S-061A1: Statistical and Psychometric Methods for Educational Measurement
Harvard Graduate School of Education
August 31 – October 17, 2016

Class meets Mondays and Wednesdays, 10:10-12:00, in Longfellow 229
Course website: https://canvas.harvard.edu/courses/18969/

Instructor: Andrew Ho
455 Gutman Library
Andrew_Ho@gse.harvard.edu

TF: Maria (Masha) Bertling
mbertling@g.harvard.edu

Instructor Office Hours: After class or email Wendy_Angus@gse.harvard.edu for appointments.

Description

This is the first of two sequential modules on quantitative methods for educational measurement. Students will learn and apply techniques essential for the design and analysis of educational and psychological assessments, including reliability, generalizability theory, validation, differential item functioning, item response theory, scaling, linking, standard setting, and adjustments for measurement error. Contexts of assessments include small-scale educational and psychological assessments for targeted research studies as well as large-scale district, state, and national assessments for formative, summative, and evaluative purposes. In this first module, the emphasis will be on learning and applying methods in class and through completion of data analytic assignments. In the second module, S-061A2, which students are required to enroll in subsequently, methods training will continue, with greater emphasis on reading and critiquing recent research in educational measurement and the development of a research proposal that has promise for advancing the field.

Prerequisite: S-052 or at least two semesters of applied statistics that includes estimation and interpretation of logistic regression coefficients. Enrollment in S-061A2 in the same semester is required. This course complements S-043 and S-090, and students may enroll in these courses in any order. Students who do not meet the prerequisite may enroll instead in S-011, which provides a nontechnical introduction to educational measurement.

Grading

The requirements of the course include regular attendance (5%), regular participation in class and in out-of-class Google Doc discussions (15%), satisfactory completion of 3 assignments (collectively 40% of the unadjusted course grade), and a take-home final exam (40% of the unadjusted course grade). These weights are approximate—the final course grade may factor in improvement over time and exemplary performance on one or more dimensions. Students are required to complete assignments in pairs. The final exam must be completed individually, without assistance from any "animate" resources.

This course is letter-grade-only; students may not take this course on a Satisfactory/No Credit basis. Registered students must submit a course evaluation form at the end of the semester in order to fulfill the requirements of the course. In-person auditing of this course is not allowed—all attendees must be registered students.

Support

Our TF, Masha Bertling, will hold weekly office hours by appointment. Occasionally, Masha may also offer optional discussion sections as she sees necessary; these may be most common around assignment due dates. Participation is strongly recommended. Office hours with the instructor are typically available after class meetings and are also available by appointment, Contact my assistant, at Wendy_Angus@gse.harvard.edu to schedule appointments. Occasional one-on-one check-ins with the instructor are strongly recommended.

CALENDAR

September 2016						
Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
28	29	30	31 – Class 1	1	2	3
4	5 – Holiday	6	7 – Class 2	8 – Asgn 1 Released	9	10
11	12 – Class 3	13	14 – Class 4	15	16 – Asgn 1 Due 5PM	17
18	19 – Class 5	20	21 – Class 6	22 – Asgn 2 Released	23	24
25	26 – Class 7	27	28 – Class 8	29	30 – Asgn 2 Due 5PM	1

October 2016						
Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
2	3 – Class 9	4	5 – Class 10	6 – Asgn 3 Released	7	8
9	10 – Holiday	11	12 – Class 11	13	14 – Asgn 3 Due 5PM	15
16	17 – Class 12	18	19– S-061A2 Begins… Class 13	20 – Take home final out: 12PM	21 – Take home final due: 5PM	22

READINGS

I require students to respond to online, Google Doc, discussion questions to central readings by 10PM the night before each class. A good response will demonstrate that the student has read and carefully considered the central reading and spent time considering what the discussion question is asking. Central readings that are very technical need not be mastered in detail, but do pay attention to notation and the underlying motivation of derivations. I intend noncentral readings as additional context and citations for future reference. Links will work on campus and, if you are off campus, if you have a VPN connection. Other readings are available via the iPa© tab on the Canvas website

There are two required textbooks that are available for purchase:

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. I recommend purchasing this via APA/AERA/NCME, where members can get a discount, and you'll also get an electronic text (very useful). The Coop is also an option.

- Koretz, D. (2009). *Measuring up: What educational testing really tells us*. Cambridge: Harvard University Press. See Amazon or the Coop.

August 31 – Class 1: Validation. SAT/MCAS example (read 1 and 3, skim 5, 7, 10, 11, glance at 2, 4, 6, 7, 8, 9)

1) Koretz (2008), Chapter 2, pp. 16-34. (required text)
2) Koretz (2008), Chapter 9, pp. 215-234. (required text)
3) AERA/APA/NCME Standards, Chapter 1, pp. 11-31. (required text)
4) Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. link
5) Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement*, *50*(1), 84-90. link
6) Ho, A. (2013). The epidemiology of modern test score use: Anticipating aggregation, adjustment, and equating. *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 64-67. link
7) Atkinson, R. C., & Geiser, S. (2015, May 5). The big problem with the new SAT. *The New York Times*, pp. 23. link
8) Geiser, S., & Studley, W. R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8*(1), 1-26. link
9) Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665-676. link
10) Linn, R. L. (2009). Comments on Atkinson and Geiser: Considerations for college admissions testing. *Educational Researcher*, *38*(9), 677-679. link
11) 2014 MCAS Technical Report Section 3.9: link

September 7 – Class 2: Reliability and Classical Test Theory… an MCAS example (read 1 and 5, skim 3, glance at 2 and 4)

1) AERA/APA/NCME Standards, Chapter 2, pp. 33-47. (required text)
2) Koretz (2008), Chapter 7, pp. 143-178. (required text)
3) Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37-45. link
4) Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group. Chapter 6: Reliability and the classical true score model pp. 105-113, 122-124; Chapter 7: Procedures for estimating reliability pp. 131-150. (iPa©)
5) 2014 MCAS Technical Report Section 3.7: link
6) Haertel, E. (2006). In R. Brennan (Ed.) *Educational measurement*. Westport, CT: Praeger. Chapter 3, 65-87. (iPa©)

September 12 – Class 3: Generalizability Theory (read 1, skim 3, glance at 2)

1) Shavelson, R.J. & Webb, N.W. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications. 1-24. (iPa©)
2) Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag. 1-19. (iPa©)
3) Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34. link

September 14 – Class 4: Generalizability Theory (read 2 skim 1)

1) Ho, A.D. & Kane, T.J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation. link
2) Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64. link

September 19 – Class 5: Scaling and Item Response Theory (read 2-5, skim 1)

1) Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.), 103*(2684), 677-680. link
2) AERA/APA/NCME Standards, Chapter 5, pp. 95-109. (required text)
3) Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35-41. link
4) 2014 MCAS Technical Report Section 3.6: link
5) Stata Manual (version 14), pp. 1-17. http://www.stata.com/manuals14/irt.pdf

September 21 – Class 6: Item Response Theory (read 1, skim 2)

1) Yen, W. M., & Fitzpatrick, A. R. (2006). In R. Brennan (Ed.), *Educational measurement*. Westport, CT: Praeger. pp. 111-118. Sections 1-2.4 (iPa©)

2) Ho, A. D., & Yu, C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement, 75*(3), 365-388. link

September 26 – Class 7: Item Response Theory – Polytomous Items. Linking (read 1 skim 2)

1) Stata Manual (version 14), pp. 34-80. http://www.stata.com/manuals14/irt.pdf
2) Holland, P. W., & Dorans, N. J. (2006). In R. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education, Praeger Publishers. pp. 197-201. (iPa©)

September 28 – Class 8: Bias, Differential Item Functioning, and Accommodations (read 1 and 2, skim 3 and 4)

1) AERA/APA/NCME Standards, Chapter 3, pp. 49-72. (required text)
2) Stata Manual (version 14), pp. 27-32. http://www.stata.com/manuals14/irt.pdf
3) Koretz (2008), Chapter 11, pp. 260-280. (required text)
4) Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English Language Learners: Implications for policy-based empirical research. *Review of Educational Research, 74*, 1-28. link

October 3 – Class 9: Standard Setting and Criterion-Referenced Reporting (read 3, skim 1 and 2)

1) Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4), 237-261. link
2) McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting. *Educational Researcher, 42*(2), 78-88. link
3) Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under no child left behind. *Educational Researcher*, *37*(6), 351-360. link

October 5 – Class 10: Growth and Vertical Scaling (read 1 and 3, skim 2 and 4)

1) Kolen, M. J. (2006). In R. Brennan (Ed.), Educational measurement (4th ed.). Westport, CT: Praeger Publishers. pp. 171-180. (iPa©)
2) Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers. link
3) Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*(2), 204-226. link
4) Pearl, J. (2014). Lord's paradox revisited - (Oh Lord! Kumbaya!). University of California. Technical Report R-36 (October). link

October 12 – Class 11: Test-Based Policy Metrics (read 1 and 2, skim 3)

1) AERA/APA/NCME Standards, Chapter 12 and 13, pp. 183-213. (required text)

2) Koretz (2008), Chapter 10, pp. 235-259. (required text)
3) Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice, 26*(4), 11-20. link

October 17 – Class 12: Psychometric Frontiers (read 1 and 3, skim 2)

1) Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35*(1), 5-25. link
2) Thissen, D. (2016). Bad questions: An essay involving Item Response Theory. *Journal of Educational and Behavioral Statistics, 41*, 81-89. link
3) Ho, A. D. (2016). The new (educational) statistics: Properties of scales that matter. *Journal of Educational and Behavioral Statistics, 41*, 94-99. link

Additional Optional Texts

All students should consider #1, below, for their reference library. Application-oriented students should consider #3 and #6 for Generalizability Theory and IRT, respectively. More technically oriented students should consider #4 and #5. Students interested in practical methods for large-scale testing should consider #2.

1) Educational Measurement, 4th Edition. ISBN: 978-0275981259

http://www.amazon.com/Educational-Measurement-American-Council-Education/dp/0275981258

2) Test Equating, Scaling, and Linking, 2nd Edition. ISBN: 978-1441923042

http://www.amazon.com/Test-Equating-Scaling-Linking-Statistics/dp/1441923047

3) Generalizability Theory: A Primer. ISBN: 978-0803937451

http://www.amazon.com/Generalizability-Theory-Measurement-Methods-Science/dp/0803937458/

4) Generalizability Theory. ISBN: 978-1441929389

http://www.amazon.com/Generalizability-Theory-Statistics-Behavioral-Sciences/dp/144192938X/

5) Fundamentals of Item Response Theory. ISBN: 978-0803936478

http://www.amazon.com/Fundamentals-Response-Measurement-Methods-Science/dp/0803936478/

6) Applications of Item Response Theory to Practical Testing Problems. ISBN: 978-0898590067

http://www.amazon.com/Applications-Response-Practical-Testing-Problems/dp/089859006X/

Statistical and psychometric computing

Statistical computing is an integral part of S-061. I will be using Stata this year, and you will require Stata 14 to use Item Response Theory methods. I assume that everyone is comfortable using a computer to perform basic statistical analysis, although I don't necessarily assume that you've used Stata.

I do not teach programming during class time, although code is threaded through the lecture slides. We provide resources to help you learn how to program on your own at your own pace. Masha may also cover coding issues in their sections.

There are two ways you can access Stata. The least expensive option is to use one of the networked workstations available in the Learning Technology Center (LTC) on the 3rd floor of Gutman Library and elsewhere on the HGSE campus (e.g., on the 2nd and 4th floors of Gutman Library). For students who would like to use Stata on their own PCs, you may purchase Stata following this link: http://www.stata.com/order/new/edu/gradplans/student-pricing/. Stata/IC, which will be sufficient for this course, is available for $75 for a 6-month license and $198 for a perpetual license.

Collaboration and study groups

Many people learn best when working in a group, and I encourage collaborative learning. To mimic statistical and psychometric work in the real world and to provide a chance for you to use this language actively, I mandate completion of assignments in pairs throughout the course, excepting only the final exam.

We mandate collaboration for at least three reasons. First, learning statistical and psychometric methods is like learning a language. To learn it, one must "speak" it actively and in a genuine context with other individuals. Second, collaborative quantitative analysis is the norm and individual work is the exception in the world of practice. Third, my experience has been that, on average, students who work in pairs and groups both perform better and enjoy themselves more than students who work individually. Statistical and psychometric collaboration is a case where the whole is greater than the sum of its parts.

Beyond pairs, study groups can be helpful to you as you prepare to do the assignments, both in terms of how to approach the work (including how to use the computer effectively) and in terms of how to think about important concepts. **However, students must turn in work as pairs or individuals where specified above, not group work. Papers should be written in the pair's own words—your text should reflect your own understanding of the material.**

Each group will undoubtedly develop its own structure; nevertheless, here are a few suggestions:

- Groups with six or more members become less useful and may be harder to organize because finding common meeting times becomes increasingly problematic.

- Plan at least one session of 1½ to 2 hours (early enough so that there is sufficient time if an additional session is necessary). After 2 hours of statistics, everyone's eyes will be glazing over.

- Schedule the meetings so that you have sufficient time afterwards to write in pairs or individually. When we read your assignments, we focus on what you say and how you say it. The assignments have been devised to require not only computation and programming skills, but skills in analyzing and reporting the material.

- Use the groups to ask questions, try out interpretations, and so on—you each represent each others' resources. Often one person can explain something that makes you see something in a new way—

or the other way around. Different people have different insights and strengths – some are good programmers, some ask good questions, others value contextual analysis—and you can learn from listening to what others in a group have to offer.

- **Be careful about sitting in groups at laptops or computers and simultaneously composing text.** You and your partner must write your own paper, on your own, using your own language. **Your papers should be written in your own words, not those of your study group.**

- Be sensitive to the distinction between collaboration to plan for and interpret the assignment and collaboration to write up the assignment. The former is encouraged; the latter is forbidden beyond, when applicable, your partner. If the distinction begins to feel murky, refocus your group's work on lecture content and course materials. And see me if you have any questions at all.

Accommodations

Students needing accommodations in instruction or testing must notify the instructor early in the semester, and HGSE's policies must be followed. Late requests for accommodations will not be honored unless there is a pressing reason, such as a recent injury.

A Note on Plagiarism

Please read the School's policy on plagiarism in the HGSE Student Handbook, which includes the statement, "Students who submit work either not their own or without clear attribution to the original source, for whatever reason, ordinarily will be dismissed from the Harvard Graduate School of Education." Attention to this policy is particularly important in a course like S-061, in which collaboration with other students is often required and generally encouraged. If you work closely with other students or partnerships—a process that I encourage and fully support—recognize the other students' contributions explicitly in your written account (a footnote is fine for this purpose). This helps avoid the natural questions that arise when similarities are detected at grading. **If you have any questions about what constitutes appropriate collaboration, or how to define what constitutes your own work, please see me or a Teaching Fellow.**

I cannot overemphasize the need for all students to monitor their own behavior. Assignments are structured such that you can receive feedback on your and your partner's understanding of the material. The consequences for plagiarism are appropriately severe.

Other Writing Resources

- HGSE Academic Writing Services: Gutman Library
- APA Online Tutorial: http://isites.harvard.edu/icb/icb.do?keyword=apa_exposed
- Writing Resources (including *Writing Like an Educator* Course and Reference Materials): http://isites.harvard.edu/icb/icb.do?keyword=awrs&pageid=icb.page48297
- Sign-up for Individual Sessions at the Writing Center: http://www.appointmentquest.com/provider/2030159020

Finally, some tips from last year's S-061 student cohort follow. These are the responses to my favorite question (#9) from our course evaluations. I have not edited or omitted any responses.

9. What advice would you give to students who are thinking of taking this course (about its level, the amount of work required, any prior training needed, ways to get the most out of the course, etc.)?

Ask questions often, and especially feel free to ask for clarifications. The material is sometimes quite dense, and just one explanation might not cut it. But an understanding of the content is invaluable.

This is a fantastic course, but be prepared to work hard and maintain a fast pace.

Make sure you have the time required!

This is a very natural continuation of S-052 (even though S-052 is not necessarily required) - expectations are about the same, the workload is manageable, the amount of readings is light (yet relevant), there is very little pressure on you

---

Budget a significant amount of time outside class to work through class lectures and examples, and homework

---

This is intensive course that requires much of background in statistics and material. Be prepared to invest your time in learning the material.

This is a great introductory psychometrics class that gives a high level overview of the major topics within the field of educational measurement. The instructor and the quality of delivered materials are outstanding. Professor Ho is a very skilled and experienced teacher. I especially liked the assignments and class discussions that built heavily upon the most recent and relevant practical issues in the field, as well as the key texts and articles by leading researchers instead of a "one size fits all" textbook. Some potential drawbacks of the class are its module structure and the coverage of a very large number of topics that provide a limited opportunity to dive deep into any specific issue. Additionally, the time investment and the workload is very high. One needs to spend at least 20 hours on a weekly basis and the number will increase during the weeks when assignments are due. I think it would be great to have such a class as a year long sequence to ensure deeper coverage of the content. For some topics, such as IRT, it would be very beneficial if HGSE considered to offer a separate class, which (without doubts) would have a high demand. The need for advanced and applied quantitative methods within HGSE is evident.

It's best to have a well developed idea about the project early on.
The first part of the course provides a nice time to think about different papers, and this module should not be spent thinking about what to do. it's a short 6 weeks.

Read as much as you can and be prepared for a tough and high value course

---

Start looking for a project topic early, and check in with the teaching team often about your ideas. They are fantastic at helping you to decide what's the best course of action given some data or some interest.

This course may help you get your paper prepared for publication.

This is a fantastic course for people interested in learning advanced psychometric statistical techniques such as IRT and G-theory. Expectations from the instructor are appropriately high, and there is a lot of work required for the course, but it is well worth it.

Come with a good paper idea and with your own data-set.
