=1 and <1.5 SD below the mean	…B-	67	88	69	83
>=2 SDs below the mean	…C+ or lower	62	86	62	81

The instructor may (and often does) adjust a student's final letter grade on the course based on his/her improvement over time and exemplary performance on one or more dimensions, so the actual distribution of letter grades is never determined exclusively by the cutoff scores above.

All grades posted at the end of the semester are final and the instructor will not discuss grades over e-mail. Students interested in better understanding their grades after they are posted are welcome to make an appointment with the instructor at the start of the following semester. There will be no exceptions to ensure no students are given an unfair advantage over others.

Grading criteria for assignments: After each problem set is graded, the instructor will post the answer key, scoring criteria, and student exemplars (i.e., anonymized problem sets with top scores, with students' permission). Students are strongly encouraged to consult these documents to ask the teaching team any questions they might have on the material.

A student may ask for his/her problem sets and/or mid-term exam to be regraded if—after carefully reviewing the answer key, scoring criteria, and exemplars—they do not believe that his/her grade is correct. Students who wish to request a regrade should e-mail the instructor no later than one week after scores have been posted. The instructor will conduct all regrades. He will regrade the entire problem set or exam, not just the questions being disputed. Therefore, regrades may result in a lower, equal, or higher scores than the ones originally awarded. The final exam and project scores are final (i.e., not subject to regrades).

6. Classroom policies and expectations

Laptops and tablets: Evidence from multiple randomized experiments indicates that students who take notes on their laptops or tablets learn less and earn worse grades than those who take notes using pen/pencil and paper. They are also more likely to adversely affect their peers' learning and grades. (See Prof. Susan M. Dynarski's summary of the evidence at: http://brook.gs/2vS6I3e). Therefore, laptop and tablet use are discouraged during lectures. Exceptions may be made, especially for devices that are not connected to the Internet.

The instructor will bring printouts of lecture slides, and students may bring printouts of any materials that they may need during lecture (e.g., assigned readings, prepared questions, etc.) Students who require financial support to print out such materials should notify the instructor. Note that the instructor often edits slides (e.g., to correct typos or incorporate aspects that arose during class discussions) and posts final versions after each lecture. Students should use those final versions as reference for course assignments.

Cell phones: Cell phone use (for making or receiving calls and sending or receiving text messages) is prohibited during lectures. There will be no exceptions.

Eating and drinking: Students who need to eat during class should clean after themselves to avoid creating additional work for maintenance workers who clean the university's spaces. Students may also bring water bottles or coffees/teas in covered containers.

Late assignments: Students should budget enough time to submit all course assignments well ahead of each deadline. Late assignments, regardless of how late they are (even a minute past the deadline), will not be accepted for three main reasons. First, the class already has a built-in system to account for unanticipated events: dropping the lowest problem-set score (see Grading section above). Second, the process of granting exceptions is inevitably inequitable: for every student who requests an extension, there are often many others who would have also benefited from such an extension but were too shy to request it. In the instructor's experience, it is often students from more advantaged backgrounds who fall into the first group and those from disadvantaged backgrounds who fall into the second group, perpetuating pre-existing trends in inequality in academic socialization. Third, the teaching team has no way to determine whether some circumstances are more meritorious of an extension than others. For all these reasons, there will be no exceptions.

Surveys: The instructor will invite students to complete two surveys during the semester: a "student survey" (at the beginning of the semester), which will allow him to get to know students better, and a "feedback survey" (midway through the semester), which will allow students to provide feedback on what is working well and what could be improved in the course. The instructor takes feedback surveys very seriously and it will make a good-faith effort to address the concerns raised by students.

All surveys are optional and there will be no repercussions for students who choose not to answer them. The student survey will ask for identifying information (to avoid asking questions for which the instructor already has information), but the feedback survey will be anonymous.

None of the surveys will be considered in students' course grades. All data survey responses will be deleted at the end of the course and it will not be used for other purposes.

7. Statistical programming

All students will need to get access to Stata, a statistical package, to complete the problem sets for this course. All the example code to be provided by the instructor will be written in Stata 15, so students should get access to Stata 15 or above.

Students may get access to Stata on campus, through the computers at Data Services (on the fifth floor of Bobst Library), the Student Technology Centers (LaGuardia Co-op, Kimmel Center Lab, and Third Avenue Lab; see http://bit.ly/2xgqvHg), or the High Performance Computing's Prince cluster (see https://bit.ly/31Rr4Wq).

Students may also get access to Stata off campus through the Virtual Computer Lab at: http://www.nyu.edu//it/vcl.

Finally, students may purchase Stata at a discounted rate through Stata Campus GradPlan at: http://bit.ly/2w1DrCc. An annual license for Stata/IC (the version for mid-sized datasets), which will be sufficient for this course, is $125.

Lectures will not be used to teach students how to code, but the instructor will upload step-by-step guides with all the commands that students will need for the problem sets to the course site. Students are encouraged to attend office hours to ask coding questions.

Additionally, students can seek help with coding from Data Services (on the fifth floor of Bobst Library) either by signing up for their Stata tutorials (see calendar at https://guides.nyu.edu/DS_class_calendar) or by making an appointment for a one-on-one meeting with a consultant (see https://library.nyu.edu/departments/data-services/.)

Students who believe that they would benefit from a book on Stata are encouraged to consult:

- Kohler, U. & Keuter, F. (2009). *Data analysis using Stata (2nd Edition).* College Station, TX: Stata Press.

Students who believe that they would benefit from an introductory book to probability are encouraged to consult:

- Blitzstein, J. K. & Hwang, J. (2019). *Introduction to Probability (2nd Edition).* Free online access: https://bit.ly/38Thki5. Print copies: https://www.crcpress.com.

Students may also consult the introduction to probability course at Harvard University: https://projects.iq.harvard.edu/stat110.

8. Writing

The problem sets, exams, and projects will involve a fair amount of writing (e.g., to define key concepts or explain results from statistical analyses). Students should not take this writing lightly; an important part of becoming a researcher is learning to convey arguments clearly.

Students are expected to review their assignments for typos and grammatical errors before submitting them. They should also take full advantage of the various on-campus resources to help them improve their writing, including the Writing Center (https://bit.ly/2PMe13x) and the University Learning Center (https://bit.ly/2hBrgX0).

9. Plagiarism

Students taking this course are expected to have read in full and agreed to NYU-Steinhardt's statement on academic integrity (http://bit.ly/2vSt2JR).

As the statement specifies, "plagiarism is failure to properly assign authorship to a paper, a document, an oral presentation, a musical score and/or other materials, which are not your original work." Therefore, any student who works together with or receives help from others on the problem sets should recognize their contributions appropriately (instructions for doing so will be provided in each problem set). This will help the instructor understand any similarities in assignments submitted by different students.

Students who have questions about what constitutes appropriate collaboration in this course should contact the instructor at least 24 hours before they submit their problem sets.

If the instructor suspects that a student has committed plagiarism, disciplinary action may be taken following the department procedure or through referral to the Committee on Student Discipline, through the Office of the Associate Dean for Student Affairs. Please, see the statement on academic integrity for details on the steps involved in each procedure.

10. Accommodations

Any student who needs an accommodation due to a chronic, psychological, visual, mobility and/or learning disability, or who is deaf or hard of hearing, should register with the Moses Center for Students with Disabilities (www.nyu.edu/csd) at 212 998-4980, 726 Broadway, 2nd and 3rd Floors.

Students should also notify the instructor within the first week of the semester. Late requests for accommodation will not be honored except in special circumstances (e.g., injury during the semester).

11. Calendar

This course calendar is tentative. The instructor may adjust the topics to be covered in each class based on how students respond to the material during the semester. Students are expected to check the latest version of the calendar on the course site before every lecture.

Date	Topics	Readings	Assignments
Jan 23	Lecture #1: Introduction to the course • What are the objectives and components of the course? • What is measurement? • How is psychological measurement different? • What is Stata?	Required: • Thorndike & Thorndike-Christ, Ch. 1 (especially pp. 2-7 on the history of educational measurement and pp. 16-20 on current issues in measurement) • Crocker & Algina, Ch. 2 (make sure you understand the statistical concepts for test theory; otherwise, come see me in office hours) • Wilson, Ch. 1 (make sure you understand the four "building blocks" of instrument development) • Raykov & Marcoulides, Chs. 1 and 2, pp. 13-21 (these readings go in depth into the concepts presented during lecture) • Koretz, Ch. 2 (especially, pp. 21-27 on the "sampling principle of testing") Recommended: • Duckworth, A. (2016). "Don't grade schools on grit." *New York Times*. March 26, 2016. • Koerth, M. & Wolfe, J. (2019). "Most personality quizzes are junk science. Take one that isn't." *FiveThirtyEight*. January 16, 2019. • John, O. P. & Srivastava, S. (1999). "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives." In L. A. Pervin & O. P. John (Eds.) *Handbook of personality: Theory and research (Vol. 2)*, pp. 102-138. New York, NY: Guilford Press. • Leonhardt, D. (2024) "The misguided war on the SAT." *New York Times*. January 7, 2024. • Radiolab (2019). "G series." (six-episode podcast series on the measurement of intelligence). *Radiolab.* June 7-July 30, 2019.	• Student survey posted

Jan 30/ Feb 6	Lectures #2-3: How can we know if we can use an instrument for a given purpose? (Validity and validation) • What is a construct map? • What is validity and validation? • How can we describe the relationship between two variables? (bar graphs, scatterplots, and correlations) • What are the different sources of validity evidence? (content, construct, and criterion validity) • What are threats to validity? (construct underrepresentation and construct-irrelevant variance)	Required: • Wilson, Ch. 2 (skim pp. 29-38; read the rest carefully) • Koretz, Ch. 9 (focus on definition of validity, different types of validity, construct underrepresentation v. construct-irrelevant variance, and different approaches to validation) • AEA/APA/NCME (2014). "Validity," *Standard for educational and psychological testing*. Washington, DC: American Educational Research Association. (skim pp. 17-24; read the rest carefully) • Raykov & Marcoulides, Ch. 8, pp. 183-192 • Duckworth, A. L. et al. (2007). "Grit: Perseverance and passion for long-term goals," *Journal of Personality and Social Psychology*, 92(6), 1087-1101. (do not worry about the methods/terms that you have not yet learned) Recommended: • Molina, E. et al. (2020). "Measuring the quality of teaching practices in primary schools: Assessing the validity of the Teach observation tool in Punjab, Pakistan," *Teaching and Teacher Education*, 96, 103171. • Ahluwalia, R. et al. (2023). "Phone-based assessment data: Triangulating schools' learning outcomes," *Ideas for India*. January 11, 2023. • Papay, J. P. (2011). "Different tests, different answers: The stability of teacher value-added estimates across outcome measures," *American Educational Research Journal*, 48(1), 163-193. • Lajaj, R. & Macours, K. (2021). "Measuring skills in developing countries," *Journal of Human Resources*, 56(4), 1254-1295. • Danon, A. et al. (2024). "Cognitive and socioemotional skills in low-income countries: Measurement and associations with schooling and earnings," *Journal of Development Economics*, 168, 103132. • Koepp, A. E. et al. (2021). "Measuring children's behavioral regulation in the preschool classroom: An objective, sensor-based approach," *Developmental Science*, e13214. • Kane, M. T. (2006). "Validation," *Educational measurement (4th edition)*. NCME and ACE. Praeger.	• Student survey due

Feb 13	Lecture #4: How can we check whether items in an instrument measure a construct? (Factor analysis, part 1) • How can we design items? • How can we explore whether item responses are caused by one or more constructs? (exploratory factor analysis)	Required: • Wilson, Ch. 3 • Raykov & Marcoulides, Chs. 3 (read pp. 52-59 without focusing on the SPSS code or output, since we will use Stata) Recommended: • Ganimian, A. J. et al. (2020). "Hard cash and soft skills: Experimental evidence on combining scholarships and mentoring in Argentina," *Journal of Research on Educational Effectiveness*, 13(2), 380-400. • Duckworth, A. L. & Yeager, D. S. (2015). "Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes," *Educational Researcher*, 44(4), 237-251.	• Problem set 1 posted
Feb 20	Lecture #5: How can we know if an instrument yields consistent results? (Reliability) • How can we determine possible scores for items? • What is reliability? • What is the most commonly used approach to measure reliability? (classical test theory) • How can we measure inter-item reliability? (split-half reliability, the Spearman-Brown formula, and Cronbach's alpha) • How can we measure inter-rater reliability?	Required: • Wilson, Ch. 4 (skim through examples, read the rest) • Koretz, Ch. 7 (focus on definition of measurement error and reliability) • AEA/APA/NCME (2014). "Reliability and errors of measurement," *Standard for educational and psychological testing*. Washington, DC: American Educational Research Association. (skim pp. 31-36; read the rest carefully) • Raykov & Marcoulides, Chs. 5, pp. 115-123; 6, pp. 137-143 and 144-146, and 7, pp. 147-152 and 154-158 Recommended: • Haertel, E. H. (2006). Ch. 3, pp. 65-79 • Barrera-Osorio, F. & Ganimian, A. J. (2016). "The barking dog that bites: Test score volatility and school rankings in Punjab, Pakistan," *International Journal of Educational Development*, 49, 31-54. • Singh, A. (forthcoming). "Improving administrative data at scale: Experimental evidence on digital testing in Indian schools," *Economic Journal*.	• Problem set 1 due

	(inter-rater agreement and Cohen's kappa)	• Singh, A. & Berg, P. (2023). "Myths of official measurement: Limits to test-based education reforms with weak governance," *Unpublished manuscript*, Stockholm, Sweden: Stockholm School of Economics. • Haertel, E. H. (2006). "Reliability" (pp. 65-79), *Educational measurement (4th edition)*. NCME and ACE. Praeger.	
Feb 27	Lecture #6: How can we know if an instrument yields consistent results? (Generalizability, part 1) • What is a more general approach to measure reliability? (generalizability theory) • How can we measure reliability across multiple facets of error? (G-studies with "crossed" designs) • How can we use estimates of reliability to improve measurement procedures? (the D-studies) • The G-study in Stata • The D-study in Excel	Required: • Brennan, R. L. (1992). "Generalizability theory." *Instructional Topics in Educational Measurement*. • Shavelson & Webb, Chs. 1 and 2 Recommended: • Hill, H. C. et al. (2012). "When rater reliability is not enough: Teacher observation systems and a case for the Generalizability study," *Educational Researcher*, 41(2), 56-64.	• Final-project proposal due
Mar 5	Lecture #7: How do we know if an instrument yields consistent results? (Generalizability, part 2) • How can we measure reliability when there are crossed designs with two or more facets of error?	Required: • Shavelson & Webb, Chs. 3 and 4 • Shavelson & Webb, Chs. 6 and 7 Recommended: • Kane, T. J. & Staiger, D. O. (2012). "Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains," Seattle, WA: Bill & Melinda Gates Foundation.	• Feedback survey posted

	• How can we measure reliability when facets of error are "nested" within individuals? (G- and D-studies with nested designs)	• Ho, A. D. & Kane, T. J. (2013). "The reliability of classroom observations by school personnel," Seattle, WA: Bill & Melinda Gates Foundation.	
Mar 12	Lecture #8: How can we score achievement tests and personality scales to account for differences across items? (Item response theory, part 1) • How can we translate item responses into scores? • What is item response theory (IRT)? • What are the different types of IRT models? (1-, 2-, and 3-PL models) • What are the main assumptions of IRT? (local independence and unidimensionality) • What are two commonly used graphs about each test that we can obtain from IRT models? (test characteristic curves and test information curves)	Required: • Wilson, Ch. 5 (pp. 85-103; read "more than two score categories" only if it is of specific interest to you/your project) • Harris, D. (1989). "Comparison of 1-, 2-, and 3-parameter IRT models." *Instructional Topics in Educational Measurement*. Philadelphia, PA: National Council for Measurement in Education. • Yen, W. M. & Fitzpatrick, A. R. (2006). "Item response theory" (pp. 111-115), *Educational measurement (4th edition)*. NCME and ACE. Praeger. Recommended: • Andrabi, T. et al. (2002). "Test feasibility survey. Pakistan: Education sector," *Unpublished manuscript*, Claremont, CA: Pomona College. • Das, J. & Zajonc, T. (2010). "India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement," *Journal of Development Economics*, 92(2), 175-187. • Muralidharan, K. et al. (2019). "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 109(4), 1426-1460.	• Feedback survey due • Problem set 2 posted
Mar 19	*[Spring break – no class]*		

Mar 26	Lecture #9: How can we score student achievement tests to account for differences across items? (Item response theory, part 2) • Classical-test theory in Stata • 1PL, 2PL, and 3PL IRT models in Stata	• No readings assigned for this week.	• Problem set 2 due
Apr 2	Lecture #10: How can we map the results of two or more student achievement tests onto a common scale? (Linking and equating) • How do we typically compare the results of two tests? (predicting) • How can we collect data to allow for better comparisons? (single, equivalent, counterbalanced, and common-item anchor test designs) • How can we analyze data to allow for better comparisons? (mean, linear, and equipercentile linking) • Under what conditions can we treat the scores from two linked tests as	Required: • Kolen, M. J. & Brennan, R. L. (2010). Chs. 1 and 10 • Holland, P. W. & Dorans, N. J. (2006). Ch. 6, pp. 197-201 Recommended: • Sandefur, J. (2018). "Internationally comparable mathematics scores for fourteen African countries." *Economics of Education Review*, 62, 267-286. • Angrist, N. et al. (2021). "Measuring human capital using global learning data." *Nature*, 592, 403-408. • Bau, N. et al. (2021). "New evidence on learning trajectories in a low-income setting," *International Journal of Educational Development*, 84, 102430.	• Problem set 3 posted

	interchangeable? (equating)		
Apr 9	Lecture #11: How can we know if an item works differently across groups of respondents? (Differential item functioning) • How can we know if an item works differently for two groups of examinees with similar overall performance? (differential item functioning) • DIF in Stata • DIF in Stata using IRT	Required: • AEA/APA/NCME (2014). "Fairness in testing and test use." *Standard for educational and psychological testing*. Washington, DC: American Educational Research Association. (skim pp. 31-36; read the rest carefully) • Koretz, Ch. 11 • Camilli, G. (2006). "Test fairness", *Educational measurement (4th edition)*. NCME and ACE. Praeger. (read pp. 229, 236-239 only)	• Problem set 3 due
Apr 16	Lecture #12: How can we check whether items in an instrument measure a construct as expected? (Factor analysis, part 2) • How can we confirm that item responses are caused by one or more constructs? (confirmatory factor analysis)	Required: • Raykov & Marcoulides, Ch. 4 (read pp. 61-87 only, trying to understand the Mplus output but remembering we will use Stata) • Kline, Ch. 9 (focus on the intuition of the concepts covered; do not worry about understanding every aspect of notation) Recommended: • Duckworth, A. L. & Quinn, P. D. (2009). "Development and validation of the Short Grit Scale (Grit-S)." *Journal of Personality Assessment*, 91(2), 166-174. • Sandilos, L. E. et al. (2014). "Measuring quality in kindergarten classrooms: Structural analysis of the Classroom Assessment Scoring System (CLASS K-3)," *Early Education and Development*, 25(6), 894-914.	• Final-project first draft due
Apr 23	Lecture #13: How do we know if an instrument yields	Required: • Shavelson & Webb, appendix 4.2 (alternative nested, two-facet, random designs)	

	consistent results? (Generalizability, part 2) •	Recommended: • Shavelson & Webb, Chs. 5 and 8 (G- and D-study for models with fixed, instead of random, facets)	
Apr 30	Lecture #14: Review for the final exam • What are the key concepts and analytical strategies that we have learned? • How can we make decisions about research drawing on these concepts and strategies?	• No readings assigned for this week.	• Final take-home exam posted
May 7			• Final take-home exam due • Final-project paper due