

S-061A1: Statistical and Psychometric Methods for Educational Measurement
Harvard Graduate School of Education
September 2 – October 23, 2015

Class meets Mondays and Wednesdays, 10:10-12:00 in Longfellow 229

Instructor: Andrew Ho
455 Gutman Library
Andrew_Ho@gse.harvard.edu
617-496-2408

TF: John Hansen
jdh488@mail.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%), active participation in class and in out-of-class 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.

Support

Our TF, John Hansen, will hold weekly office hours by appointment. Occasionally, John may also offer optional discussion sections as he 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 2015						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
30	31	1	2 – Class 1	3	4	5
6	7 – Holiday	8	9 – Class 2	10 – Asgn 1 Released	11	12
13	14 – Class 3	15	16 – Class 4	17	18 – Asgn 1 Due 5PM	19
20	21 – Class 5	22	23 – Class 6	24 – Asgn 2 Released	25	26
27	28 – Class 7	29	30 – Class 8	1	2 – Asgn 2 Due 5PM	3

October 2015						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
4	5 – Class 9	6	7 – Class 10	8 – Asgn 3 Released	9	10
11	12 – Holiday	13	14 – Class 11	15	16 – Asgn 3 Due 5PM	17
18	19 – Class 12	20	21– S-061A2 Begins... Class 1	22 – Take home final out: 12PM	23 – Take home final due: 5PM	24

READINGS

All readings must at least be opened and glanced over. I intend noncentral readings as additional context and citations for future reference. Other readings are central and should be read closely. 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 will be attempting to provide discussion questions for central readings in advance. Links will work on campus and, if you are off campus, if you have a VPN connection.

There are two required textbooks that are available for purchase at the Coop:

- 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.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge: Harvard University Press.

Readings from these textbooks will be indicated in the syllabus as Required Texts (RT).

September 2 – Class 1: Validation. SAT/MCAS example (1, 3, 5, 7, 10, 11 central)

- 1) Koretz (2008), Chapter 2, pp. 16-34. (RT)
- 2) Koretz (2008), Chapter 9, pp. 215-234. (RT)
- 3) AERA/APA/NCME Standards, Chapter 1, pp. 11-31. (RT)
- 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) MCAS Tech Report Section 3.9:
<http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2013/2013%20MCAS%20and%20MCAS-Alt%20Technical%20Report.pdf>

September 9 – Class 2: Reliability and Classical Test Theory... an MCAS example (1, 3, 5 central)

- 1) AERA/APA/NCME Standards, Chapter 2, pp. 33-47. (RT)
- 2) Koretz (2008), Chapter 7, pp. 143-178. (RT)
- 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) MCAS Tech Report Section 3.7:
<http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2013/2013%20MCAS%20and%20MCAS-Alt%20Technical%20Report.pdf>
- 6) Haertel, E. (2006). In R. Brennan (Ed.) *Educational measurement*. Westport, CT: Praeger. Chapter 3, 65-87. (iPa©)

September 14 – Class 3: Generalizability Theory (1, 3 central)

- 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 16 – Class 4: Generalizability Theory (2 central)

- 1) Ho, A.D. & Kane, T.J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.
http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf
- 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 21 – Class 5: Scaling and Item Response Theory (2, 3, 4 5 central)

- 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. (RT)
- 3) Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, 8(1), 35-41. [link](#)
- 4) MCAS Tech Report Section 3.6:
<http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2013/2013%20MCAS%20and%20MCAS-Alt%20Technical%20Report.pdf>

- 5) Stata Manual (version 14), pp. 1-17. <http://www.stata.com/manuals14/irt.pdf>

September 23 – Class 6: Item Response Theory (1 central)

- 1) Yen, W. M., & Fitzpatrick, A. R. (2006). In R. Brennan (Ed.), *Educational measurement*. Westport, CT: Praeger. pp. 111-118. Sections 1-2.4, pp. 111-118. (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 28 – Class 7: Item Response Theory – Polytomous Items. Linking (1 central)

- 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 30 – Class 8: Bias and Differential Item Functioning (2 central)

- 1) AERA/APA/NCME Standards, Chapter 3, pp. 49-72. (RT)
- 2) Stata Manual (version 14), pp. 27-32. <http://www.stata.com/manuals14/irt.pdf>
- 3) Koretz (2008), Chapter 11, pp. 260-280. (RT)

October 5 – Class 9: Standard Setting and Criterion-Referenced Reporting (2, 3 central)

- 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 7 – Class 10: Growth and Vertical Scaling (1, 3 central)

- 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. <http://www.ccsso.org/Documents/2013GrowthModels.pdf>
- 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). http://ftp.cs.ucla.edu/pub/stat_ser/r436.pdf

October 14 – Class 11: Test-Based Policy Metrics (1, 2 central)

- 1) AERA/APA/NCME Standards, Chapter 12 and 13, pp. 183-213. (RT)
- 2) Koretz (2008), Chapter 10, pp. 235-259. (RT)
- 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 19 – Class 12: Psychometric Frontiers (1 central)

- 1) Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, 35(1), 5-25. [link](#)
- 2) Thissen (in press, TBD)
- 3) Ho (in press, TBD)

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. John 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>