Abstract Title Page
Not included in page count.

Title: Assessing the conditional reliability of state assessments.[*]

Author(s): Henry May (UPenn), Russell Cole (MPR), Josh Haimson (MPR), and Irma Perez-Johnson (MPR)

[*] The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education, through the Analytical and Technical Support for Advancing Education Evaluations contract. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education

Background/context:

There has been growing interest in the use of state achievement tests as outcomes in educational evaluation research. When conducting education evaluations where the goal of the intervention is to improve student achievement, researchers can either administer a study specific assessment as an outcome indicator, or can take advantage of available administrative data, if appropriate achievement data exist. In a review of National Center for Education Evaluation (NCEE) funded evaluations from the past 5 years, May et al. (2009) identified 58 ongoing or completed studies, of which 21 had planned to or had used use state assessments as a source of outcome measures. There are a number of reasons why education researchers are increasingly turning to state assessments as outcome measures, rather than administering their own study specific tests (see May et al., 2009, for a more thorough treatment of the subject).

Given the requirements of the No Child Left Behind (NCLB) Act of 2001, nearly all states have begun testing students in grades three through eight in Mathematics and English Language Arts (ELA). This mandate has created an opportunity for education researchers to utilize available state assessment data as outcome measures rather than administer a study-specific achievement test for their intervention studies. States and districts have developed electronic databases containing student demographic and achievement data, and these data can be obtained at a much lower cost than administering a separate study-specific test.

In addition to the cost-benefit of using state assessment data, the quality of the information obtained in these high-stakes assessments may be higher than what would be possible when study-specific assessments are used. The adequate yearly progress (AYP) requirements stated in NCLB provides school staff with strong incentives to encourage student attendance during test administration and serious test taking, which ultimately create high quality indicators of student academic achievement.

Finally, given the increased prevalence of state assessments, it has been suggested that taking and administering additional study-specific assessments are an unnecessary burden for students and teachers. There is an opportunity cost when students in intervention studies are pulled out from their regular classroom instruction to take a study-administered examination.

We have summarized several reasons why state assessments are potentially desirable outcomes for education research studies, taking the perspective that these tests are available, low cost, and minimally intrusive. However, unless the content of the assessment well aligned with the intervention being studied, these tests may not be appropriate outcome measures. When the content of state tests aligns with the content of the education intervention, these state assessments can be considered valid outcomes.

While having valid outcome measures is a necessary condition for studies of education interventions, it is also necessary that the outcome indicator is reliable. An outcome can be considered reliable if repeated measurements (on the same individual, in the same situation) produce similar results on the instrument. In statistics, reliability of a test is often defined by the amount of measurement error in the outcome relative to the total variance in the outcome, where lower measurement error implies higher reliability. According to true-score theory, the reliability of a test is a constant, defined by these two terms (Lord and Novick, 1968).

Most state tests are developed and scored using an Item Response Theory (IRT) framework, rather than from a true score perspective, with 40-50 multiple choice items per subject test (Webb, 2007). While multiple choice tests have been criticized for their inability to measure higher order skills, they are lauded for their reliability (Darling-Hammond 2007; Bracey 2002; Kohn 2000). In IRT scaled assessments, the reliability of a test is not assumed to be a

constant. Rather, in IRT scaled assessments, a conditional (or local) reliability can be calculated as the reliability of the test at a given ability level (Wasserman & Bracken, 2003). In these IRT based state tests, the measurement error of the test increases as average student ability moves away from the mean towards the tails of the distribution. Test reliability is inversely related to measurement error, and thus, the conditional reliability of IRT tests also decreases as the average ability of samples of test-taking students becomes more disparate from the population mean.

The attenuation in the conditional reliability of state assessments as student ability moves towards the tails of the distribution has research implications for studies that focus on low and high performing samples of students. In these samples, where the measurement error of state assessments may be relatively large, intervention researchers must consider the implications of reduced reliability as it pertains to prospective power analysis and the precision of impact estimates. While this concept of attenuated reliability in IRT based tests is not new, researchers at Mathematica have a unique opportunity to quantify the magnitude of this issue using state assessment data from 5 states (and we are in the process of obtaining permission from a 6^{th} state and a large school district). This will provide opportunity to compare different state tests with different populations and to provide researchers with empirical benchmarks against which to ground their own prospective studies.

The actual calculation of conditional reliability from an IRT perspective requires knowledge of the item parameters and the item response patterns, which is an unrealistic expectation for researchers using administrative achievement data (du Toit, 2003). An alternate measure of conditional reliability could be the commonly used Cronbach's α (1951). However, this measure requires knowledge of individual item responses, which again are typically unavailable from large scale state assessment databases. Pretest-posttest correlations are an indication of the consistency of relative ranks of students in a sample, and can be considered a signal of the conditional reliability of a test for the given sample. This study uses the pretest-posttest correlations of assessments for samples defined by pretest achievement level as its indicator of conditional reliability, which can be calculated using available state assessment data.

Purpose / objective / research question / focus of study:

The purpose of this study is to provide empirical benchmarks of the conditional reliabilities of state tests for samples of the student population defined by ability level. Given that many educational interventions are targeted for samples of low performing students, schools, or districts, the primary goal of this research is to determine how severe the attenuation in conditional reliability is, depending on the test, subject matter, and average ability level of the students in the research sample. This objective is addressed, along with implications for future research, in research questions 1 and 2 below:

1. What are the conditional reliabilities (defined as the median and 95% CI for pretest-posttest correlations) of given state tests for subsets of the population defined by achievement level?
2. What are the implications for power analysis when state assessments are the outcomes of interest for samples of students taken from the tails of the achievement distribution?

The results of research question 1 will be used to explore the implications of conditional reliability in the prospective power analysis for a study using state assessments as outcomes.

Setting:

The student-level data for this study will be taken from the National Research Study on Charter Management Organization Effectiveness, a study funded by the Bill and Melinda Gates

foundation. Mathematica serves as a research partner on this study which spans 12 states. The purpose of the Charter Management Organization (CMO) study is to estimate the impact of nonprofit CMOs on student outcomes, including student achievement.

For the purposes of the current study, we have secured longitudinal data for students in recent school years from four states (Texas, Arkansas, Louisiana, and Indiana), and are awaiting approval to use data from Arizona and the Chicago School district. The goal of this methodological study, funded by IES, is to calculate the conditional reliability scores for samples of the ability distribution of the populations defined by the students in these states. Given that administrative data are to be used for this study, and that the conditional reliabilities will be calculated using pretest and posttest data, student records from grades 3-8 will be used in these analyses. These are the grades that are commonly assessed for AYP purposes, and the adjacent grades being assessed will provide an opportunity for pretest-posttest correlation analyses.[†]

At the time of this proposal, these data sources are in the process of being organized and cleaned, and are not yet available for analysis. To provide some results for this structured abstract, simulated data were used for the empirical analysis. The data generation procedure for the population of simulated students is described in the following section. For the actual paper, we anticipate being able to present the results of the study using the actual data … we may include these simulated results to provide additional information and context.

Population / Participants / Subjects:

A population of 200,000 students was simulated for this purpose, and student ability levels ($\theta_i \sim N(0,1)$) were generated for each student in the population. To generate assessment data according to an IRT perspective, a one-parameter (Rasch) model was employed. A 50 item pretest and a 50 item posttest were created by drawing item difficulties (b_j) from a random normal distribution with mean 0 and standard deviation 1. If the item difficulty parameters were greater than 2 standard deviations from the mean, the difficulty parameter was re-drawn.

For each item on the test, the probability (p_{ij}) of student *i* answering item *j* correctly was calculated by equation (1) below:

$$(1)\ p_{ij} = P(X_{ij} = 1 \mid \theta_j) = \frac{1}{1 - e^{(\theta_j - b_i)}}$$

Then, the response for each item (X_{ij}) was drawn from a Bernoulli distribution, with probability p_{ij}, ($X_{ij} = Bern(p_{ij})$).

Once all of the item responses were generated, we used ML to estimate a $\hat{\theta}_j$ for each student on both the pretest and posttest, given the student's item responses (X_{ij}) and the known item difficulties (b_j). These $\hat{\theta}_j$ were then rescaled to have a mean of 500 and a standard deviation of 100 for both the pretest and the posttest. In this large population, there were some students who answered every question incorrectly, and some who answered every question

[†] The NCLB requirement for high school assessment is a single test (for both mathematics and ELA) sometime during 10th-12th grade. As a result, a comparison of pretest-posttest correlations is not viable with administrative high school assessment data.

correctly on a given examination. The MLE $\hat{\theta}_j$ for these extreme scores were pulled in, and the ultimate floor and ceiling for all scores ranged between 200 and 800.

Scatterplots of the relationship between posttest and pretest observed scores for a random sample of 10,000 observations from the population are presented as Figures 1 and 2 (random noise was added to the observed scores to Figure 2 to help illustrate the overall density of the relationship). Of note is the heteroscedastic relationship between posttest and pretest, which would not be observed if the test data were simulated using a true score framework.

< Insert Figures 1 and 2 here >

Data Collection and Analysis:

In the final paper, the following series of analyses will be performed for each available assessment in a given year, for a particular state. For the purposes of the current study, these analyses were only performed for the simulated population, and the single pretest and posttest. The primary goal of this study was to determine the conditional reliability of state assessments for samples of the population defined by student ability level. In an effort to guard ourselves against sampling error when drawing samples for each ability subgroup, a bootstrap sampling procedure was performed to define an empirical distribution for conditional reliability for ability level subgroups. Our methods are elaborated below:

Bootstrap Sampling: To determine conditional reliability based on pretest achievement level, samples of students are randomly selected from a population for a variety of desired pretest achievement levels. One thousand bootstrap sample simulations containing approximately 500 students per sample are selected for a given ability level. For this sampling procedure, the probability of a student being selected into a sample is maximized when the student's pretest achievement score aligns with the desired pretest ability level (ability levels will range from two standard deviations below the mean to two standard deviations above the mean, in .25 standard deviation unit intervals). As a result, these seventeen mean pretest ability levels serve as the reference points of interest for the empirical comparisons of conditional reliability.

Conditional Reliability: A pretest-posttest correlation is calculated for each bootstrap sample at each ability level. The distribution of correlations at a given ability level are presented as a series of boxplots (one for each of the seventeen ability levels from -2 to +2 SDs), along with a tabular presentation of the empirical median correlations and 95% confidence intervals resulting from the repeated simulations.

In the current study, this analysis is only performed for the simulated population, however, for the main paper, this analysis will be replicated across all available assessment data in all available states and large districts to illustrate the variability in pretest-posttest correlations for various ability levels.

Findings / Results:

When calculations were performed for the full population of 200,000 students, the pretest-posttest correlation of observed scores was estimated to be .94 (and the correlation between true ability and both the pretest and posttest was .97). The average conditional reliability for each test was .87, with maximum conditional reliability of both tests equal to .91 when $\theta = 0$. This indicates that in this population, the tests exhibit high levels of overall reliability and maximal reliability at the center of the ability distribution.

Figure 3 shows the variability in pretest-posttest correlations for the 1000 bootstrap samples of students, when sample pretest average varied from 2 standard deviations below the

mean to 2 standard deviations above the mean. As expected, the pretest-posttest correlation coefficients are markedly lower in the tails of the distribution than they are in the center of the distribution, indicating that conditional reliability is maximized at the mean of the distribution. In addition, the boxplots indicate that there is greater variability in conditional reliability when samples are selected in the tails of the population distribution, as evidenced by the increasing inter-quartile range as samples are selected away from the mean. These results are also presented as point estimates and empirical confidence intervals in Table 1. The median conditional reliability coefficients (the pretest-posttest correlations) range from a low of .61 in the lower tail of the distribution to a high of .87 in the center of the distribution.

< Insert Figure 3 and Table 1 here >

We have performed similar analyses with preliminary versions of the actual state population data, and we have seen that the conditional reliability of some state tests exhibit a similar upside down "U" shape seen in Figure 3. However, our preliminary findings also suggest that tests (and subjects) differ in the extent to which conditional reliability varies in samples of the ability distribution, where some tests only exhibit reliability attenuation in one tail of the distribution. We do not present any empirical results from these tests here, as our results are based on incomplete data, and will change as data are finalized.

Conclusions:

IRT based state-assessments are an attractive and low cost outcome measure for education interventions. The current study demonstrates that the reliability of these assessments may be compromised when low and/or high ability segments of the student population are to be sampled for the intervention. The simulation results posed above indicate that if a study were to be performed on a sample of low performing students (students 2 standard deviations below the mean), the use of a state test would be far less appropriate than if the intervention was focused on students whose ability levels were closer to the average ability level of the population as a whole.

These findings have important implications for prospective power analysis and the precision of impact estimates for randomized trial designs. Consider a balanced randomized trial were to be performed on a sample of 500 students, where state assessment data were selected as the desired outcome and a pretest were to be administered to improve power to detect effects. In this design, the minimum detectable effect size (assuming $\alpha = .025, \beta = .20$) would be .20 for the low performing students, versus just .12 for the average performing students (based on the formulae in Bloom, 2006, where the squared conditional reliability estimates indicate the proportion of variance explained by the pretest). If only a small impact (ES < .20) from the intervention was expected, the researchers may be advised against using the state assessment data in this example.

The final version of this paper will provide a table of conditional reliability estimates for the state assessments (stripped of any identifying information, for confidentiality purposes for states and test developers), which can be used as a set of benchmarks for future power calculations. This technical appendix of correlations will be a contribution for researchers interested in selecting appropriate outcomes for their ability defined sample, and can help to provide guidance for prospective power analyses.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bloom, H.S. (2006). The Core Analytics of Randomized Experiments for Social Research. MDRC working paper.

Bracey, G.W. Put to the Test: An Educator's and Consumer's Guide to Standardized Testing (second edition) (2003). Bloomington, IN: Phi Delta Kappa International.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3),* 297-334.

Darling-Hammond, L. (2007). "Testimony Before the House Education and Labor Committee on the Re-Authorization of No Child Left Behind." Washington, DC, September 10, 2007.

Du Toit, M., (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT., Lincolnwood, IL.: Scientific Software International, Inc.,

Hambleton, R.K., H. Swaminathan, and H. J. Rogers. (1991) *Fundamentals of Item Response Theory*. Newbury park, CA: Sage Press.

Kohn, A. (2000). The Case Against Standardized Testing: Raising the Scores, Ruining the Schools. Portsmouth, NH: Heinemann.

Lord, F.M. and M.R. Novick (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Welsley Publishing Company.

May, H., I. Perez-Johnson, J. Haimson, S. Sattar, and P. Gleason. (2009). "Making Good Use of State Tests in Education Experiments." Draft report. Princeton, NJ: Mathematica Policy Research, Inc.

Wasserman, J.D. & Bracken, B.A. (2003). Psychometric Characteristics of Assessment Procedures. In "Handbook of Psychology: Volume 10 – Assessment Psychology", J.R. Graham and J.A. Naglieri, Eds., pp 43-66.

Webb, N.L. (2007). Issues Related to Judging the Alignment of Curriculum Standards and Assessments. Applied Measurement in Education, vol. 20, no. 1, pp. 7-25.

Appendix B. Tables and Figures

Not included in page count.

Figure 1: Scatterplot of a random sample of 10,000 Posttest (Y) and Pretest (X) test scores

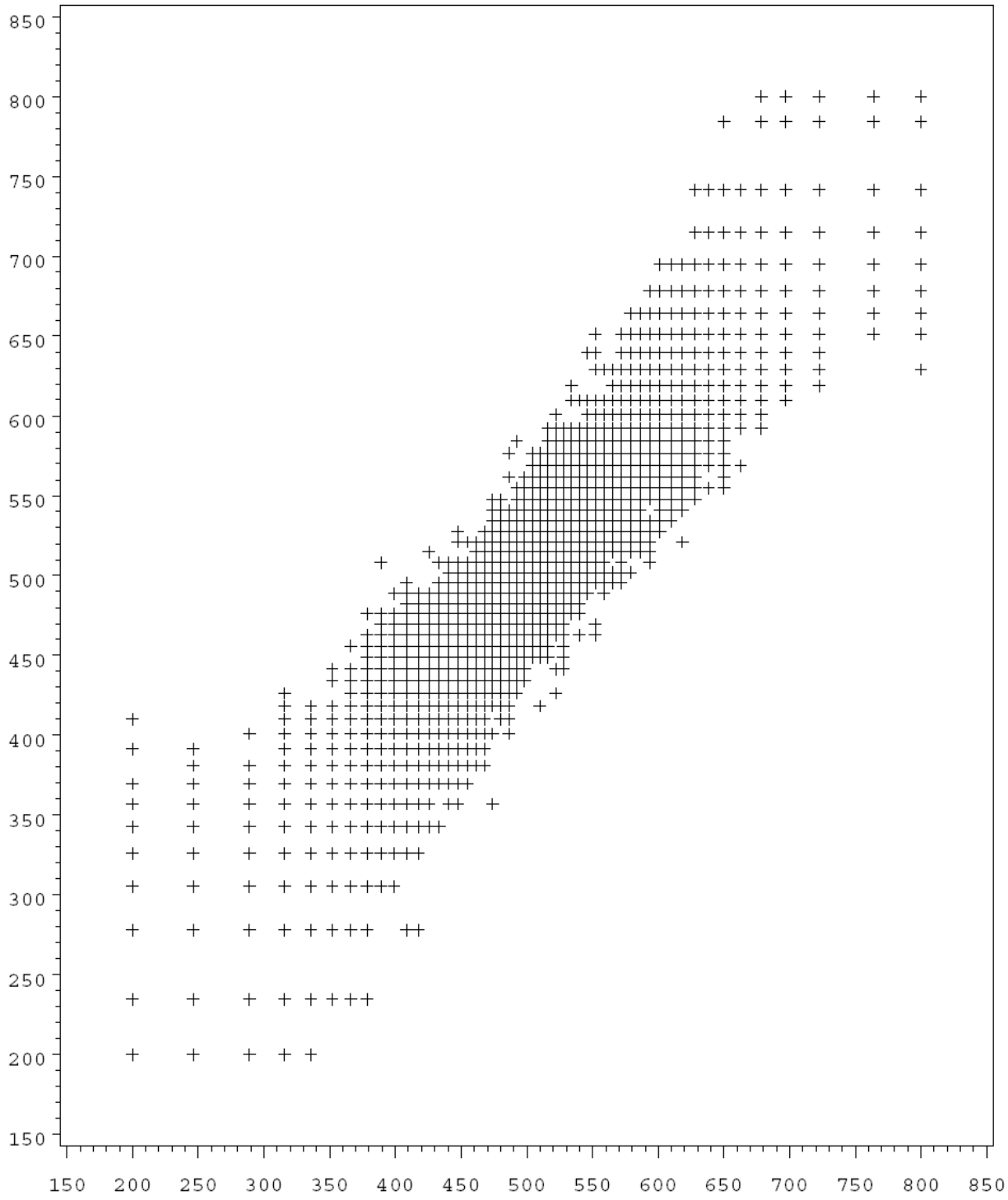


Figure 2: Scatterplot of a random sample of 10,000 Posttest (Y) and Pretest (X) test scores (observed scores jiggled)

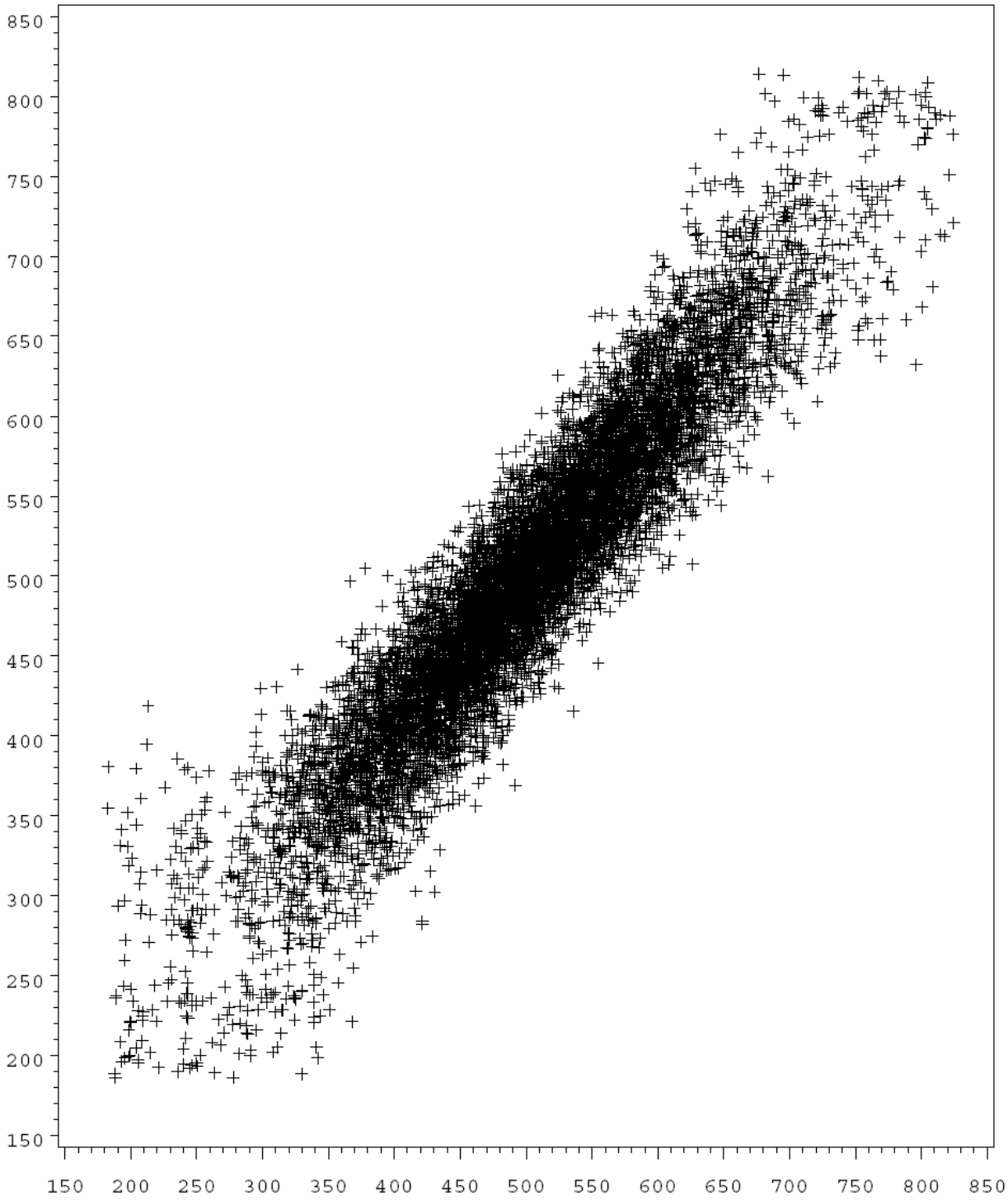
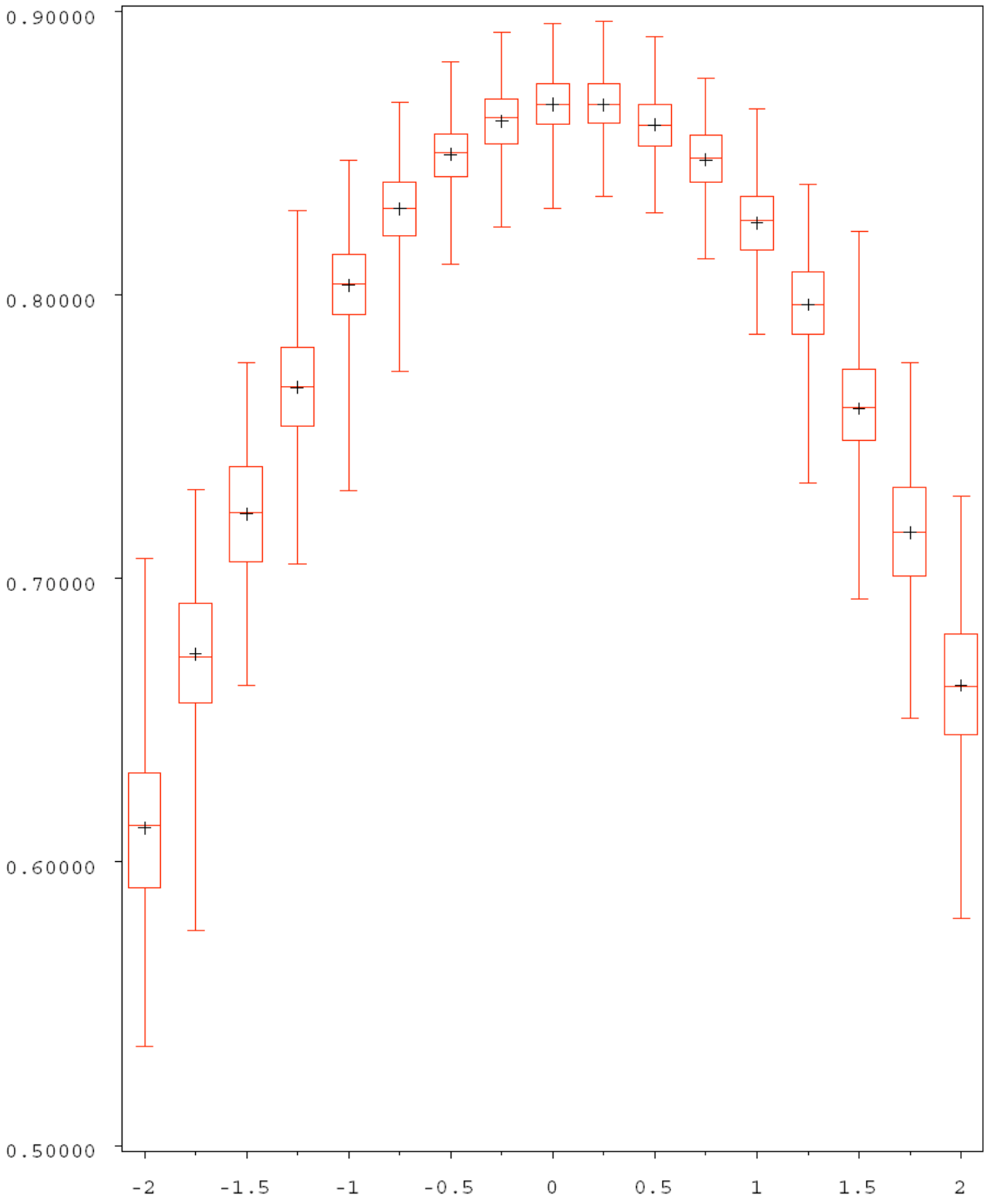


Figure 3: Boxplots of pretest-posttest correlations resulting from 1000 bootstrap samples of ~500 students for each average ability level.



Note: The X-axis represents the sample average of the pretest, in z-score units.

Table 1: Median pretest-posttest correlation coefficient (conditional reliability estimates) and empirical 95% confidence intervals based on 1000 bootstrap samples of 500 students.

Pretest Mean Achievement Level (in z-score units)	Median Correlation	Lower 95% CL	Upper 95% CL
-2.00	0.61	0.55	0.67
-1.75	0.67	0.63	0.72
-1.50	0.72	0.68	0.76
-1.25	0.77	0.73	0.80
-1.00	0.80	0.77	0.83
-0.75	0.83	0.80	0.86
-0.50	0.85	0.83	0.87
-0.25	0.86	0.84	0.88
0.00	0.87	0.85	0.89
0.25	0.87	0.85	0.89
0.50	0.86	0.84	0.88
0.75	0.85	0.82	0.87
1.00	0.83	0.80	0.85
1.25	0.80	0.76	0.83
1.50	0.76	0.72	0.79
1.75	0.72	0.67	0.76
2.00	0.66	0.61	0.71