Measuring Knowledge: Two Modern Statistical Tools for Better Discerning What Data on Student Performance Can (and Cannot) Tell Us

Guada Lozano, Spring 2015

Mathematical performance data has been leveraged to measure what students and teachers “know” for quite some time. The SAT and GRE tests, for example, as well as the more recent MKT items (for K-12 teachers) are well-known sources of mathematical “knowledge” data that has been extensively analyzed.

In the last decade or two, theories about what knowing mathematics really is, and about how mathematics know-how can be learned, have motivated the creation of specialized mathematical knowledge measures called concept inventories. These instruments, designed to measure discipline-specific conceptual knowledge, exist in a variety of STEM subjects, including differential calculus. Variation in performance data on concept inventories across universities has led to a well-known claim: that concept inventory gains are positively correlated with specific types of active learning instruction. However, finding proper evidence of this claim is difficult. It requires not only reliable methods for measuring instruction, but also statistical tools that are properly sensitive to the information we hope to discern.

In this RTG we will very briefly introduce two different methods from modern statistics (HLM and IRT, see below) and explore each of their affordances for studying the connection between mathematical knowledge of calculus and instruction, as an example. Our exploration will be based on undergraduate performance data on the Calculus Concept Inventory collected at two large research universities. As much as possible, I will compare and contrast HLM and IRT with well-known but less sensitive methods from classical test theory, many of which are still used today in a fair number of fields.

Hierarchical Linear Models (HLM) are statistical modeling tools sensitive to clustered data (eg: teachers within schools, students within classrooms, etc.) and enable us to distinguish scores that may seem identical if one is blind to the underlying cluster structure. HLM may hence enable us to detect and quantify cluster effects (eg: exam performance effects that are due to differences between instructors rather than due to differences among students).

Item Response theory (IRT) is a measurement framework that enables the determination of item parameters (eg: how difficult each test item is) independently of the ability of the respondent sample, for example. As a result, IRT is a useful tool for detecting “item bias,” the tendency of one or more test items to “favor” specific student populations, due to characteristics likely unrelated to mathematical ability (eg: men vs. women, American vs. non-American students). Our RTG exploration of IRT will be motivated by my intent to measure item bias to pin-point the impact of instruction on student performance.