A Question of Numeracy: Is Self-Assessed Competency Registered on Knowledge Surveys Meaningful?

Dr. Edward Nuhfer
Professor of Geology
Director of Faculty Development
Director of Educational Effectiveness (retired)
enuhfer@earthlink.net, Niwot, CO

Dr. Karl Wirth
Associate Professor of Geology
Macalester College
St Paul, MN 55105
wirth@macalester.edu

Dr. Steven Fleisher
Instructional Faculty,Psychology
California State University Channel Islands, Camarillo CA
93012
steven.fleisher@csuci.edu

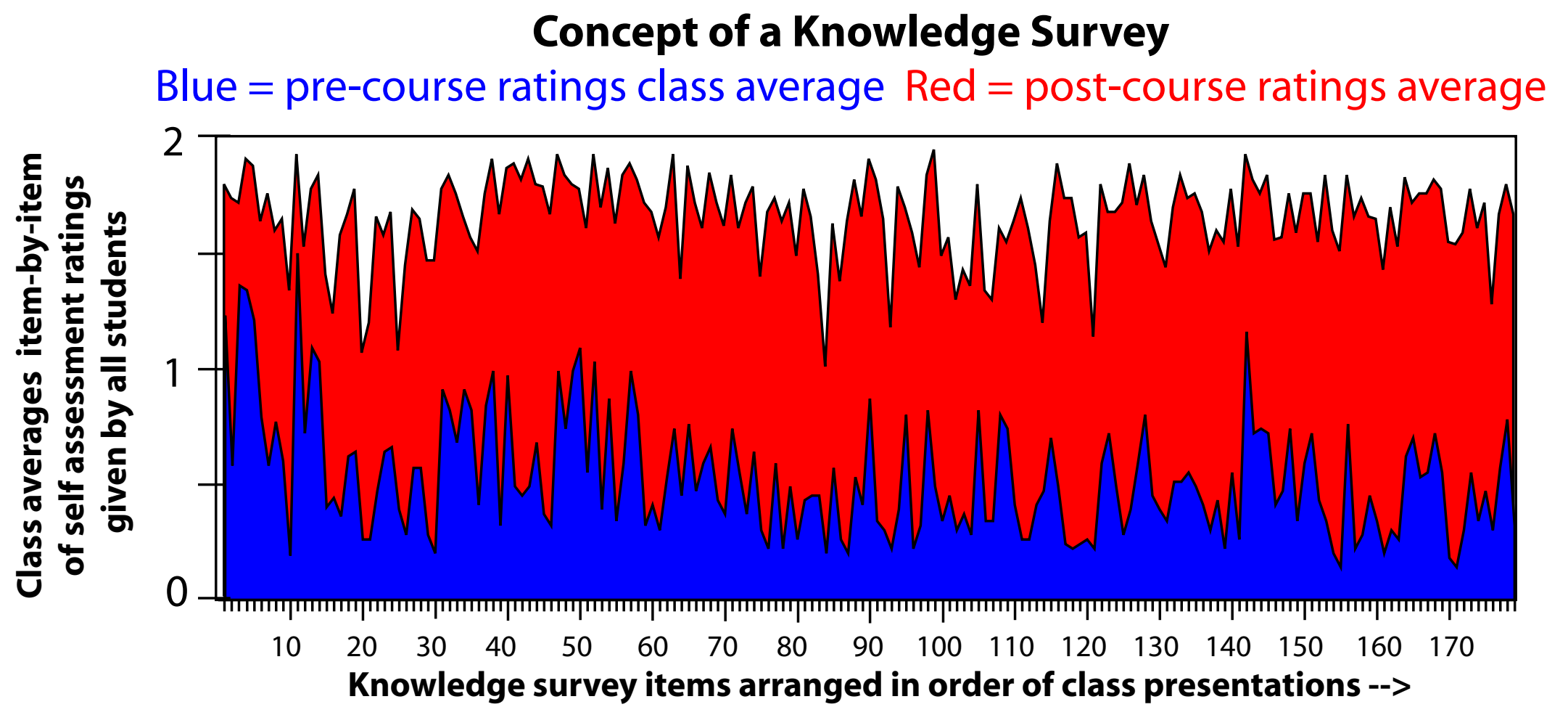
Dr. Christopher B. Cogan
Independent Consultant in Environmental Science and GIS,
Camarillo CA 93012
cbcmapper@gmail.com

Dr. Eric Gaze
Director of the Quantitative Reasoning Program, and
Lecturer in Mathematics, Bowdoin College
Bowdoin, ME 04011
egaze@bowdoin.edu

ABSTRACT

Geoscientists often use knowledge surveys to collect self-assessed competency data about learning and learning gains. If people believe that they can do something, how well can they actually do it? At first glance, quantifying the accuracy of a person's self-assessment of competency appears simple. It involves comparing direct measures of self-assessed confidence taken by one instrument, such as a knowledge survey, with direct measures of competence taken by another instrument, usually a test. In accurate self-assessments, the scores on both measures would be about equal. Disparities from this perfect score would register as measures of either over-confidence or under-confidence. However, deducing self-assessment accuracy is not simple. Both instruments used to obtain paired measures must have sufficient reliability to permit good comparisons, and both must measure the same learning construct. Competence and confidence have no established units, so the default measures are scores reported in percents. These constitute arrays bounded by 0% and 100%, a fact that introduces complications. Sorting of data needed to report results in aggregate imparts bias, and the probability of overestimating or underestimating is not uniform across all participants. To deduce this, we employed reliable, tightly aligned instruments to measure self-assessed competency (a knowledge survey of the Science Literacy Concept Inventory) and actual competency (Science Literacy Concept Inventory) of 1154 participants in understanding the nature of science. We used random number simulations to discover how mathematical artifacts can be (and have been in published literature) mistaken as human measures of self-assessed competencies. Innumeracy leads to misinterpretations so severe as to contradict what the data actually reveals. In our study, knowledge survey self-assessments of competence proved strongly related to actual performances.

What are Knowledge Surveys?



Self assessment scores given item-by-item by each student:

2 = I have current ability to address this challenge very well.
1 = I have partial knowledge/skill and can now only partially address the challenge.
0 = I currently have insufficient skill/knowledge to address the challenge.

Figure 1. Concept of a knowledge survey consisting of 179 items from an introductory course.

Knowledge surveys (Nuhfer and Knipp, 2003; see http://www.merlot.org/merlot/viewMaterial.htm?id=437918) query individuals to self-assess by rating their present ability to meet the challenge expressed in each item by responses on a three-point multiple-choice self-assessment. Simple three-item choice formats appear psychometrically sound (Landrum, Cashin, and Theis, 1993; Rodriguez, 2005; Baghaei and Amrahi, 2011) and expedite quick, clear distinctions.

To produce this poster, we employed a database generated from 1154 participants (undergraduates, graduate students, and professors) who completed a test of competency (the 25-item multiple choice Science Literacy Concept Inventory or SLCI) after performing a self-assessment of competency in the form of a knowledge survey based on the identical 25 SLCI items (KSSLCI). The SLCI has a Cronbach Coefficient Alpha reliability of R = 0.84; the KSSLCI has reliability of R = 0.93.

Noise and Signal

We consider human self-assessment measures as blends of the self-assessment signal that researchers seek to measure and random variation or "noise" that always accompanies the signal.

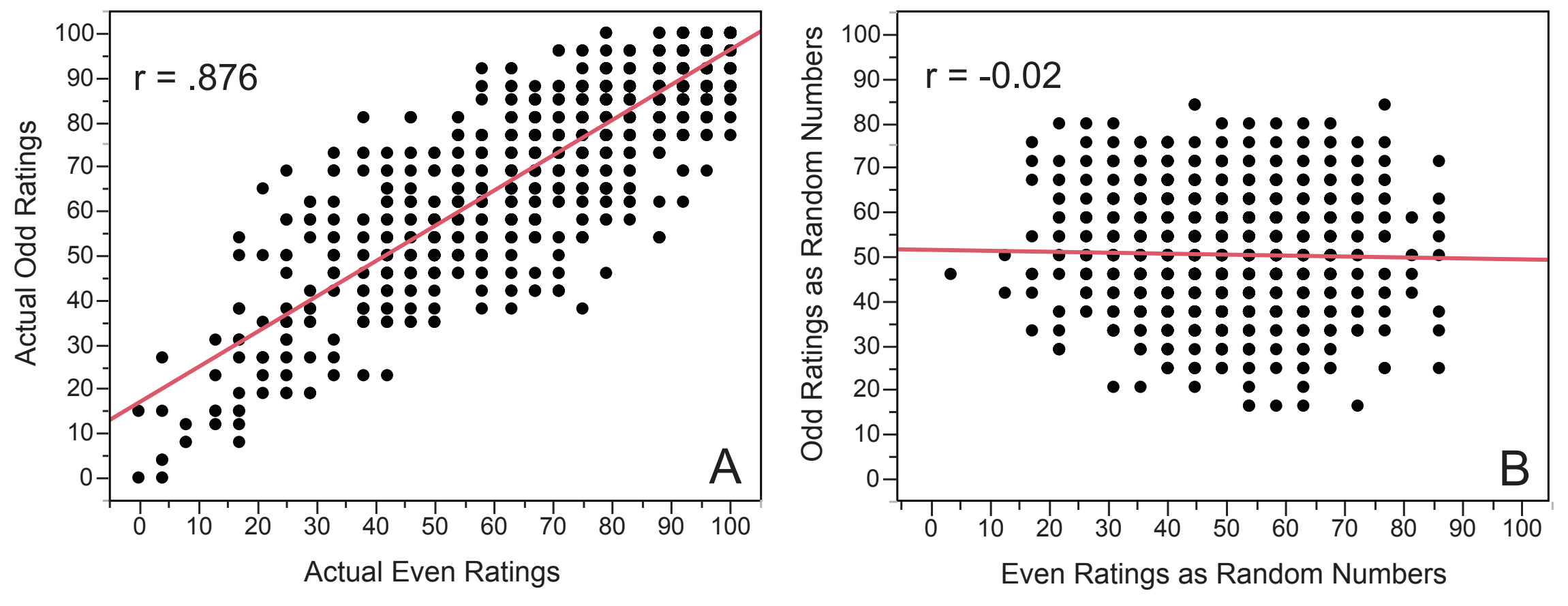


Figure 2. (A) Actual data from our 25-item KSSLCI, and (B) depicts the same data simulated by random numbers. Its generally circular pattern depicts randomness with no trend that differentiates individuals by high or low self-assessment confidence. Self-assessments from knowledge surveys depict a dominantly valid signal with some noise, but are distinct from purely random noise.

Map of Self-Assessed Competency

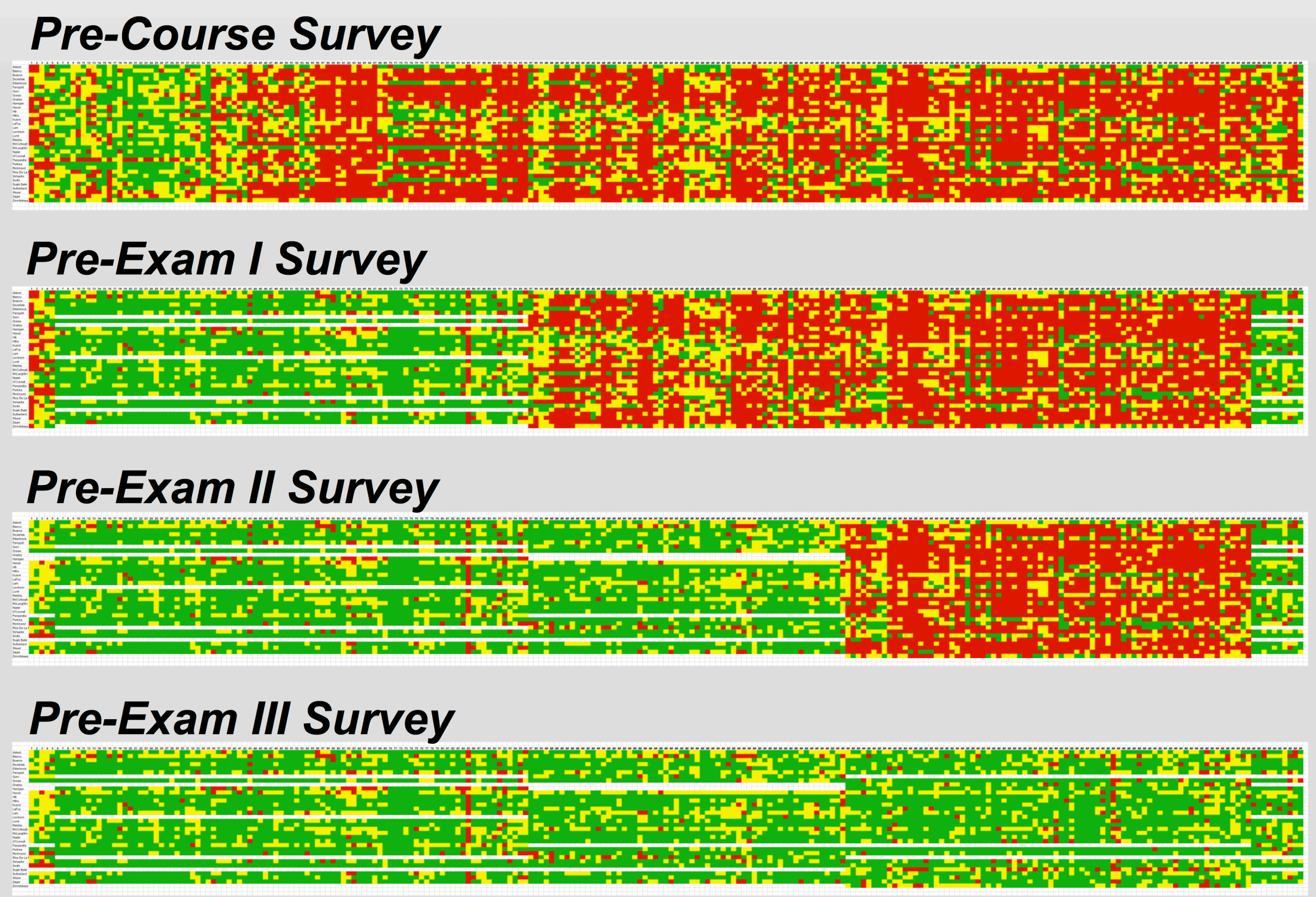


Figure 3. Color coded knowledge survey results from across a single semester. For each administration of the knowledge survey, each colored cell represents an individual students' (rows) self-assessed competency in response to an expressed challenge (columns). Colors: Red: self-report of inability to address the challenge; Yellow: self-report of partial knowledge; and Green: self-report of confidence to address a challenge when tested. Knowledge surveys reflect progressive change from red to green as knowledge surveys are taken periodically through any well-taught course. This change is not explained by random chance.

How well does a test of known reliability correlate with a knowledge survey of known reliability?

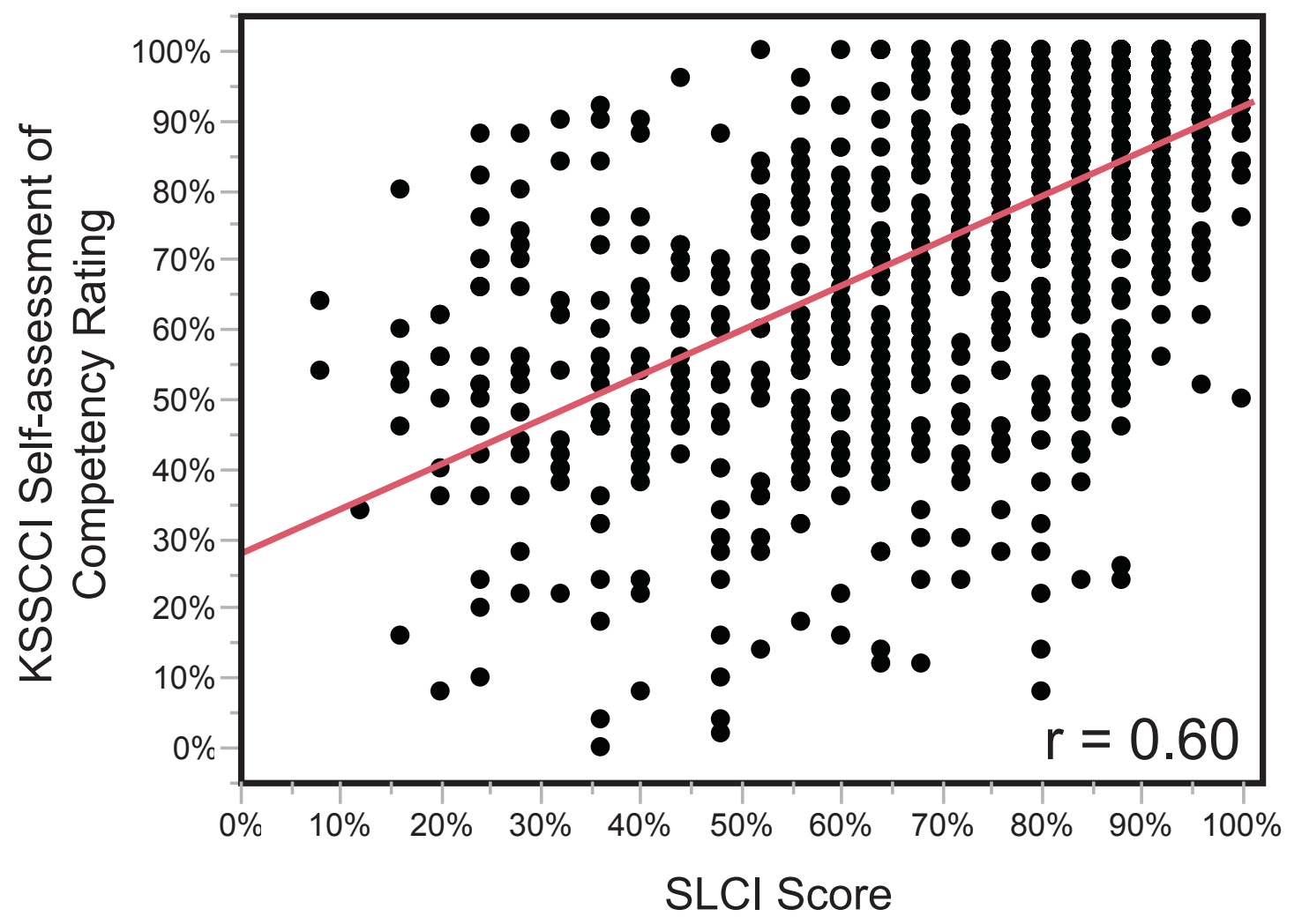


Figure 4. Correlation between actual performance on the SLCI and self-assessed competency through the KSSLCI for 1154 participants. Probability > F <0.0001.

What limits the correlations?

Reliability (R) = 2r/(1+r) (Spearman-Brown relationship). The degree to which data yielded by any instrument can correlate with itself (r) limits the degree to which it can correlate with data yielded by another instrument. Correlations between instruments of unknown reliability qualify as non-studies. The internal correlation (r) of our least reliable instrument (in this case r = 0.73 from the SLCI) limits the maximum correlation that we can expect between the SLCI and another measure (the KSSLCI). The actual correlation (Figure 4) was r = 0.60.

Required: good knowledge survey items

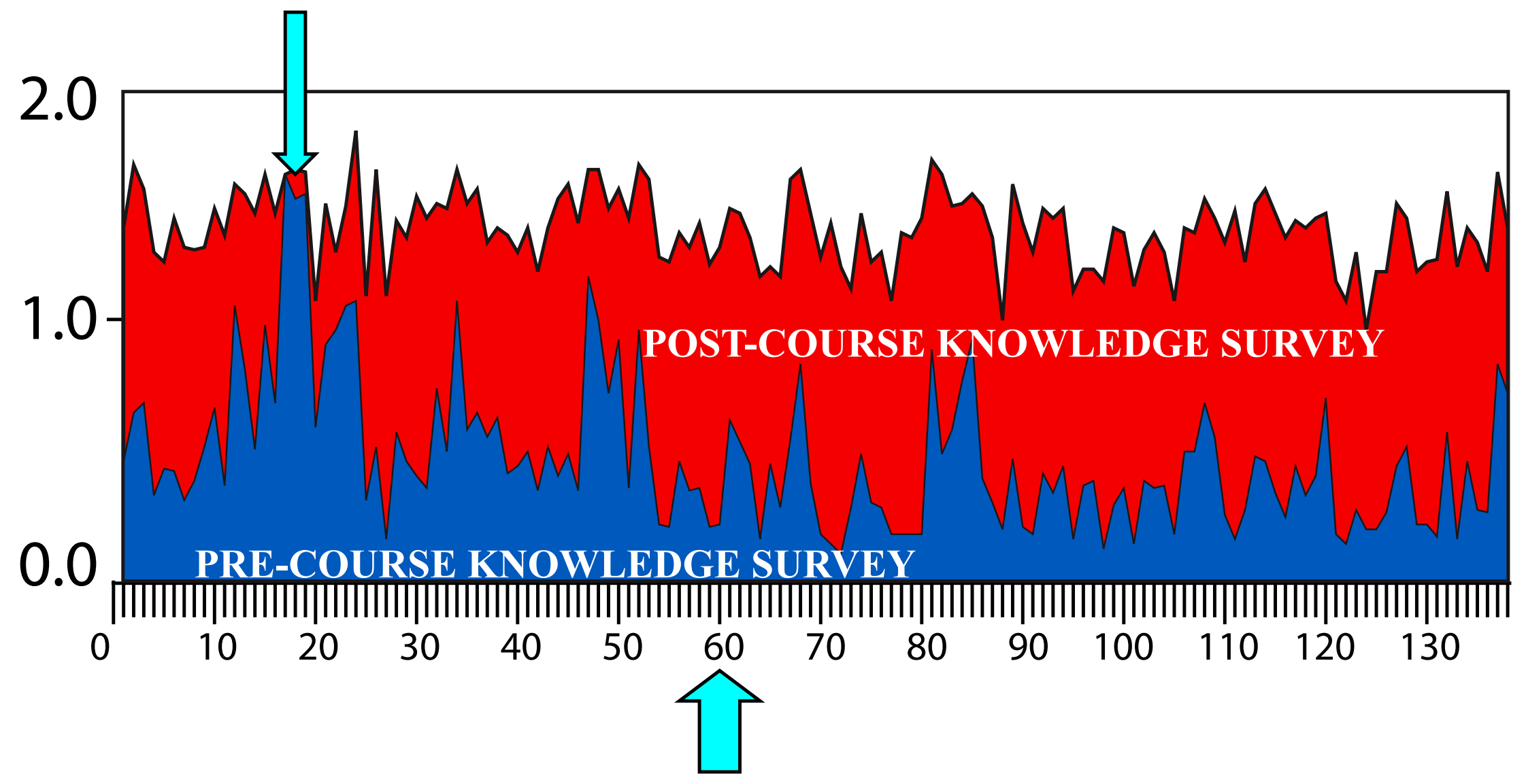


Figure 5. Broad goal statements are not appropriate items for self-assessments. Note reactions to questions #17: "I can think clearly and logically." and #18: "I can find and critically examine information." Good knowledge survey items are specific and assessable outcome statements, such as question 60: "I can outline Piaget's four main stages of cognitive development, and comment on how children's thinking changes during these four stages."

Required: an adequate database

Often, attempts to correlate knowledge surveys with direct assessments fail because the database is too small to achieve reliable measures. Figure 6 shows a commonly used graphical convention that began with the famous 1999 Kruger and Dunning paper, "Unskilled and Unaware of It…." This convention doesn't provide a metric for assessing graphics built from databases that are too small to achieve reliability.

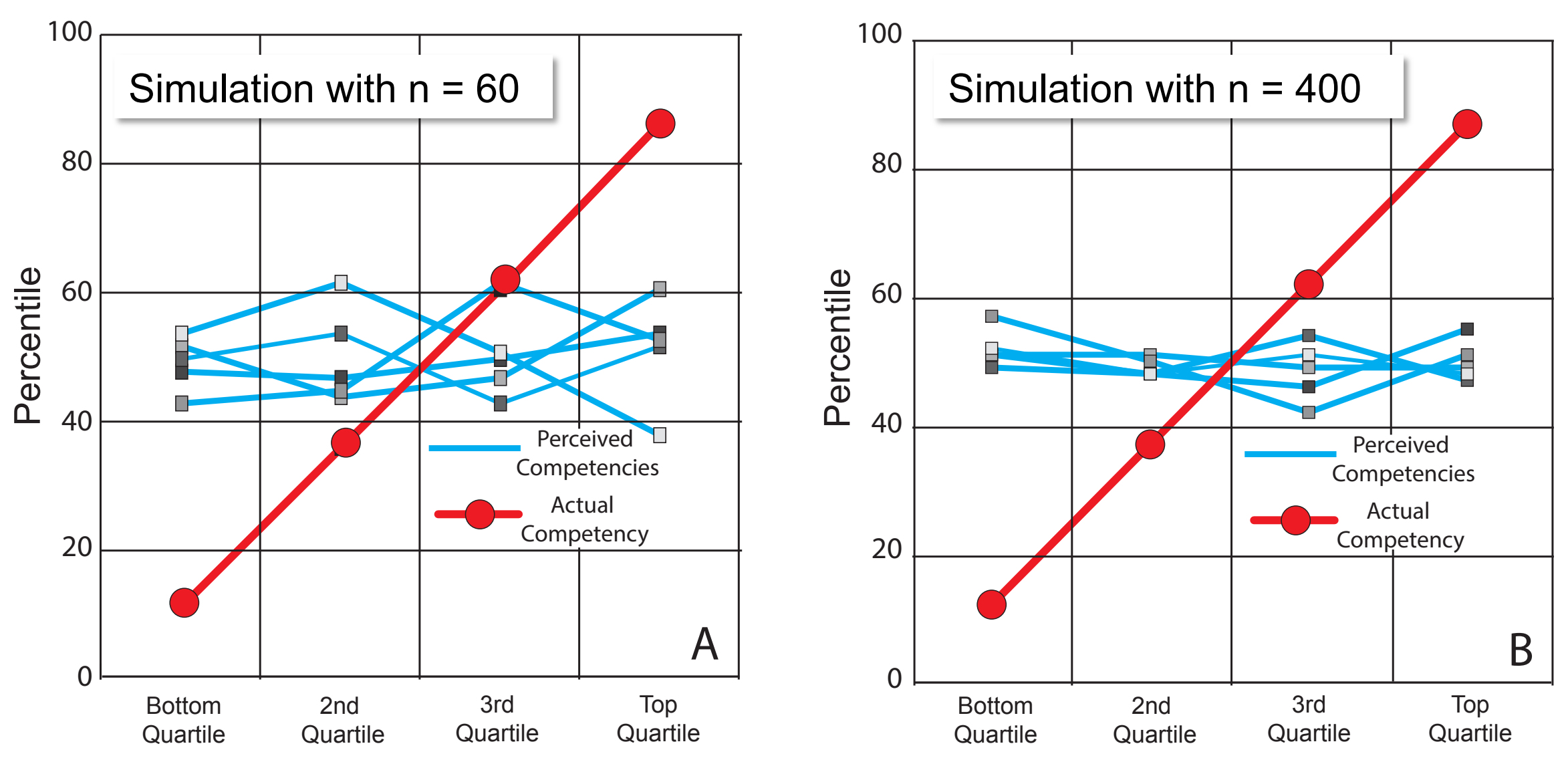


Figure 6. Random number simulations of two self-assessment studies of varied sizes showing five replications of each study. Fig. 6A simulates a study with 60 participants. The mean values vary greatly in each replication's perceived competencies as tabulated by quartiles, which shows the database is too small to achieve good reliability. Fig. 6B shows how raising the study populace to 400 participants allows any single replication to better represent the actual mean quartiles' values with more reliability.

Mistaking noise for signal

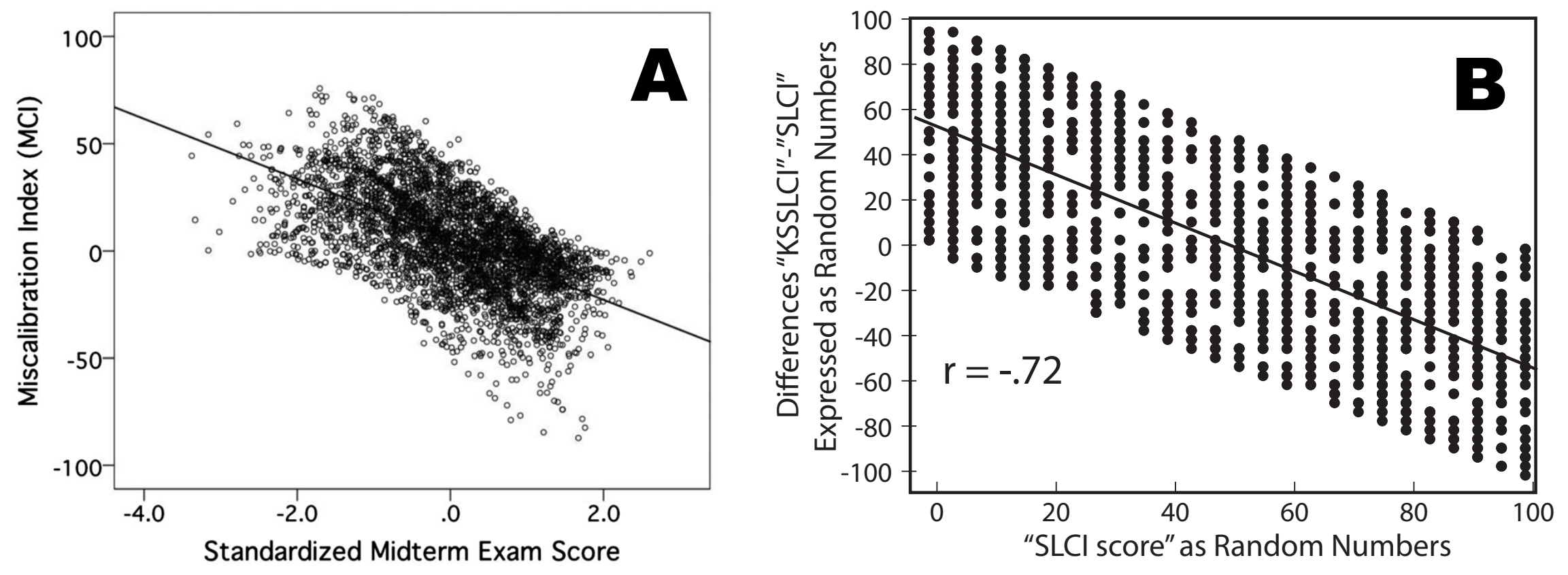


Figure 7. A few graphical conventions make noise look like a self-assessment signal. Fig. 7A is Figure 5 from Pazicni and Bauer (2014) who reported a correlation of mis-calibrated self-assessments as "strongly correlated with exam performance, Pearson r (two-tailed) = -0.587 at **r** < 0.001." However, Fig. 7B, presents 1154 data pairs modeled as random numbers (random noise). The apparent correlation is caused by a ceiling effect, and is nothing more than noise easily confused with a meaningful self-assessment measure.

Attenuating noise; clarifying signal

If the noise in actual measurements is mostly random, then attenuating its effects by averaging the measures should be possible. Given a sufficient database, averaged data should cancel out random noise and allow the definition of the signal to improve.

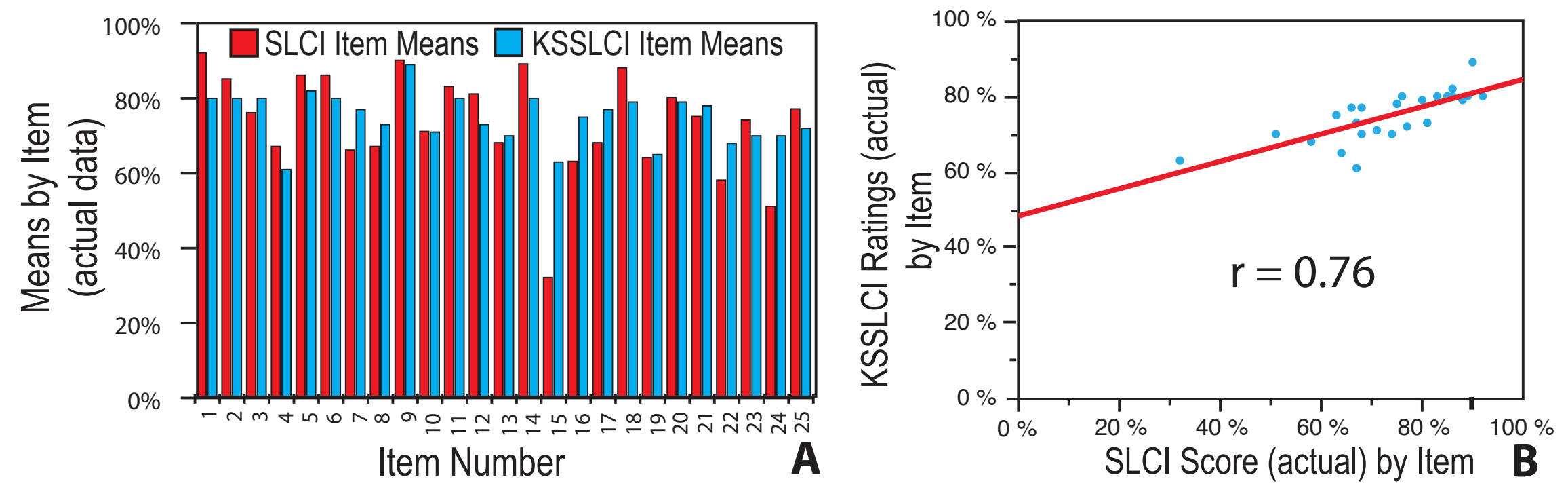


Figure 8. Means from 1154 participants of SLCI scores and KSSLCI self-assessment ratings on each item (Fig. 8A). A plot of the collective item-by-item averages of the KSSLCI ratings versus item-by-item average SLCI scores (Fig. 8B) reveals a highly significant correlation. The averaging of collective measures cancelled out noise and permitted the self-assessment signal to emerge clearly. This is the kind of averaging used to produce Figs. 1, 2, and 5 and to determine learning gains in pre- post- course class knowledge surveys.

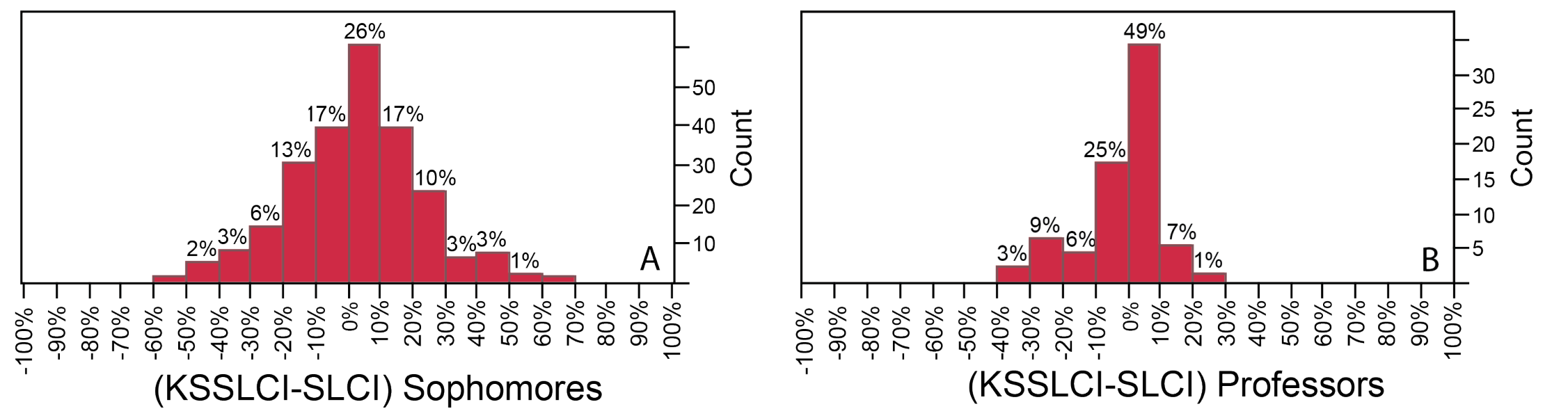


Figure 9. Histograms of KSSLCI-SLCI data for sophomores (n = 235) and professors (n = 69). A greater proportion of professors (74%) accurately self-assessed their competency (±10%) compared with only 43% of sophomores. Histograms offer a better way to evaluate the contributions of noise and signal to self-assessment data.

Conclusions

Ours is the first report using a large database to a knowledge survey of known reliability with scores from a well-aligned competency measure of known reliability. Other research that attempted to quantitatively document the relevance of self-assessments to actual performance likely failed because investigators: (a) failed to recognize reliability as fundamental, (b) failed to acquire an adequate database, (c) mistook patterns produced by noise for patterns produced by the signal or (d) failed to carefully align the paired instruments to measure the same construct.

Collective results from pre- and post- course self-assessments such as knowledge surveys should provide meaningful assessments of course content mastery, providing that the course knowledge surveys align well with what instructors test and with what they teach.

This poster is distilled from a more detailed paper now under review.