Numeracy
Advancing Education in Quantitative Literacy

Volume 9 | Issue 1 Article 4

2016

Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency

Edward Nuhfer
California State University (retired), enuhfer@earthlink.net
Christopher Cogan
Independent Consultant, cbcmapper@gmail.com
Steven Fleisher
California State University - Channel Islands, steven.fleisher@csuci.edu
Eric Gaze
Bowdoin College, egaze@bowdoin.edu
Karl Wirth
Macalester College, wirth@macalester.edu

Recommended Citation

Nuhfer, Edward, Christopher Cogan, Steven Fleisher, Eric Gaze, and Karl Wirth. "Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency." *Numeracy* 9, Iss. 1 (2016): Article 4. DOI: http://dx.doi.org/10.5038/1936-4660.9.1.4

Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency

Abstract

Self-assessment measures of competency are blends of an authentic self-assessment signal that researchers seek to measure and random disorder or "noise" that accompanies that signal. In this study, we use random number simulations to explore how random noise affects critical aspects of self-assessment investigations: reliability, correlation, critical sample size, and the graphical representations of self-assessment data. We show that graphical conventions common in the self-assessment literature introduce artifacts that invite misinterpretation. Troublesome conventions include: (*y* minus *x*) *vs.* (*x*) scatterplots; (*y* minus *x*) *vs.* (*x*) column graphs aggregated as quantiles; line charts that display data aggregated as quantiles; and some histograms. Graphical conventions that generate minimal artifacts include scatterplots with a best-fit line that depict (*y*) vs. (*x*) measures (self-assessed competence *vs.* measured competence) plotted by individual participant scores, and (*y*) *vs.* (*x*) scatterplots of collective average measures of all participants plotted item-by-item. This last graphic convention attenuates noise and improves the definition of the signal. To provide relevant comparisons across varied graphical conventions, we use a single dataset derived from paired measures of 1154 participants' self-assessed competence and demonstrated competence in science literacy. Our results show that different numerical approaches employed in investigating and describing self-assessment accuracy are not equally valid. By modeling this dataset with random numbers, we show how recognizing the varied expressions of randomness in self-assessment data can improve the validity of numeracy-based descriptions of self-assessment.

Keywords

self-assessment, Dunning-Kruger Effect, knowledge surveys, reliability, graphs, numeracy, random number simulation, noise, signal

Cover Page Footnote

Edward Nuhfer served as Director of Faculty Development and Educational Assessment and tenured Professor of Geology at four universities. His research interests are in metacognitive self-assessment, the role of the affective domain, and curricular design for reflective, higher-level thinking. He continues actively in writing, research, and assessment.

Christopher Cogan is an independent consultant and practices in Environmental Science and Geographic Information Systems. He was a researcher at the Alfred Wegener Institute, a member of the California State University design team for the Science Literacy Concept Inventory, and a winner of the best teaching award at CSU Channel Islands. His interests are in GIS applications in the study of wildlife and in teaching for exceptional learning.

Steven Fleisher is Instructional Faculty in Psychology at California State University Channel Islands. His expertise is in teacher-student relationships and instructional methodologies that support student autonomy and learning. His research focus is on metacognition, self-regulated learning, positive affective environments, self-assessment and reflective thinking, and the neurobiology of learning.

Eric Gaze directs the Quantitative Reasoning (QR) program at Bowdoin College, is Chair of the Center for Learning and Teaching, and is a Lecturer in the Mathematics Department. He is the current President of the National Numeracy Network (2013 – 2015) and an associate editor of *Numeracy*. Eric has given talks and led workshops on the topics of Quantitative Reasoning course development and assessment.

Karl Wirth is an Associate Professor of Geology at Macalester College. His research focuses on metacognition, motivation, and undergraduate research experiences in support of best practices in teaching and learning in undergraduate STEM. As assessment coordinator for the Keck Geology Consortium, he seeks to improve undergraduate research experiences through the development of intentional curricular structures and mentoring practices.

Introduction

Self-assessment is a personal judgment of one's capacity to perform competently with present skills and knowledge. How accurately do peoples' self-assessments of their competency predict their demonstrated competency when they engage a challenge? Answering this question is the essence of studies that seek to measure self-assessment skill.

Self-assessment is a valuable metacognitive skill that improves through instruction (Kruger and Dunning 1999; Caputo and Dunning 2005). Bell and Volckmann (2011) suggested that early identification of students with poor self-assessment skills could allow timely training in self-assessment that might help these students to have greater success in college. Because self-assessment appears to be valuable and teachable, instructors from varied disciplines increasingly seek to measure their students' ability to accurately self-assess their competencies.

Self-assessment differs from self-efficacy (Bandura 1997), which refers to metacognitive confidence in one's abilities to acquire the capacity for competence through future preparation. Good self-assessment skills help to improve learning by building self-efficacy and "contribute to higher student achievement and improved behavior" (Ross 2006).

Our incentive for this paper began in 2011 while studying how students' self-assessments of perceived science literacy compared with their performances on a test of science literacy. We adopted a graphical convention used by Bell and Volckmann (2011). Their graphs displayed patterns resulting from the measures of self-assessed competencies of chemistry students and their actual proficiencies on chemistry tests. The patterns revealed the least-competent students as those most overconfident about their competence.

As our data grew, we watched our graphical patterns generated from measures of science literacy becoming identical with those of Bell and Volckmann. We began to suspect that the graphical convention we employed in common with Bell and Volckmann might account for this convergence, and we confirmed our suspicion by graphing nonsense data generated by random numbers. In that graphical convention, random numbers generated patterns very similar to those produced by our actual data. We subsequently used random number simulations of real data to examine other graphical conventions employed in the literature of self-assessment.

At first, measuring a person's skill in self-assessment of competency appears simple. It involves comparing a direct measure of confidence to perform taken through one instrument with a direct measure of demonstrated competence taken through another instrument. For people skillful in self-assessment, the scores on both self-assessment and performance measures should be about equal.

Departures from perfect self-assessment register by degrees in overconfidence or underconfidence.

In practice, measuring self-assessment accuracy is not simple. Obtaining meaningful results that have quantitative significance requires attention to the construction of the measuring instruments. The paired instruments must address a common construct; they must be capable of acquiring reliable data, and the investigators must acquire enough data before they can produce a contribution characterized by reproducible results. Unfortunately, investigators can still graph the data acquired while ignoring these fundamentals, and they can make convincing interpretations of the resulting patterns.

Several graphical conventions unique to the self-assessment literature generate artifact patterns that are easy to mistake as offering meaningful portrayals of self-assessment.

These difficulties contribute to the current situation when "…it remains unclear whether people generally perceive their skills accurately or inaccurately" (Zell and Krizan 2014, p. 111). The purpose of our paper is to increase awareness of these aspects when interpreting existing self-assessment literature and when doing the research to produce new knowledge about self-assessment.

In investigating the relationship between self-assessed competence and actual competence, the role of mathematics lies in describing the relationships; the role of the behavioral sciences lies in explaining them. The numerical description is indispensable because it assures that a credible signal exists that can be explained. In this paper, we focus solely on the descriptive role and employ our data to advance that understanding. We reserve contributing to explanations about the nature of self-assessment for a separate paper now in preparation.

To address the descriptive role, we find it useful to view human self-assessment measures as a blend of two components. The first is a meaningful self-assessment signal that investigators seek to detect and measure. The second is random noise that accompanies the signal. This simple distinction is between the order that is characteristic of a relevant signal and the disorder characteristic of irrelevant random noise. To study the effects of random noise on self-assessment measures, we simulate random noise with random numbers. In this paper, we do not discuss noise from a behavioral science perspective. For such a discussion, see Mueller and Weidemann (2008).

We advanced our understanding of self-assessment measures by following a practice recommended by teachers of quantitative literacy: use authentic data, and employ the power of spreadsheets to model the problem (Gaze 2014). Spreadsheets are especially helpful to answer "What if…?" questions through simulations. The awareness gained by doing so aided our understanding of published literature and planning further study of our data.

Competing Positions on Self-Assessment

Disparate results from self-assessment studies give rise to competing positions on self-assessment. One position holds that self-assessments of learning and competency offer little more than the random noise that arises from responses that are mere guesses about competency. As such, students' self-ratings of their understanding should contribute little value to assessments of their actual knowledge, skills or abilities to think. This position emerges whenever researchers consider relationships measured between self-assessed competence and actual competence as insignificant (Bowers et al. 2005; Porter 2012, 2013).

Contradicting this position are two positions that consider self-assessed competence as meaningful and measurable. One of these positions holds that people tend toward overconfidence in their abilities, with many being "unskilled and unaware of it." This view arises from findings that identify the least-proficient performers as those with the most over-inflated self-assessments (Kruger and Dunning 1999; Ehrlinger et al. 2008; Bell and Volckmann 2011).

The other position holds that self-assessment ratings, overall, reflect the competence that people usually can demonstrate. This position arises when researchers consider relationships between measures of self-assessed competence and actual competence as significant (Nuhfer and Knipp 2006; Favazzo et al. 2014).

Methods

We employ a single dataset[1] derived from paired measures of self-assessed competence and demonstrated competence throughout this study. Employing common data facilitates meaningful comparisons between disparate graphical conventions.

We compare each graph of authentic data to an equivalent graph constructed from random numbers and then explain the value derived from doing the simulation. By modeling pure noise with random numbers, we show how noise sometimes mimics the signal that investigators seek to measure, how it affects portrayals of self-assessment data, and how it sometimes confounds efforts to understand self-assessment.

Instruments

The Science Literacy Concept Inventory (from here on called the SLCI) is an instrument that tests proficiency in understanding science's way of knowing. Our direct measures of competency come from 1154 participants (undergraduates,

[1] Included in the Excel workbook of Appendix A under Additional Files for this article.

graduate students, and professors) who completed the SLCI. The participants represent selective and open-admission institutions and seem representative of the American higher education community.

The SLCI offers 25 multiple-choice items with four choices consisting of three distracters and only one correct answer. The 25 items map to 12 concepts (Nuhfer et al. 2010), which are relevant to understanding science's way of knowing the physical world. The SLCI's length is similar to that exhibited by the 20-item Quantitative Literacy Reasoning Assessment (QLRA, Gaze et al. 2014) and to that of the well-established 30-item Force Concept Inventory of physics (Lasry et al. 2011). Based upon testing of over 18,000 participants, the SLCI exhibits content, construct, criterion, concurrent, and discriminant validity. We provide evidence for this validity in another paper now under review.

A 25-item Knowledge Survey of the Science Literacy Concept Inventory (from now on called the KSSLCI) provides the quantitative measures of 1154 participants' self-assessments of competencies. Ross (2006) noted that self-assessment measures usually have high reliability, and knowledge surveys have particularly high reliability (Nuhfer and Knipp 2006). Well-constructed knowledge survey items express challenges that are specific and directly assessable through observed performance. The KSSLCI yielded self-assessed competency measures with Cronbach Coefficient Alpha Reliability = .94 and Spearman-Brown prophecy = .93.

Knowledge surveys (Nuhfer and Knipp 2003; see tutorials and downloadable examples[2]) 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 scale:

> A. I can fully address this item now for graded test purposes.
>
> B. I have partial knowledge that permits me to address at least 50% of this item.
>
> C. I am not yet able to address this item adequately for graded test purposes.

The choices A through C register in data as numbers 2, 1, and 0 respectively. Simple three-item choice formats appear psychometrically sound (Landrum et al. 1993; Rodriguez 2005; Baghaei and Amrahi 2011) and expedite quick, clear distinctions.

In this study, we expressed every participant's 25-item KSSLCI rating in percent as derived from (sum of item scores*2)/100. This expression allows comparisons with the measured competency scores expressed as percent correct on the 25-item SLCI.

Our participants completed both instruments online in one sitting, and our 1154 data pairs come from completed instruments with no missing responses. The

[2] http://www.merlot.org/merlot/viewMaterial.htm?id= 437918 (accessed Dec 2, 2015)

participants were able to complete the measures at their pace so that participants could be reflective and engage with the exercise under relaxed everyday conditions rather than in a timed environment.

In our design for this study, we maximized alignment of our two instruments by having the KSSLCI and the SLCI derive their measures from the identically worded 25 items. Challenges articulated differently often communicate different meanings and can trigger different levels of comprehension (Gendall and Hoek 1990). In such cases, participants studied might have a significant self-assessment capability, but poorly aligned instruments are insufficient to capture that trait.

Reliability, Random Noise, and Random Numbers

Data acquired that are unreliable or obtained from misaligned instruments are likely to be mostly noise. Before we could begin graphing or further studying paired measures, we needed to confirm that both of our instruments collected data that revealed a signal and thus were distinct from pure noise. If such were not the case, our study could not have progressed further. In studies of self-assessment, this is particularly necessary because a position already exists that argues that human self-assessments are mostly random noise.

> … if the critics are correct, and students lack the cognitive ability to accurately answer most survey questions, then the critics are, in essence, arguing that students must be generating random self-assessment responses to survey questions (McCormick and McClenney 2012[3]) (Porter, 2013, p. 202).
>
> Theoretical models of cognition, and empirical evidence to date demonstrate that self-reported learning gains are mostly noise and cannot be used to assess student learning (Porter, 2012, p. 6).

Quantifying reliability offers a way to learn when instruments yield data that are largely random noise. The Spearman-Brown definition of reliability (Jacobs and Chase 1992), $R = 2r/(1 + r)$, offers a clear relationship between reliability (R) of an instrument to discriminate between different individuals' abilities and the internal coefficient of correlation (r), which is a measure of the instrument's ability to generate data that can correlate with itself. A rule of thumb is to accept only data with a minimum reliability (R) of .7 for research purposes (DeVellis 2003), which implies an internal correlation of about $r = .54$.

Figure 1A displays a split-halves approach to calculating a Spearman-Brown Reliability (R). One derives r as the linear correlation coefficient obtained by correlating each of our 1154 participants' scores generated from the odd-numbered items with their scores generated from the even-numbered items on the

[3] Although Porter (2013) cited McCormick and McClenney (2012) in his quotation, their 2012 text discloses reservations about self-assessments being only random guesses.

25-item KSSLCI. The internal correlation's value of $r = .876$ yielded by participants' self-assessments produced a Spearman-Brown Reliability (R) of .93.

What pattern would a scatterplot produce if the data were simply random noise? To address this question, we replaced our 1154 measures of self-assessed competency (KSSLCI) with random numbers. To produce our random number simulation, we used the RANDBETWEEN (0,2) command of EXCEL in each of the 25 cells for recording a random 3-item self-rating on the 25-item knowledge survey. We then replaced the actual self-assessment ratings for each of our 1154 respondents with the ratings calculated from these random numbers. Figure 1B is an expression of noise produced from paired data that consist of the means (even and odd items on the simulated 25-item instrument) of two sets of random numbers bounded by 0 and 100. Plotting means of each pair generates a somewhat circular pattern clustered around the theoretical mean point of (50, 50). The pattern reveals no trend that differentiates individuals by high or low self-assessment confidence. In contrast, Figure 1A reveals an ordered trend and confirms that the self-assessment signal we sought to measure is present.

Some noise is probably present in all measures of human behavior. An expression of the perfect self-assessment signal in Figure 1A would have been a correlation of $r = 1$, with all data points in Figure 1A plotting directly on the regression line. We view the difference between this theoretically perfect expression of $r = 1.0$ and our actual measured self-assessments of $r = .876$ as arising largely from some noise present in our actual data.

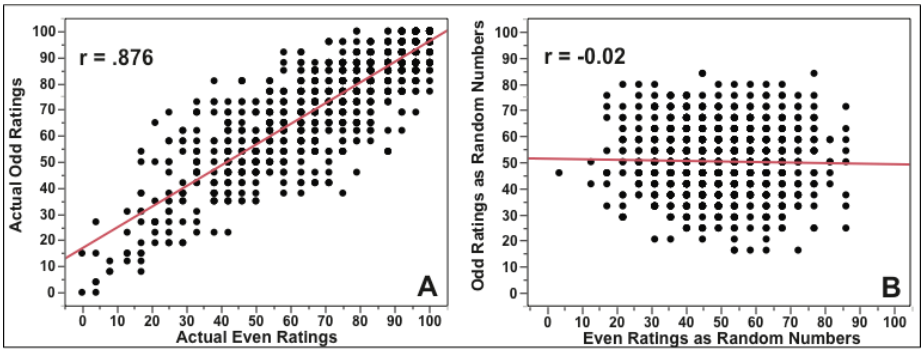

Figure 1. Comparison of split-halves (y) *vs.* (x) scatterplot patterns yielded from correlating 1154 participants' KSSLCI ratings calculated on even and odd items (A) with these same participants' ratings replaced with random numbers 0, 1 or 2 (B).

Preparing a random number model of a dataset, displaying randomness visually, and comparing the graphical patterns as done in Figure 1 may at first appear trivial. Indeed, to most readers, the results shown are probably intuitively obvious in the familiar convention of the (y) versus (x) scatterplot used in that figure. However, such results are not intuitively obvious in other graphical conventions that we address later. The value of comparing patterns of authentic

data with those yielded by simulation of the data with random numbers lies in helping to distinguish the patterns of random noise from the patterns of a strong signal in any graphical convention.

In a parallel approach, we used our dataset to measure the SLCI's reliability too, through several conventions. All proved numerically similar: Cronbach Coefficient Alpha = .85, Kuder-Richardson KR20 = .85, Kuder-Richardson KR21 = .84, and Spearman-Brown prophecy = .87. Thus, we confirmed that both of our instruments collect reliable data that are measurably distinct from random noise.

Studies performed with good instruments may still fail to achieve reliability if the database of participants is too small to allow a signal to emerge from the noise. Later in this paper, we show how random number simulations can help reveal the size of the database needed for reproducibility.

Reliability rests on the data furnished by the instrument, not the instrument itself. Therefore, it is not safe to employ a standardized instrument whose known reliability is derived from a much larger database than the dataset under investigation. Just as reliability must be established from the dataset under consideration, a random number simulation of actual data must employ the same number of participants as exist in the actual dataset.

At the end of data acquisition stage, we assigned random numbers to the 1154 lines that contained both the SLCI and their equivalent KSSLCI data. We then sorted the data by random number assignment and ran reliability estimates on successively smaller splits of our dataset. We could then see the reliability that both instruments yielded from smaller subsets. This result confirmed that our dataset of 1154 participants was several times as large as that required for establishing reproducible results.

Before we could use our data to describe the relationships between paired measures, we needed to confirm that our instruments yielded reliable data, were well aligned, that the data we collected differed substantially from random noise, and that our dataset was sufficiently large. Having done so, we next use the data to explore the issues that occur when investigating relationships between paired measures of self-assessed competency (KSSLCI) and actual competency (SLCI).

Results

Simple Scatterplots and Linear Correlation

Calculating a linear correlation coefficient between individuals' self-assessed ratings of competency and scores generated on a test of competency is a prevalent approach to determining a meaningful relationship between the two (Dunning and Helzer 2014). The most common graphical convention for expressing results is a simple (y) *vs.* (x) scatterplot with a best-fit line (Fig. 2). In this convention,

patterns depicting the order of the self-assessment signal (Fig. 2A) are distinct from the disorder of random noise (Fig. 2B).

Reliability is fundamental to understanding paired correlations. When researchers publish correlations produced by two sets of measures of undocumented reliability (Bowers et al. 2005), readers cannot evaluate the results in the absence of knowing what was correlated.

Cashin (1988, p. 2) provided brief but valuable guidelines regarding the usefulness of correlations expected from studies that utilize the instruments of the social sciences such as tests and surveys:

> Correlations between .20 and .49 are practically useful. Correlations between .50 and .70 are very useful but they are rare when studying complex phenomenon.

Considering data as mixtures of signal and noise clarifies why higher correlations are rare in the social sciences. Measures captured by a test or a survey consist of a mix of signal and noise, so data accumulated by a single test or survey instrument at best achieves an imperfect correlation with itself. The degree to which data yielded by any instrument can correlate with itself limits the degree to which it can meaningfully correlate with data yielded by another instrument. A correlation coefficient derived from pairing measures from two imperfect instruments should thus be even lower than the internal correlation of the least reliable of the paired measures.

The Spearman-Brown reliabilities of the SLCI and the KSSLCI, yield respective internal correlations of each of $r = .73$ and .87. The internal correlation of our least-reliable instrument is $r = .73$ from the SLCI. The actual correlation coefficient between the paired SLCI scores and KSSLCI ratings for our example is $r = .60$ with highly significant $p < .0001$ (Fig. 2A). In contrast, a random noise simulation of our dataset (Fig. 2B) yields $r = .02$ with insignificant $p < .5053$.

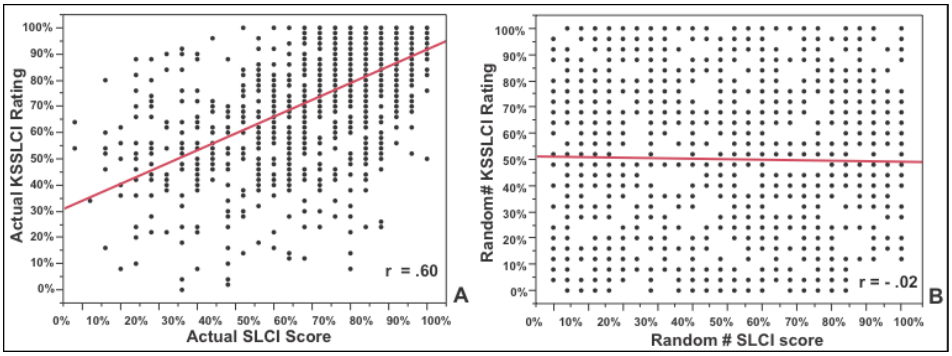

Figure 2. Scatterplots (y *vs.* x) between actual self-assessment measures (A) and simulated self-assessment measures with random numbers (B) for 1154 participants. SLCI scores measure actual competence. KSSLCI ratings measure self-assessed competence.

The pattern in Figure 2A reveals that people who self-assessed on the KSSLCI that they would do poorly on the SLCI did tend to score lower, and those who self-assessed that they would do well, as a whole, scored higher. To be sure, this general trend had many exceptions. The r of .60 reveals that the relationship between self-assessed competency and demonstrated competency does not permit prediction of one from the other at the level of individual participants. Of all of the graphical depictions of self-assessments, the scatterplot in Figure 2B expresses the pattern that most people intuitively recognize as a display of randomness with no significant trend.

Global Items Contrasted with Knowledge Surveys

Global items are single queries written to elicit self-assessments of general competence in a broad area such as humor, science, logical thinking, or overall performance on a long test. In our measures of self-assessment, we asked participants to self-assess competency through three global items and a knowledge survey consisting of 25 specific measures that map into a broader common area (Table 1). Our first global question employed a description of the SLCI (Table 1). When participants self-assessed their understanding of science literacy in response to that query, the responses produced a correlation of $r = .29$ between the participants' self-assessed competencies and their actual SLCI performance scores. While positive and significant at the 99% confidence level, taking the 25-item knowledge survey (KSSLCI) immediately after answering that global query seemed to clarify to participants the specifics of the challenges that a science-literate person should be able to meet. The correlation between the KSSLCI and the SLCI ($r = .60$) was more than double that generated by correlating the first global item with the SLCI.

Table 1.
Correlations between the Science Literacy Concept Inventory (SLCI) and Four Self-Assessments.

Self-Assessments in %	Correlation with SLCI scores in %
1. "A multiple choice test has been designed to measure how well citizens understand the thinking process that scientists employ to understand the physical world. The test is not timed and can be done online in any setting. The test does not depend upon factual recall of knowledge. Any factual information needed or meanings of any technical terms used are provided within the test itself. Based on your feelings of self-assessment at this time, what is the score in percent (Write as % an estimate between 0% and 100%) that you believe that you would obtain if you took such a test?"	$r = .29$
2. Knowledge Survey (KSSLCI): cumulative rating in % derived from all 25 items in total	$r = .60$
3. "Based only on your gut feelings established after taking this knowledge survey, what score in percent (between 0% and 100%) do you think you would obtain if you actually had to answer the twenty-five questions?"	$r = .51$
4. "Now that you have completed taking the Inventory, what score in percent (between 0% and 100%) do you think you actually obtained?"	$r = .59*$

Items 1, 3 and 4 are general global queries. Participants completed these in the numerical sequence provided.
*$N = 1154$ except for Item #4 that was added later in this study where $N = 662$.

Thereafter, the participants appeared to retain the understanding established through taking the KSSLCI in the two subsequent global assessments (Table 1). No teaching occurred to produce this increased understanding.

Most published studies that we cite in this paper employed global items to measure self-assessment accuracy. Table 1 shows that numerical results can differ depending on whether people self-assess their competence through global items or instruments like the SLCI and KSSLCI. Our self-assessment measures employed in the graphical representations in this paper derive only from the paired measures of the SLCI and KSSLCI (Item #2 in Table 1). In describing self-assessment accuracy, the graphical relationships remain the same regardless of the instruments used to generate the paired measures. Considering both of these self-assessments as equivalent may be problematic for explaining self-assessment accuracy. Such explanations are not addressed in this paper.

Computing the linear correlation coefficient between self-assessed competency and demonstrated competency is likely to be a fruitful effort when done with a sufficiently large and reliable database. The scatterplot with a best-fit line is informative, especially when we can view the patterns yielded by actual data next to patterns yielded by random number simulations of those same data. However, improved understanding of self-assessment results from going beyond calculating a correlation coefficient (Dunning and Helzer, 2014).

Kruger-Dunning Graphical Convention

Online searches for "Dunning-Kruger Effect" reveal a popular belief that the general populace is "unskilled and unaware of it," with a significant portion of the populace inclined to make self-assessments that grossly inflate their actual abilities. This belief originated from an influential paper (Kruger and Dunning 1999) that employed global queries and provided results in a graphical convention typified by Figure 3. It displays the worst-performing participants as grossly overestimating their ability to perform and the best-performing participants as having a tendency toward accurate self-assessment or toward underestimating their ability to perform by small amounts.

In this convention, the self-assessed competency ratings and actual competency scores are tabulated in two columns of a spreadsheet, followed by the sorting of both columns together in the ascending order of actual competency. This sorting is necessary for tabulating the same participants' self-assessments and performance measures as thirds (Bell and Volckmann 2011) or quartiles (Kruger and Dunning 1999; Kennedy et al. 2002; Ehrlinger et al. 2008; Pazicni and Bauer, 2013). The computational algorithm we used to construct our Kruger-Dunning graphs is as follows:

1. Use participants' responses to calculate each participant's raw test score and self-assessment score. Place into spreadsheet respectively as columns 1 and 2.
2. Convert participants' test scores and self-assessments into percentiles and place into columns 3 and 4.
3. Sort columns 1 and 2 together by ascending participants' competency scores. Later, do the same with columns 3 and 4.
4. Define the boundary scores of four quartiles in the test as raw scores. Later, do the same with the data expressed in percentiles. Note that different software packages that convert raw scores into percentiles or that define the quartile boundaries use different conventions that may produce slightly different results (Hyndman and Yanan 1996).
5. Compute the average ratings for self-assessments as percentiles by quartile; plot these points, and connect these points by lines as shown in Figures 3, 4 and 5.

Published research that employed this graphical convention showed results that almost invariably supported those of Kruger and Dunning (1999). They concluded that the poorest performers are those who greatly overestimate their abilities, and the best performers are those who tend to underestimate theirs slightly (Fig. 3).

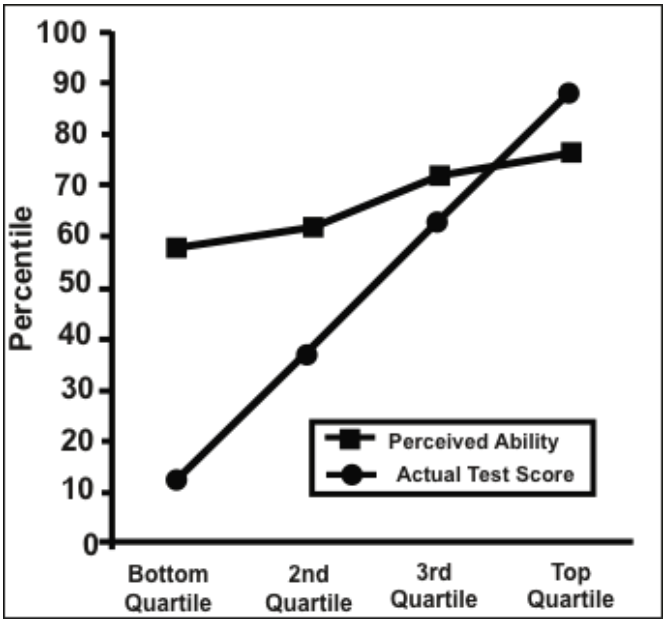

Figure 3. Kruger-Dunning convention of line chart of data aggregated by quartiles showing self-assessed competency to recognize humor as compared to actual test performance of competency in 60 participants (adapted with permission of American Psychological Association from Kruger and Dunning, 1999, Fig 1, p. 1124). The "Actual Test Score" is equivalent to our SLCI scores; the "Perceived Ability" is equivalent to our KSSLCI ratings.

Figure 4 employs the convention of Figure 3 to provide a synopsis of the full scatterplot (Fig. 2A). Whereas Figure 3 portrays a case in which overestimation greatly exceeds underestimation of abilities, Figure 4 shows only a modest difference between overestimation and underestimation.

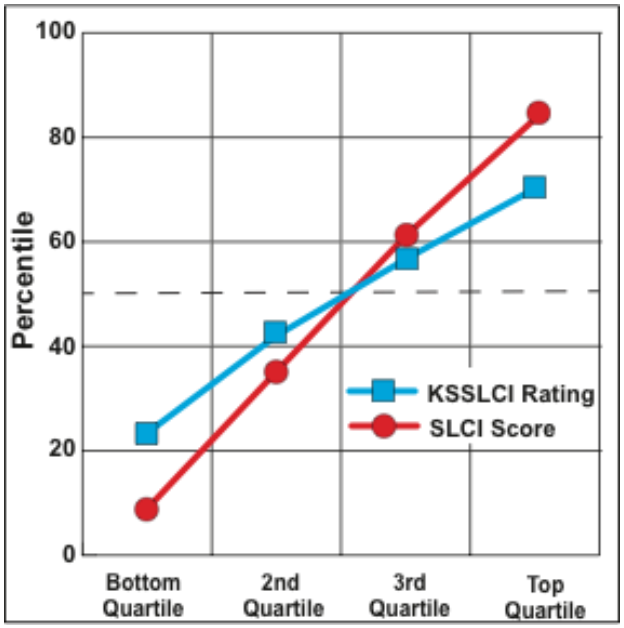

Figure 4. Line chart generated by graphing 1154 paired SLCI scores and KSSLCI ratings (perceived ability) through the Kruger-Dunning convention. Pairs expressed in percentiles are sorted by SLCI scores followed by calculating averages of SLCI scores and KSSLCI ratings within each quartile. Some investigators publish similar graphs but employ raw data as percentages (see Ehrlinger et al. 2008).

We can use random numbers to understand better this convention by knowing the patterns that our data would produce if they consisted of a perfect signal devoid of noise (Fig. 5A) or if they were only pure noise (Fig. 5B).

Consider the common case in which some people overestimate their abilities, some underestimate, and some estimate fairly accurately. After sorting the paired measures by the performance measure, the average of the top quartile of performance scores will always exceed the average of the accompanying self-assessments of the members of that quartile. Likewise, the average of the bottom quartile of performance scores will always be less than the average of the accompanying self-assessments of the members of the bottom quartile.

In the depiction of pure signal devoid of noise, both self-assessed competence and actual competence are identical. The self-assessed competence line coincides with the line that portrays the actual competency measure (Fig. 5, "SLCI Score"). In both measures, connecting the means of each quartile produces a line that slopes positively at about 45° (Fig. 5B, "KSSLCI Rating").

In graphing random noise, the plot of self-assessed competence produces a nearly horizontal line that lies along the common mean of the 50th percentile (Fig. 5B). Interestingly, Figures 5B and 2B represent the same random number data.

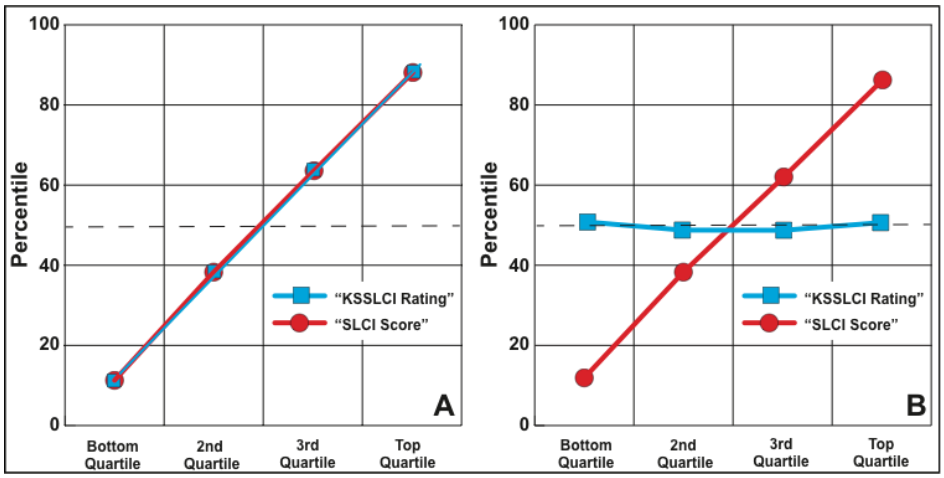

Figure 5. Kruger-Dunning type line charts of simulated data for 1154 data points (A) shows a pure self-assessment signal that results when the self-assessment KSSLCI ratings are the same as the actual SLCI performance scores. B shows the pattern of pure random noise generated by graphing random number pairs bounded by 0 and 100.

Actual self-assessment measures are blends of signal and noise, so simulations like those in Figure 5A and 5B are useful for detecting the signal-to-noise ratio in actual human self-assessment measures. We can see that the self-assessment "KSSLCI Rating" line in Figure 4 derived from our actual data is steeply inclined and seems rotated just a few degrees clockwise from its position in the simulated perfect self-assessment (Fig. 5A). Figure 4 expresses a high signal-to-noise ratio with a closer pattern semblance to Figure 5A than to Figure 5B.

Figure 5 reveals a key for detecting the degree of relative influence of noise *versus* signal in this graphical convention. In random number datasets of sufficient size, the means of the random self-assessments (Fig. 5B, "KSSLCI Ratings") for every quartile will be about equal and close to 50. Connecting these means produces a nearly horizontal line along the 50th percentile. The more that this self-assessed competence line displays horizontality, the more likely that the data producing the line is random noise. The expressed overconfidence by the least competent (Bottom Quartile) and underconfidence by the most competent (Top Quartile) are both large in the random number simulation (Fig. 5B), and the degree of inaccuracy is about the same for both.

The Kruger-Dunning convention offers a convenient way to use random number simulation to discover whether our dataset is sufficiently large to yield

reproducible results. Such simulations require using the same number of pairs of random numbers in the simulation as exists in the dataset. For an example, we will use 60, to simulate the dataset size that produced Figure 3. The theoretical mean of an unsorted set of random numbers bounded by 0 and 100 lies at the 50th percentile. Our dataset apportions 14 participants into each performance quartile. If this is sufficient, the mean of each quartile should be near the 50th percentile (dashed line, Figs. 6A, 6B). We can run several simulations with these random number data to see the reproducibility obtained from a dataset of this size and with the given bounds.

Five replications (Fig. 6A) indicate that when sorted by the data column containing the measures of actual competency, the lines that portray actual competency (red lines in Figs. 6A, 6B) in all five simulations are so consistent that their line plots are indistinguishable from one another. However, each replication produces notably different results from the accompanying measures of self-assessed competencies.

The differences are large enough to show that another study with a similar number of participants may not reproduce the initial results. Figure 6 shows the degree to which increasing our study populationfrom 60 to 400 will improve reproducibility by allowing the mean of each quartile to represent more accurately the theoretical true mean.

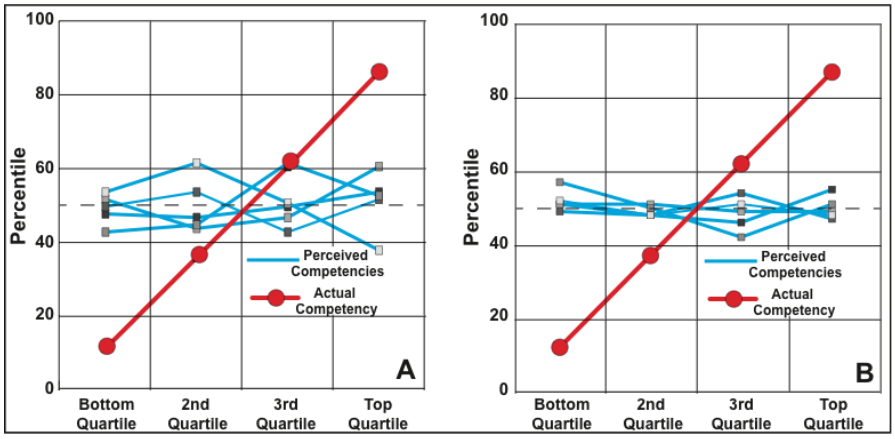

Figure 6. Kruger-Dunning type graphs of random number simulations of two self-assessment studies of varied sizes showing five replications of each study. A simulates a study with 60 participants. B shows the effect of raising the study population to 400 participants. In B, the quartile means in every replication cluster more tightly along the true mean at the 50th percentile (dashed).

The Kruger-Dunning convention can produce useful and informative graphs when sufficient and reliable data exist, but the convention carries two hazards. One is that a database too small to generate reproducible results will yield patterns that seem persuasive and meaningful to interpret. The second is that random noise

produces patterns that appear ordered and invite interpretation. Only some X-shaped patterns displayed in this convention present a meaningful self-assessment signal. Investigators who model their data with random numbers should achieve the understanding needed to avoid both hazards.

Bell-Volckmann and Pazicni-Bauer Graphical Conventions

Figures 7 and 8 employ a $(y - x)$ *versus* (x) convention that graphs the difference between self-assessment ratings and performance scores on the ordinate and the actual performance scores on the abscissa (Bell and Volckmann 2011; Pazicni and Bauer 2013). We used the following computational algorithm to present our data through this graphical convention:

1. Use participants' responses to calculate each participant's competency score and self-assessment ratings. Place into spreadsheet columns 1 and 2.
2. Calculate the difference between each participant's self-assessment score and test of competency score. Place into spreadsheet column 3.
3. Sort the three columns in ascending order of participants' competency scores.
4. Define the boundary scores of lower, mid and upper thirds (or quartiles).
5. Average the scores and differences from spreadsheet columns 1 and 3 within each third (or quartile) and plot as the column graph of the type shown in Figure 7.

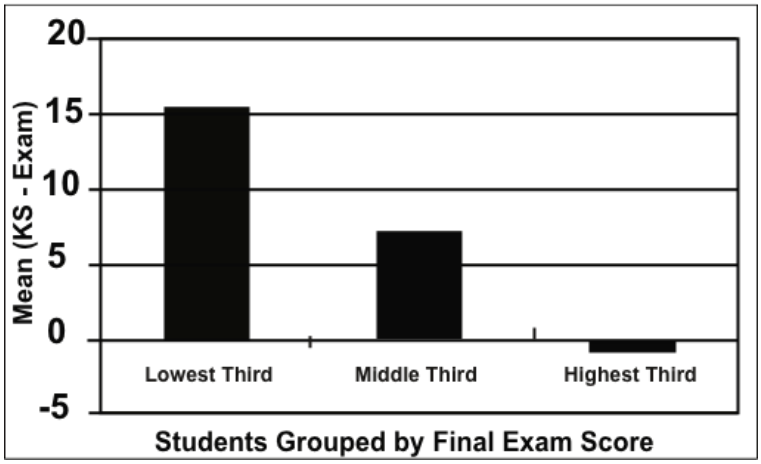

Figure 7. Column graph of $(y - x)$ *vs.* (x) type that summarizes self-assessment accuracies in an introductory chemistry class by thirds. The ordinate depicts the difference between the self-assessment ratings from a knowledge survey and actual exam scores. The abscissa portrays the exam scores. (Adapted from Bell and Volckmann 2011, p 1473, Fig. 6, with permission of American Chemical Society).

The column graph (Fig. 7) displays the worst-performing students on a chemistry final exam as those who most seriously overestimate their performance.

The scatterplot (Fig 8A) discloses our raw data by the individual participants before it becomes aggregated by performance into thirds to provide the column graphs of Figure 8B. Like our Figure 8A, Pazicni and Bauer (2013, p. 28, Fig. 5) display raw data as a $(y - x)$ vs. (x) scatterplot. Our data graphed in this convention yield a correlation of $r = -.39$, whereas Pazicni and Bauer's data yielded $r = -.587$.

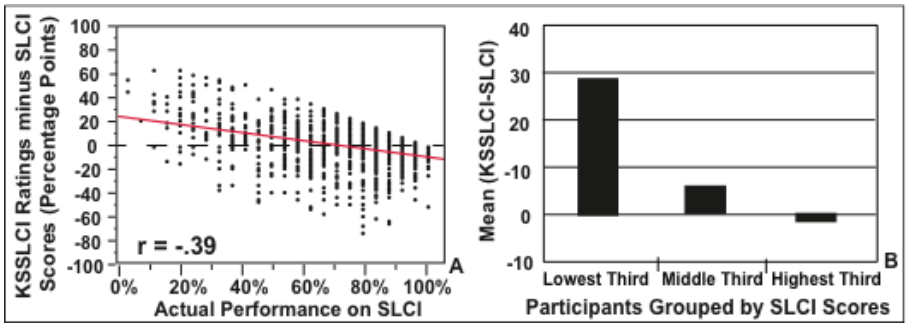

Figure 8. Actual data graphed as $(y - x)$ *vs.* (x) type scatterplot (A) and column graph (B) from 1154 participants. B employs the data from A to depict the mean accuracies of the bottom, middle, and top thirds of performers.

All figures produced by this convention invite the conclusion that the highest-performing participants are very good judges of their abilities, whereas the lowest-performing participants greatly overestimate their abilities to perform. However, arriving at this conclusion overlooks recognizing that the probability of overestimation increases from right to left in graphs like Figure 8A. The impossibility of being able to overestimate one's confidence by any percentage points (ppts) beyond a test score of 100% defines a ceiling. High performers cannot overestimate their competence by much. Numerically, they have little potential to do so.

Kruger and Dunning were aware of this problem: "If one has a low score, one has a better chance of overestimating one's performance than underestimating it." (Kruger and Dunning 1999, p.1124). By creating a random number simulation of Figure 8, we can appreciate the power of that "better chance" to influence the portrayals of data.

To create our simulation, we used Excel's command RANDBETWEEN (0, 25) to generate a random number data value for each participant's "KSSLCI rating" and "SLCI score." We then multiplied each random number by 4 to simulate the patterns yielded by a 25-item instrument in scores ranging from 0% to 100%. Graphing these random numbers in the convention of Figure 8A produces a scatterplot (Fig. 9A) that depicts a pronounced negative relationship between the simulated inaccuracy of self-assessed competency (derived by subtraction, KSSLCI– SLCI) and the simulated actual competency (SLCI). The random number simulation mimics the human-generated responses in our Figure

8A and the responses shown in the similar figure published by Pazicni and Bauer (2013, p. 28, Fig. 5).

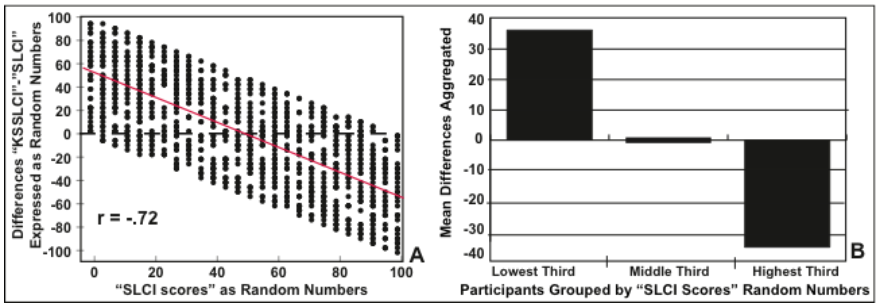

Figure 9. Scatterplot of type $(y - x)$ *vs.* (x) produced by replacing the 1154 data pairs from Figure 8A with random numbers (A). The column graph (B) displays the data from Figure A aggregated by the bottom, middle, and top thirds of "test performers."

Random number simulations (Fig. 9A) reveal how this graphical convention renders a pattern from random noise that entices researchers into describing it as a trait of human self-assessments. The significant negative correlations (Figs. 8A and 9A) that researchers interpret as the inverse relationship between performance and the degree of inaccuracy of self-assessment are maximized by random data (Fig. 9A). The closer a correlation derived from actual data approaches $r = -.7$ in the graphical convention of Figures 8A and 9A, the more the actual data resemble random noise.

Presenting such data aggregated into column graphs like Figures 7 and 8B conceals the nature of the data that were aggregated. This type of graph makes the artifact imposed by ceiling effects more difficult to discover. It seems best to avoid employing the convention shown in Figures 7, 8 and 9 in future research.

Collective Self-Assessments

Random number simulations (Figs. 5A and 9B) show that true random noise is as likely to contribute overestimation as underestimation to self-assessment measures. If the noise in actual measurements is mostly random, then, given a sufficiently large database, averaging collective data from all participants on every item should attenuate such noise and allow the signal-to-noise ratio to increase. Instruments like the KSSLCI and the SLCI that collect multiple measures that map to a single unifying construct offer an opportunity to do this on an item-by-item basis (Fig. 10).

The 1154 participants' item-by-item average ratings derived from the three-point scale of the KSSLCI proved similar to their average performance scores on the corresponding items of the SLCI (Fig. 10A). Item-by-item, the collective mean self-assessments of participants displayed a substantial positive relationship (at $r = .76$) to their mean collective performances (Fig. 10B). The signal-to-

noise ratio appears improved, indicating that the character of the noise present in our measures was largely random because averaging attenuated it.

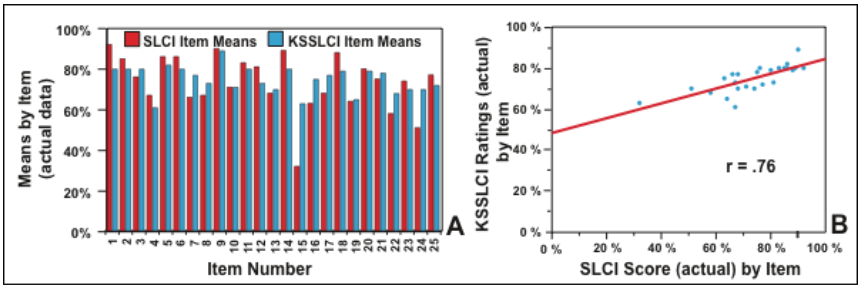

Figure 10. Means from 1154 participants of SLCI scores and KSSLCI self-assessment ratings on each item (Fig. 10A). Scatterplot of item-by-item average KSSLCI ratings *versus* item-by-item average SLCI scores reveals a strong correlation between collective self-assessment and collective performance (Fig. 10B).

For comparison, we replaced all 1154 participants' responses to each item on the SLCI and KSSLCI with appropriate random numbers (Fig. 11). In the case of the KSSLCI, we employed Excel's RANDBETWEEN(0,2) to randomize expressions of three levels of confidence. For the SLCI, we used RANDBETWEEN(0,1) to generate randomized correct (1) and incorrect (0) responses.

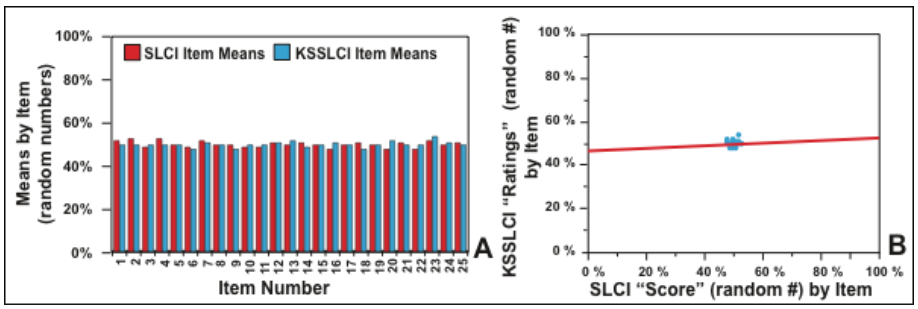

Figure 11. Simulated self-assessment ratings via the KSSLCI for 1154 "participants" and their simulated competency scores for the SLCI (A). Random numbers represent all KSSLCI and SLCI responses. Correlation between collective self-assessment and collective performance on an item-by-item basis is an insignificant $r = .06$ (B).

Figure 11A shows that when the measurements consist of purely random noise, the simulated knowledge survey ratings and inventory scores on every item become nearly identical. All converge, as expected, around the mean of 50%. The pattern in Figure 11B converges much more tightly around the point (50, 50) that marks the theoretical means than it does in Figure 1B. The tighter pattern results because the averages of 1154 randomly chosen scores and ratings bounded by 0 and 100 offer a higher probability of converging at the theoretical mean of 50 than do averages calculated from fewer ratings (12 and 13 in the case of Figure 1B). Figure 2B displays no such clustering around the mean because none of the plotted points derives from averaging.

Histograms of Self-Assessed Accuracy

Figure 10A indicates that both the KSSLCI and the SLCI instruments are measuring the same construct with scales of measures that are comparable. Therefore, it seems permissible to construct histograms of self-assessment accuracy based on differences (KSSLCI – SLCI) of the data yielded by the two instruments. The algorithm for constructing the histograms simply apportions the participants into intervals with increments of ten percentage-point (ppt) differences (Fig. 12).

Histograms seldom appear in the self-assessment literature because only large studies furnish the representative data needed to construct a meaningful histogram. Stinson and Xiaofeng (2008) employed data from 555 respondents to produce a histogram similar to our Figure 12.

Figure 12 reveals the distribution by categories of self-assessment accuracy across our 1154 participants. This representation permits disclosure of how greatly each subset of the participants errs in their accuracy of self-assessment and the proportions of the participants that populate each subset. The value of a histogram becomes apparent after considering the challenge of finding those with "good" self-assessment skills in Figure 2.

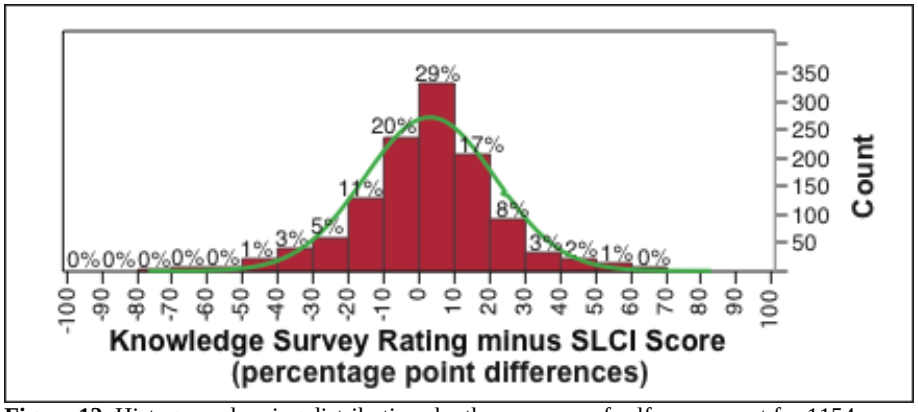

Figure 12. Histogram showing distributions by the accuracy of self-assessment for 1154 measures computed by the difference (KSSLCI-SLCI) in ten percentage-point intervals. Numbers above the intervals are the percent of the study population in each data range. Perfect self-assessment is zero (0).

We designated self-assessment accuracies within ±10% of zero as good self-assessments. We derived this designation from 69 professors self-assessing their competence, and 74% of them achieving accuracy within ±10%. On this basis, Figure 12 shows that 49% of participants achieved good self-assessment. Outside these bounds, 31% over-assessed, and 20% under-assessed.

If the distributions of 1154 participants were truly random across twenty categories, then the null hypothesis states that our 1154 participants should

distribute equally across each of the twenty intervals. Each interval would have about 58 respondents.

However, the histograms employed in self-assessment plot the differences between self-assessed ratings and test score, bounded by 0% and 100%. In computing the differences (the KSSLCI rating minus the SLCI score), there are 101 possible subtractions to generate zero (0) ppts, whereas there is only one possible subtraction that can generate either 100 ppts or –100 ppts. The probability of generating a value near zero is about two orders of magnitude greater than producing a value near the sides (±100) of these histograms.

Figure 13 is a histogram's depiction of pure noise produced by using the random numbers bounded by 0 and 100 that generated Figures 2B, 5B, 9, and 11. Because of the influence of probability, that pattern differs from the expected pattern of the null hypothesis. Because random noise is present in our actual measures, it is also an influence on Figure 12, just as it is on Figures 2A, 4, 8A, and 10.

A simple chi-square test[4] reveals that the distribution produced by random chance (Fig. 13) is statistically different from the null hypothesis at $p < .00001$. We can see from simulating our data with random numbers that Figure 13 is actually the expected distribution of 1154 random values bounded by 0 and 100. We can use the number of individuals in each of the intervals in Figure 13 as our expected values for a more informed new null hypothesis. A second chi-square test confirms that the distribution in Figure 12 is truly different from that in Figure 13 at $p < .00001$.

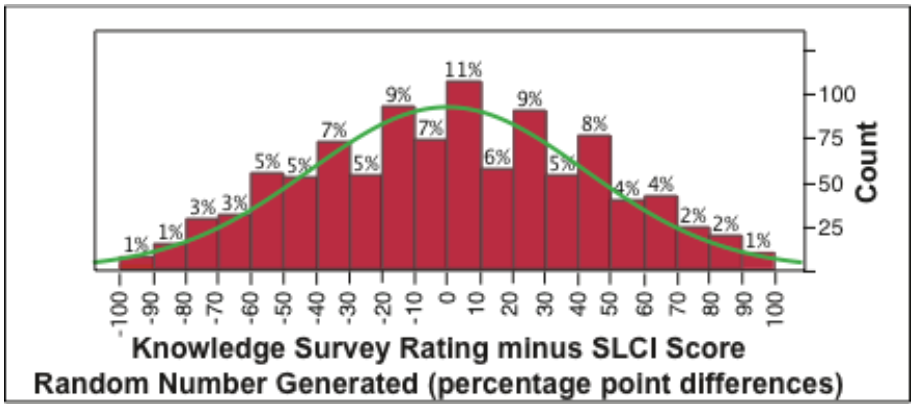

Figure 13. Histogram showing distributions of the accuracy of self-assessment computed from differences (KSSLCI-SLCI) in ten percentage point intervals. Here, the 1154 participants' responses were replaced by random numbers in increments of 4 ppts with bounds of 0 and 100 in both the KSSLCI and the SLCI. Numbers above the intervals are the percent of the population in each interval.

[4] Included in Excel workbook in Appendix A.

The pattern of Figure 13 contains no self-assessment signal. If one is unaware of what randomness looks like in a histogram, one might easily interpret Figure 13 as 18% of human participants' having good self-assessment skills, a general tendency toward accurately sensing their capabilities and no tendency toward overestimating or underestimating their competencies.

By graphing our actual data and comparing it to a random number simulation, we can recognize that Figure 12 shows 49% of participants having good self-assessment abilities. This result is greater than the 18% showing good self-assessment abilities that random chance produced (Fig. 13). The chi-square test shows that the distributions depicted in these two figures differ significantly from one another, which reveals that our real data carry a self-assessment signal that rises above the level of noise. If the data obtained from our actual study produced a distribution not significantly different from the distribution expected from random noise, further interpretation of the data would be meaningless.

Histograms appear to offer an informative presentation for displaying the distributions of participants' magnitudes and frequencies of self-assessment accuracy. Histograms that graph percentage-point differences of self-assessed performance and actual performance do portray random noise in patterns that can mimic those produced by real data. Lack of understanding the influence of random noise on histograms and how to determine when they portray a signal significantly different from noise can lead to unsound interpretations.

Conclusions

Random number simulations are useful for informing the collection of data, the graphing of data, and producing interpretations. They improve the understanding of self-assessment measures by revealing whether datasets are large enough, whether instruments produce reliable data, and whether graphical patterns express a meaningful self-assessment signal or primarily represent noise.

Of the graphical conventions that we studied, the (y) *versus* (x) scatterplots with a best-fit line that represent the self-assessment responses of each participant and the scatterplots that represent the participants' collective average responses item-by-item generated the fewest artifacts.

The problematic graphical conventions all employ calculated differences to produce the graphs. The most troublesome are the Bell-Volckmann and Pazicni-Bauer graphical conventions. These display raw data (scatterplots) and aggregated data (column charts) in $(y - x)$ *versus* (x) formats. In these, the influences of ceiling effects are so severe as to make these conventions untenable.

Construction of line charts drawn in the Kruger-Dunning convention does not require directly calculating differences. Instead, these line charts require users to estimate the differences from the distances between the lines in order to interpret

the graphs. This convention carries the influence of ceiling effects wherein the quantile containing the least-competent people overestimate their competency the most, simply because they can. The top quantile represents the most-competent participants who, by definition, simply cannot overestimate by as much. A strength of this convention is that it can reveal the signal-to-noise ratio of the measures and allow estimates of the critical size of a dataset needed to generate reproducible results.

Histograms of self-assessment accuracy also employ differences but in a more-constrained way. Unlike broad aggregates of thirds or quartiles, the histogram intervals group only those participants within a narrow range of self-assessment skill. Every interval of a histogram has the same range of percentage points (10 ppts in the case of Figs. 12 and 13). The portrayal of all such intervals together as a histogram generates a detailed picture that reduces the influence of the ceiling effect. However, histograms introduce a second illusory pattern wherein more participants may appear to have good self-assessment skills than is the case. This illusion occurs because the random noise present in imperfect measures of self-assessment imparts a strong probability toward producing a normal distribution centered at the value of perfect self-assessment. The degree to which histograms can factually represent actual self-assessment skills depends greatly on the signal-to-noise ratios in the measures. When the ratio is large, it may be the most informative way to portray human self-assessment. When the ratio is small, it may offer one of the most deceptive portrayals of any convention.

Not all numerical approaches employed to describe the relationship of self-assessed competence to actual competence are equally valid. Some numerical approaches do not offer valid descriptions of the relationship. Random number simulations allowed us to discover unanticipated idiosyncrasies associated with collecting data and describing the results of self-assessment measures. Performing such simulations should likewise be helpful to other investigators.

References

Baghaei, P. and N. Amrahi. 2011. The effects of the number of options on the psychometric characteristics of multiple-choice items. *Psychological Test and Assessment Modeling* 53: 192–211.

Bandura, A. 1997. *Self-efficacy: The exercise of control*. New York: Freeman.

Bell, P. and D. Volckmann. 2011. Knowledge surveys in general chemistry: Confidence, overconfidence, and performance. *Journal of Chemical Education* 88: 1469–1476. http://dx.doi.org/10.1021/ed100328c

Bowers, N., M. Brandon, and C. Hill. 2005. The use of a knowledge survey as an indicator of student learning in an introductory biology course. *Cell Biology Education* 4: 311–322. http://dx.doi.org/10.1187/cbe.04-11-0056

Caputo, D. and D. Dunning. 2005. What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology* 41: 488–505. http://dx.doi.org/10.1016/j.jesp.2004.09.006

Cashin, W. E. 1988. Student ratings of teaching: A summary of the research. *IDEA Technical Report No. 20*; Center for Faculty Evaluation and Development; Kansas State University: Manhattan, KS.

DeVellis, R. F. 2003. *Scale development: Theory and applications*. Los Angeles: Sage Publications.

Dunning, D. and E. G. Helzer. 2014. Beyond the correlation coefficient in studies of self-assessment accuracy: Commentary on Zell & Krizan. *Perspectives on Psychological Science* 9 (2): 126–130. http://dx.doi.org/10.1177/1745691614521244

Ehrlinger J., K. Johnson, M. Banner, D. Dunning, and J. Kruger. 2008. Why the unskilled are unaware: Further explorations of absent self-insight among the incompetent. *Organizational Behavior and Human Decision Processes* 105: 98–121. http://dx.doi.org/10.1016/j.obhdp.2007.05.002

Favazzo, L., J. D. Willford, and R. M. Watson. 2014. Correlating student knowledge and confidence using a graded knowledge survey to assess student learning in a general microbiology classroom. *Journal of Microbiology & Biology Education* 15 (2): 251–258. http://dx.doi.org/10.1128/jmbe.v15i2.693 (accessed October 18, 2015).

Gaze, E. 2014. Teaching quantitative reasoning: A better context for algebra. *Numeracy* 7 (1): Article 1. http://dx.doi.org/10.5038/1936-4660.7.1.1 (accessed October 18, 2015).

———, A. Montgomery, S. Kilic-Bahi, D. Leoni, L. Misener, and C. Taylor. 2014. Towards developing a quantitative literacy/reasoning assessment instrument. *Numeracy* 7 (2): Article 4. http://dx.doi.org/10.5038/1936-4660.7.2.4 (accessed October 18, 2015).

Gendall, P., and J. Hoek. 1990. A question of wording. *Marketing Bulletin* 1: 25–36.

Hyndman, R. J. and F. Yanan. 1996. Sample quantiles in statistical packages. *The American Statistician* 50 (4): 361–365. http://dx.doi.org/10.1080/00031305.1996.10473566

Jacobs L. C. and C. I. Chase. 1992. *Developing and using tests effectively: A guide for faculty*. San Francisco: Jossey-Bass.

Kennedy, E. J., L. Lawton, and E. L. Plumlee. 2002. Blissful ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education* 24 (3): 243–252. http://dx.doi.org/10.1177/0273475302238047

Kruger, J. and D. Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77: 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Landrum, R. E., J. R. Cashin, and K. S. Theis. 1993. More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement* 53: 771–778. http://dx.doi.org/10.1177/0013164493053003021

Lasry, N., S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef. 2011. The puzzling reliability of the Force Concept Inventory. *American Journal of Physics* 79 (9): 909–912. http://dx.doi.org/10.1119/1.3602073

McCormick, A. C., and K. McClenney. 2012. Will these trees ever bear fruit? A response to the special issue on student engagement. *Review of Higher Education* 35 (2): 307–333. http://dx.doi.org/10.1353/rhe.2012.0010

Mueller, S. T. and C. T. Weidemann. 2008. Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review* 15 (3): 465–494. http://dx.doi.org/10.3758/PBR.15.3.465

Nuhfer, E. B., and D. Knipp. 2003. The knowledge survey: A tool for all reasons. In *To improve the academy: Resources for faculty, instructional, and organizational development* 21, ed. C. M. Wehlburg and S. Chadwick-Blossey, 59–78. San Francisco: Jossey-Bass.

———. 2006. Re: The use of a knowledge survey as an indicator of student learning in an introductory biology course. *Life Sciences Education* 5 (4): 313–314. http://dx.doi.org/10.1187/cbe.06-05-0166 (accessed December 20, 2015).

Nuhfer, E., J. Clifford, C. Cogan, A. Goodman, C. Kloock, B. Stoeckly, C. Wheeler, G. Wood, and N. Zayas. 2010. Multi-campus project: Promoting and assessing science literacy in general education science courses: *California State University Institute for Teaching and Learning Connections* 3 (4). https://www.calstate.edu/itl/newsletter/10-summer.shtml (accessed October 18, 2015).

Pazicni, S. and C. F. Bauer. 2013. Characterizing illusions of competence in introductory chemistry students. *Chemistry Education Research and Practice* 15: 24–34. http://dx.doi.org/10.1039/C3RP00106G (accessed October 18, 2015).

Porter, S. R. 2012. Using student learning as a measure of quality in higher education. In *Context for Success: Measuring Colleges' Impact*; HCM Strategists: Washington DC, 2012. http://www.hcmstrategists.com/contextforsuccess/papers/PORTER_PAPER.pdf (accessed October 19, 2015).

———. 2013. Self-reported learning gains: A theory and test of college student survey response. *Research in Higher Education* 54 (1): 201–226. http://dx.doi.org/10.1007/s11162-012-9277-0

Rodriguez, M. C. 2005. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice* 24: 3–13. http://dx.doi.org/10.1111/j.1745-3992.2005.00006.x

Ross, J. A. 2006. The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation* 11 (10). http://pareonline.net/getvn.asp?v= 11&n= 10 (accessed October 19, 2015).

Stinson, T. A., and Z. Xiaofeng. 2008. Unmet expectations: Why is there such a difference between student expectations and classroom performance? *Journal of College Teaching & Learning* 5 (7): 33-42.

Zell, E., and Z. Krizan. 2014. Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science* 9 (2): 111–125. http://dx.doi.org/10.1177/1745691613518075