Numeracy
Advancing Education in Quantitative Literacy

Volume 10 | Issue 1 Article 4

2017

How Random Noise and a Graphical Convention Subverted Behavioral Scientists' Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives

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

Follow this and additional works at: https://scholarcommons.usf.edu/numeracy

Part of the Arts and Humanities Commons, Life Sciences Commons, Physical Sciences and Mathematics Commons, and the Social and Behavioral Sciences Commons

Recommended Citation

Nuhfer, Edward, Steven Fleisher, Christopher Cogan, Karl Wirth, and Eric Gaze. "How Random Noise and a Graphical Convention Subverted Behavioral Scientists' Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives." *Numeracy* 10, Iss. 1 (2017): Article 4. DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

Authors retain copyright of their material under a Creative Commons Non-Commercial Attribution 4.0 License.

How Random Noise and a Graphical Convention Subverted Behavioral Scientists' Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives

Abstract

Despite nearly two decades of research, researchers have not resolved whether people generally perceive their skills accurately or inaccurately. In this paper, we trace this lack of resolution to numeracy, specifically to the frequently overlooked complications that arise from the noisy data produced by the paired measures that researchers employ to determine self-assessment accuracy. To illustrate the complications and ways to resolve them, we employ a large dataset (N = 1154) obtained from paired measures of documented reliability to study self-assessed proficiency in science literacy. We collected demographic information that allowed both criterion-referenced and normative-based analyses of self-assessment data. We used these analyses to propose a quantitatively based classification scale and show how its use informs the nature of self-assessment. Much of the current consensus about peoples' inability to self-assess accurately comes from interpreting normative data presented in the Kruger-Dunning type graphical format or closely related (y - x) *vs.* (x) graphical conventions. Our data show that peoples' self-assessments of competence, in general, reflect a genuine competence that they can demonstrate. That finding contradicts the current consensus about the nature of self-assessment. Our results further confirm that experts are more proficient in self-assessing their abilities than novices and that women, in general, self-assess more accurately than men. The validity of interpretations of data depends strongly upon how carefully the researchers consider the numeracy that underlies graphical presentations and conclusions. Our results indicate that carefully measured self-assessments provide valid, measurable and valuable information about proficiency.

Keywords

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

Creative Commons License

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License

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.

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.

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.

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.

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.

This article is available in Numeracy: https://scholarcommons.usf.edu/numeracy/vol10/iss1/art4

Introduction

Measuring whether or not people are good judges of their abilities rests largely on numbers that result from simple arithmetic, namely subtraction. To quantify our abilities to self-assess accurately, we select a challenge, provide an estimate of our self-assessed ability to meet that challenge, and complete a direct measure of our competence in engaging the challenge. Obtaining our self-assessment accuracy requires nothing more than computing the difference between the two measures. While the computation could scarcely be simpler, the simplicity belies a surprisingly complex numeracy required to derive meaning from these numbers.

Demands for numeracy arise at every step of self-assessment research. Such numeracy enlists number sense, reading and interpreting graphs, basic probability and statistics, and reasoning. These concepts are emphasized on the Quantitative Literacy Reasoning Assessment (Gaze et al. 2014). The steps themselves include recognizing the assumptions involved in the paired measurements, preparing the data for analyses, presenting the data graphically, interpreting the patterns that the data produce on graphs, and deducing what these results reveal about our collective abilities to self-assess. Within these steps, inattention to numeracy produces (a) measures of undocumented reliability, (b) paired measures from poorly aligned instruments, (c) data from studies of insufficient size to achieve reliability or reproducibility, (d) data produced from vague questions, (e) failures to recognize ceiling and floor effects in paired data and (f) mistaking graphical patterns of random noise for patterns that depict the self-assessment signal (Nuhfer et al. 2016a).

This paper is our second in *Numeracy* that addresses the challenges of quantifying self-assessment. In our first paper (Nuhfer et al. 2016a), we focused on insights produced by considering self-assessment data as mixtures of signal and noise. The self-assessment signal manifests as a valid relationship between self-assessed ratings of competence and direct measures of competence, but the presence of noise interferes with detection of the sought-after signal, much as static interferes with clear radio reception.

Three competing hypotheses about self-assessment follow from that first paper. Proponents of the first hypothesis indeed do argue that measures of self-assessment yield meaningless nonsense.

1. No meaningful relationship exists between self-assessed competence and demonstrable competence. Self-assessed competence is mostly random noise. (Porter 2012, 2013).

This first hypothesis is arguably a null hypothesis to our second and third hypotheses, which are:

2. The relationship between self-assessed competence and demonstrable competence is meaningful and measurable. Studies confirm that people have a strong propensity toward overestimating their abilities. Those least competent have the greatest overconfidence in

their actual abilities. Those most competent tend toward accuracy or slight underconfidence in estimating their actual abilities. (Representative sources are Kruger and Dunning 1999; Kennedy et al. 2002; Ehrlinger et al. 2008; Stinson and Xiaofeng 2008; Bell and Volckmann 2011; Pazicni and Bauer 2013)

3. The relationship between self-assessed competence and demonstrable competence is meaningful and measurable. Some people exhibit significant overconfidence or underconfidence, but overall, people's self-assessed competence is in accord with a competence that they can demonstrate. (Ackerman, Beier and Bowen 2002; Nuhfer and Knipp 2006; Favazzo, Willford and Watson 2014; Handel and Fritzsche 2016; Nuhfer et al. 2016a; this paper)

In Nuhfer et al. (2016a), we ascertained that obtaining good measures of self-assessment ability requires great care. Studies done without such care produce questionable results, and such results, when published, contribute to beliefs that self-assessment is a nebulous human quality. After working to attend to the numeracy issues outlined above, we generated a dataset from which we could easily distinguish the numerical character of self-assessment measures from the character of randomness. Our results (Nuhfer et al. 2016a) required us to reject the first (null) hypothesis that consigned self-assessment to random noise.

In this paper, we determine which one of the remaining two hypotheses best explains human self-assessment. The prevalent consensus in the peer-reviewed literature supports the second of the three hypotheses. We trace the origins of this consensus to the seminal paper of Kruger and Dunning (1999), and we explain in this paper how eighteen years of replicating the procedures introduced in the founding paper have produced the prevalent consensus. Our study, however, shows merit in using alternative procedures, which, we have found, produce results that contradict the established consensus about the nature of human self-assessment.

To convey how attention to numeracy might eliminate misconceptions about self-assessment requires providing detailed explanations supported by examples. To offer a more concise report in this ("main") paper, we provide the explanations with examples in Appendix A. The omission of such explanations in earlier papers may account for the prolonged duration of misconceptions about self-assessment. To allow others to test our procedures and conclusions, we also share our dataset in a separate appendix (B). It augments the dataset shared in our first paper with some additional demographic information that we reserved for our completion of this study.

In this study, we are not so much disputing behavioral scientists' conclusions about the nature of self-assessment as we are questioning the numeracy that underlies these conclusions. Indeed, our greatest concern in questioning the numeracy is that readers might see our work as intentionally detracting from the pioneering contribution that Kruger and Dunning (1999) made to behavioral science. We note, however, that Kruger and Dunning (1999, p. 1132) clearly

anticipated the difficulties inherent in the area of study that they opened:

> Although we feel we have done a competent job in making a strong case for this analysis, studying it empirically, and drawing out relevant implications, our thesis leaves us with one haunting worry that we cannot vanquish. That worry is that this article may contain faulty logic, methodological errors, or poor communication.

Our navigating the numeracy of self-assessment measures revealed a path replete with unanticipated pitfalls and barriers. The trepidation that Kruger and Dunning expressed in 1999 remains shared by us in 2017. Yet, if our work confirms that earlier conclusions have less support from quantitative reasoning than investigators recognized, then justification exists for reevaluation of the consensus established from nearly two decades of self-assessment literature.

To provide continuity with our earlier paper, we employ the same dataset that we collected for Nuhfer et al. (2016a). Because we provided the methods section for collecting this data in the first paper, we do not repeat it here. Each of the 1154 participants in our study produced a measure of demonstrated competence in science literacy from his/her score on the Science Literacy Concept Inventory (SLCI, reliability R = .84) and a self-assessed competency rating to address this challenge through a knowledge survey of the Inventory (KSSLCI, R = .93). Both instruments furnish data that contain signal mixed with noise. We verified that the data had sufficient reliability to allow us to extract clear expressions of the signal from the noise (Nuhfer et al. 2016a).

We proceed next to clarify why measuring metacognitive self-assessment is worth the effort; distinguish between the several kinds of metacognitive self-assessments currently addressed by researchers; and visit considerations of what we are actually measuring. After that, we assess the influential Kruger-Dunning graphical presentation of self-assessment data and explain why we believe that future studies must employ alternative approaches.

Background

Why Measure Metacognitive Self-Assessment?

“Metacognition refers to one's knowledge concerning one's own cognitive processes or anything related to them…” (Flavell 1976, p.232). A primary aim of higher education is to produce graduates with abilities to increase their capacity for effective learning and thinking throughout their lives. Developing students' metacognitive self-assessment skills may be key to producing such graduates.

Self-assessment is a metacognitive skill that includes the capacity to assess accurately one's own ability to meet immediate cognitive and social challenges with present skills and knowledge. The exercise of self-assessment is more intuitive than analytical and occurs by accessing one's *affective* feeling of capacity

to meet the challenge.

Summaries of the history that have led to confirming the value of metacognition (Dunlosky and Metcalfe 2009) and the affective domain (Damasio 1999) reveal earlier periods during which behavioral scientists disrespected the two topics and regarded each as unworthy of serious study. Research eventually established both as important to learning. Self-assessment, however, which draws from metacognition and affect, remains viewed with suspicion. We concur with the observation of Zell and Krizan (2014) that continued disagreement still exists about whether people, in general, perceive their skills accurately or inaccurately.

Kruger and Dunning (1999) presented the first serious effort to quantify the accuracy of peoples' self-assessment. They concluded that relatively unskilled people suffer illusory superiority and mistakenly assess their abilities to be much higher than they are. Conversely, persons who demonstrate high ability accurately or modestly underestimate their competence.

Subsequent studies replicated Kruger and Dunning's results, and in less than a decade, many accepted that their results applied to the general populace, as typified by the following statement.

> People are typically overly optimistic when evaluating the quality of their performance on social and intellectual tasks. In particular, poor performers grossly overestimate their performances because their incompetence deprives them of the skills needed to recognize their deficits (Ehrlinger et al. 2008, p. 98).

A Web search for "Dunning-Kruger Effect" reveals that Kruger's and Dunning's discovery reached the lay populace where it engendered beliefs that people were mostly incapable of accurate self-assessment. At least one scholar went so far as to proclaim measures of self-assessed learning as meaningless noise (Porter, 2012; 2013). We note here that deprecating the value of self-assessment conflicts with the views expressed by Kruger and Dunning (1999) who recognized the value of self-assessment skill and documented that instruction could improve it.

Other workers furnished results that emphatically assigned value to metacognitive self-assessment. Ertmer and Newby (1996, p. 1) studied the characteristics of expert learners and listed these as "strategic, self-regulated, and reflective." All three characteristics have metacognitive qualities that we now recognize incorporate self-assessment. They further noted that expert learners use specific strategy "to deliberately select, control, and monitor strategies needed to achieve desired learning goals."

Isaacson and Fujita (2006, p. 39) confirmed the value of self-assessment when they deduced that the most successful college students possess metacognitive skills. Highly successful students were "more accurate at predicting their test results; more realistic in their goals; more likely to adjust their confidence in-line with their test results...." Dunlosky and Rawson (2012) offered

DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

related evidence that linked students overconfidence to their underachievement. Current researchers (Wittmann et al. 2016) identify specific areas of the brain activated during self-assessment, and they credit self-assessment to originating as one of the important human survival skills.

McMillan and Hearn (2008, p. 40) may be two of the strongest proponents for the educational value of developing students' ability to accurately self-assess:

> In the current era of standards-based education, student self-assessment stands alone in its promise of improved student motivation and engagement and learning. Correctly implemented, student self-assessment can promote intrinsic motivation, internally controlled effort, a mastery goal orientation, and more meaningful learning. Its powerful impact on student performance—in both classroom assessments and large-scale accountability assessments—empowers students to guide their own learning and internalize the criteria for judging success.

In summary, college instructors should measure self-assessment because the skill is valuable, measurable and teachable. Gaining self-assessment skill seems to increase the capacity for improved learning, problem-solving and decision-making. Improvement of students' self-assessment skill could be a universal educational outcome that transcends all disciplines.

Kinds of Self-Assessment

Scholars identify several kinds of self-assessment. Kruger and Dunning's (1999) seminal paper addressed participants' *predicted* ability to meet a cognitive challenge before confronting it. Scholars also refer to predicted abilities as "first-order judgments" (Dunlosky, Serra, Matvey and Rawson 2005). Kruger and Dunning (1999) also addressed results from a second kind of self-assessment subsequently termed "postdicted performance judgment" (Händel and Fritzsche 2016). In *postdicted* self-assessments, each participant expresses a summative estimate of how successfully she/he has addressed a recently completed cognitive challenge.

Kruger and Dunning, as well as later researchers, asked students to rate their relative performance on a test as compared to other participants' test scores. These estimates demand that the participants rate, not just self-assessed competence, but also the relative competence of others. There are conditions under which the competence and performance of other participants are available (Wittmann et al. 2016), but such was not the case in our study. In the absence of substantial information about other participants, estimates of self-competence relative to others seem based on little substance (see Hartwig and Dunlosky 2014). Self-assessed competence in the context of estimated comparisons with competence demonstrated by others might be registering each individual's relative sense of self-esteem rather than self-assessed competence.

Related research literature recognizes additional kinds of self-assessment. One is "*meta-metacognition*" or "second-order judgment." Here, participants

estimate the degree with which they have successfully provided an accurate self-evaluation of their performance to a cognitive challenge (Dunlosky, Serra, Matvey and Rawson 2005; Buratti and Allwood 2012). Another type of self-assessment described in some related literature is "*metacomprehension*" (Dunlosky and Lipko 2007), a term primarily employed in studies of reading skill. It refers to metacognitive awareness of readers about their learning and understanding produced while accessing text materials.

Self-assessment queries take global and granular forms. *Global* queries are singular statements that are broad and general. An example asks participants to rate their degree of competency in a broad conceptual area such as humor, critical thinking, writing or science in response to a query similar to "I understand… (humor, science, etc.)." Kruger and Dunning's original paper employed global queries, and so did most of the self-assessment studies such as Ehrlinger et al. (2008) and Pazicni and Bauer (2013) that subsequently built on Kruger and Dunning's work.

Granular self-assessment instruments employ a battery of specific items, all of which map to a broad conceptual area. Knowledge surveys (Nuhfer and Knipp 2003; Bell and Volckmann 2011; Favazzo et al. 2014; Nuhfer et al. 2016a), which ask respondents to estimate their ability to address many specific cognitive or skill challenges, constitute granular assessments. For example, the composite rating derived from all 25 items of the knowledge survey (KSSLCI) provides a granular self-assessment of the degree to which a participant understands science's way of knowing (Nuhfer 2015). As another example, about 200 items on a course-based knowledge survey might map to the general understanding of psychology or geology as provided by an introductory college course (Nuhfer et al. 2010).

Some workers treat self-assessments derived from global and granular queries as equivalent (Bell and Volckmann 2011). Our study participants furnished a total of four separate self-assessment ratings as registered by three global questions and the granular KSSLCI. When we compared global and granular self-assessments that addressed the same cognitive construct, our study revealed that some global queries could yield a different kind of self-assessment from that provided by granular instruments (Nuhfer et al. 2016a, Table 1; this paper, Appendix A, Fig. A1-6).

In this paper, we address predicted self-assessment and touch briefly on postdicted self-assessment. We do not address second-order-type judgments, metacomprehension, or any self-assessments that request that participants estimate their ability to address a challenge relative to others' abilities.

What Are We Measuring?

Since 1999, studies of self-assessment accuracy have typically employed *two* measures expressed as percentages or percentiles bounded by 0 and 100. *One*

DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

quantifies cognitive competence as expressed by a test score. *The other* manifests in responding to "*I can*… (do the specified challenge)… now, with my present knowledge and skills." Such responses express affective feelings. Such feelings can range from well informed to completely uninformed by cognitive knowledge and relevant experiences (Caputo and Dunning 2005).

By an informally accepted convention, researchers quantify self-assessment accuracy by *subtracting* the demonstrated competency score from the self-assessed competency rating. We follow that convention. By expressing both measures as percentages, the differences between paired measures register in *percentage points* (ppts). In our studies, we used the knowledge survey (KSSLCI) matched to the Concept Inventory (SLCI) to calculate accuracy:

$$Self\text{–}assessed\ Accuracy\ =\ KSSLCI\ rating\ -\ SLCI\ score.$$

By this procedure, perfect self-assessed accuracy is 0. Increasingly positive values denote increasing overconfidence. Increasingly negative values denote increasing underconfidence.

The act of computing self-assessment accuracy by subtraction assumes that we are calculating the difference between two measures with like qualities. Such subtractions begin to question the nature of distinctions often made between the cognitive and affective domains of thought and learning. We could view self-assessment accuracy as subtracting a direct score on a test of cognitive understanding from a quantified rating of affective feelings. We initially questioned what the remainder generated by subtracting a measure of competence from a measure of confidence expressed and whether the computation was justifiable.

Given the nature of our study, we opted to consider the calculation as justified by considering both measures as addressing the same competence in a well-defined area, one as a measurement and the other as an estimate. Our self-assessment instrument (KSSLCI) furnishes the estimate. It addresses the same construct as the cognitive competency measure (SLCI) because the 25 challenges employed in both are identical. The two instruments generate similar numerical results in percentage points for each of these challenges (Nuhfer et al. 2016a, Fig. 10; Nuhfer 2015, Fig. 1). Of course, one could argue the nature of such estimates as cognitive, affective or a combination of both.

Support for considering these as both comes from recognizing affective confidence and cognitive competence as two properties produced by multiple regions of the brain, which contribute cognitive and affective components to a common thought (Phan et al. 2004). If true, then it seems impractical to distinguish two properties of the same thought with separate units, such as we might do for the distinctly different physical properties of a physical object. Still, to invoke this justification for the subtraction *requires* two well-aligned

instruments. Studies that employ non-identical challenges or different constructs for queries of the confidence and competence measures risk computing self-assessment accuracy from subtracting two nonequivalent measures. Doing so offers an unsound basis for further interpretations or conclusions.

If separate units were viable, we might be able to quantify self-assessment through ratios of confidence-to-competence instead of through difference. In our early research, we experimented with trying to use ratios. However, measures of both expressed as percentages rendered the use of ratios impractical.

We turn next to examining how researchers' embracing of a common graphical convention may have produced the current consensus about the nature of self-assessment.

Influence of the Kruger-Dunning Graphical Convention

The numeracy associated with the Kruger-Dunning graphical convention (Fig. 1) is fundamental to understanding the prevalent consensus views about the nature of self-assessment. This convention constitutes the most influential graphic in the self-assessment literature, and many researchers from Kruger and Dunning (1999) through present (e.g., Miller and Geraci 2011; Handel and Fritzsche 2016) have employed it to portray their results and substantiate their conclusions.

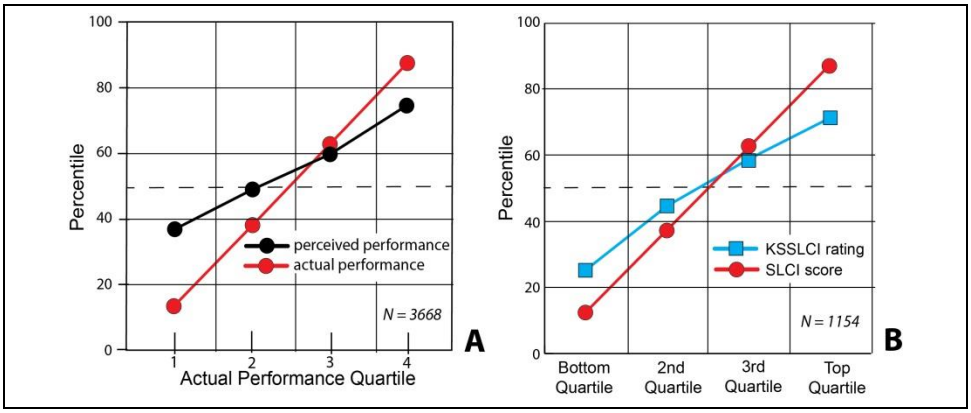

Figure 1. Self-assessment data rendered in the Kruger-Dunning graphical convention. The figure shows self-assessed competence compared with measured competence from two studies with two of the largest databases currently registered in the self-assessment literature. A, which is redrawn from Pazicni and Bauer (2013, Fig. 1), displays performance on a mid-term chemistry test and perceived performance obtained from a single global postdicted self-assessed rating of performance on the test. B displays actual competence as measured by performance on the 25-item Science Literacy Concept Inventory (SLCI) and anticipated performance computed as the average of self-assessment ratings from the 25-item knowledge survey of the Inventory (KSSLCI). The datasets employed for A and B both contain a strong self-assessment signal as shown by the steeply inclined perceived performance and KSSLCI rating lines (see Appendix A Fig. A1-6 and Nuhfer et al. 2016a, Fig. 5).

The published graphics consistently display the X-shaped patterns that show unskilled people as self-assessing their abilities to be much higher than they are. (See bottom quartiles in Fig. 1A and B.) This replication across many studies certainly provided confidence in the prevalent view that affirms a tendency for people who lack skill to overestimate their abilities.

The patterns usually show members of the top quartile as more accurately assessing their actual abilities, and tending toward underestimating their actual performance. Certainly, it is logical to expect that those with expertise in an area are in a much better position to accurately self-assess their abilities in that area than are those with little or no expertise. However, we raise two questions.

(1) Do the data depicted through the Kruger-Dunning convention offer sufficient quantitative evidence for confirming the expectation?

Our answer to this first question is "no." Our study caused us to realize that the Kruger-Dunning graph offers insufficient information needed for characterizing human self-assessment. Since 1999, assumptions based on interpretations made from that graph's characteristic patterns exemplified in Figure 1 have led to the current consensus view. We justify this "no" answer in detail in Appendix A, Part 1, where we address the following six overlooked aspects of numeracy on which such interpretations rest.

1. Random noise can generate X-shaped patterns in Kruger-Dunning-type graphs, and researchers can easily misinterpret these patterns as meaningful measures of self-assessment.
2. The Kruger-Dunning type graphs present patterns that appear meaningful from datasets too small to offer reliability.
3. In $(y - x)$ vs. (x) graphs, Sets of (x) and (y), both bounded by 0 and 100, generate strong ceiling and floor effects that researchers easily misinterpret as meaningful measures of self-assessment (addressed in Nuhfer et al. 2016a, Figures 7, 8 and 9).
4. Sorting data pairs by one member of the pair invariably produces the "X-shaped" pattern of Kruger-Dunning graphs and, sorting data by percentile rank renders all expressions of performance as norm-referenced rather than criterion-based.
5. Kruger-Dunning graphs cannot show the distributions of varied self-assessment skills in a populace.
6. Kruger-Dunning graphs fail to reveal the degree of correlation that exists between self-assessed competence and demonstrated competence on a participant-by-participant basis.

Artifact patterns generated by noise are particularly troublesome because they mimic those that researchers might reasonably expect as patterns produced from the self-assessment signal. This similarity of patterns generated by artifacts and expectations invites attributing the graphical patterns that random noise produces as patterns that describe the character of human self-assessment.

(2) Do the data in total from all studies best support the second or the third of the three hypotheses that we listed above?

Scholars established the prevalent view, which supports the second hypothesis, by interpreting patterns depicted by the Kruger-Dunning-type graph. If our data represented in Figure 1B followed the same conventional interpretation, then our study would also support the second hypothesis. However, we proceed next to explain why we portray and interpret our data differently. We will show how an analysis based on careful considerations of numeracy better supports the third hypothesis.

Results

Results from Categorical Data: Comparing Experts with Novices

The consensus that favors the second hypothesis rests upon the process of sorting participants' data by demonstrated competency scores in ascending order and constructing interpretations from a Kruger-Dunning-type graph like Figure 1. This approach is analogous to the norm-referenced practice of "grading on a curve," wherein participants gain access into the top quartile by being *relatively* more proficient than members of the lower quartiles.

This section describes a different approach. Here, we present a criterion-referenced study based on categories defined by qualifications of expertise to meet a cognitive challenge. The value of such an approach lies in avoiding reliance on numerically sorted data and gaining a way to study the degree to which the self-assessment characteristics of members of the top and bottom quartiles defined by norm-referenced scoring reflect the criterion-referenced characteristics that typify experts and novices.

Our categories consist of qualified novices (lower-division undergraduates), developing experts (upper-division undergraduates and graduate students) and experts with significant qualifications (professors). The SLCI measures cognitive competence in the ability to recognize and understand science as an evidence-based way of knowing, and knowing factual content did not advantage participants in this particular challenge (Nuhfer et al. 2016b). Our experts in this study became qualified as such through achieving advanced degrees that required demonstrable evidence-based reasoning. The mean *competence* values (SLCI score averages) calculated for each category of academic rank confirm highly significant differences in demonstrated competence between novices and experts (Appendix Part 2, Fig. A1-7).

The graphical convention employed in this section to display the categorical data (Fig. 2) offers more information than a Kruger-Dunning graph like Figure

DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

1B. Figure 2 shows the confidence intervals of the means of each category, the significance of differences between these means, and the spread or range of variance of participants within each category. The different results that come from graphing unsorted categorical data (Fig. 2) and sorted data (Fig. 1B) account for the two graphical conventions offering a basis for two contradictory interpretations. Opting to present the data as percentiles (Fig. 2) or as raw scores in percentages (Fig. 3) further complicates the interpretations. Researchers employ Kruger-Dunning-type graphs that present data as either percentages or percentiles to render interpretations (see Ehrlinger et al. 2008 for examples), but the use of percentiles is prevalent.

The figures that display categorical data (Figs. 2, 3 and 5) have dimensionless abscissas that simply plot the $(y - x)$ scores by categories. This yields a graph of the form $(y - x)$ *vs. categories*. The norm-referenced data aggregated by quartiles appears in Figures A1-2, A1-3, and A1-5 in the Appendix. These latter figures have scaled abscissas that display increasing SLCI scores, which places these graphs in the category of $(y - x)$ *vs*. (x) formats that we noted (Nuhfer et al. 2016a) as particularly troublesome because they generate severe ceiling effects.

The prevalent consensus in the self-assessment literature asserts that the people who are most lacking in competence are those who most severely overestimate their abilities, whereas people who possess the greatest competence are more accurate in their estimates and usually tend to underestimate their competence by modest amounts. Researchers (typified by Burson, Larrick and Klayman 2006; Ehrlinger et al. 2008; Bell and Volckmann 2011; Pazicni and Bauer 2013) corroborate that assertion through displaying their data in the Kruger-Dunning type graphs.

From the patterns presented in Kruger-Dunning-type graphs (Fig. 1) and the prevalent consensus derived from such graphs, we expected that the average self-assessments of confirmed novices would exhibit a pronounced overestimation of abilities and be less accurate as a whole than the average self-assessment of confirmed experts. However, the mean self-assessment accuracies (as registered by $KSSLCI\ rating - SLCI\ score$) differ little across the categories of academic rank (Figs. 2 and 3). Figure 2 shows the mean estimates of all academic ranks as plotting close to the perfect self-assessment value of $KSSLCI\ rating - SLCI\ score = 0$. In this graphical presentation, experts even appeared less accurate in their collective self-assessments than did novices (Fig. 2), although this appearance could be a product of some floor and ceiling effects, as discussed further below.

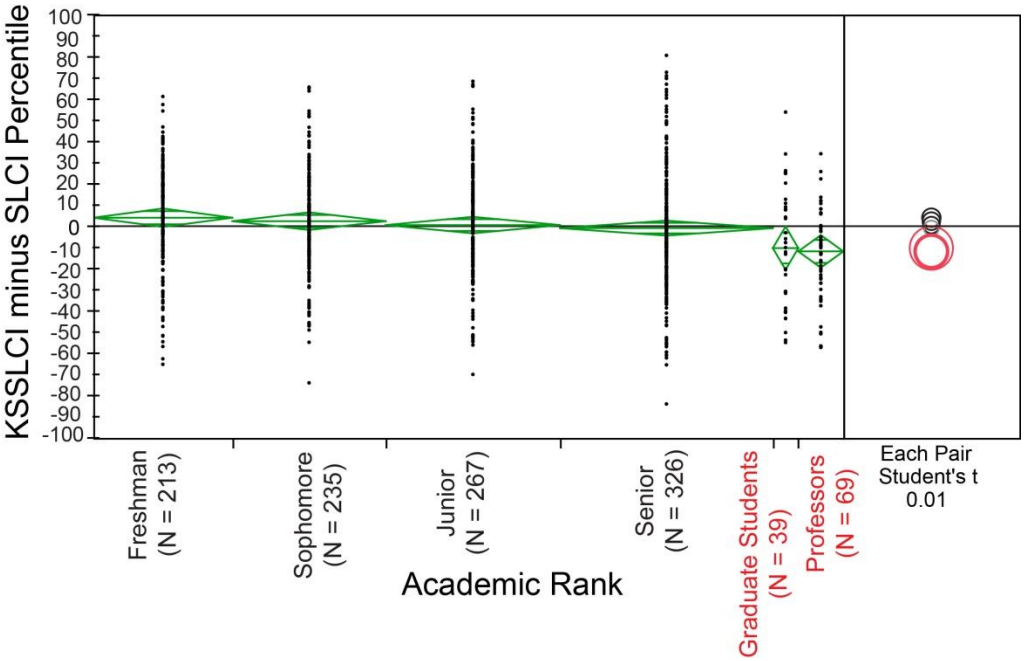

Figure 2. Categorical self-assessment accuracies plotted as percentiles. The abscissa is dimensionless and simply displays self-assessment accuracies by academic rank. Black dots show the respondents' distributions of accuracy expressed in percentile-rank differences in each academic rank. The height of the green diamonds reflects the bounds of the 99% confidence level of the mean; width of diamonds reflects the numbers in each rank category. Box to the right depicts the significant differences between ranks as expressed by t-testing. Diameters of the circles are the bounds of the 99% confidence interval. Separation of circles shows that the means of professors and the means of graduate students differ significantly from those of undergraduates. Overlapping of circles reflects a lack of significant differences between undergraduate ranks. Graph produced by SAS Institute's JMP 11.2 software.

In Figures 2 and 3, the data points plotted for each category reveal that members of each academic rank tend to overestimate and underestimate with similar frequency. This accounts for the category means of ($KSSLCI\ rating - SLCI\ score$) all being close to zero. In Figure 1, the members of each quartile also overestimate and underestimate with about the same *frequency* (see Appendix Fig. A1-2 for supporting evidence), but the clustering of all of the lowest scores in the bottom quartile dictates that the probability for larger *magnitudes* of over-assessment are greater for members of the bottom quartile. For those in the top quartile, the probability for larger magnitudes of under-assessment is greater. Thus, the calculated mean self-assessment inaccuracies are highest in the bottom quartile and lowest in the top quartile, but that's because of the probability situation and not because of a quality inherent to human self-assessment. This situation produces the ceiling and floor effects mentioned above (described in more detail in Nuhfer et al. 2016a) and in the Appendix of this

paper. These findings show that the means of the quartiles are not useful for distinguishing human self-assessment abilities.

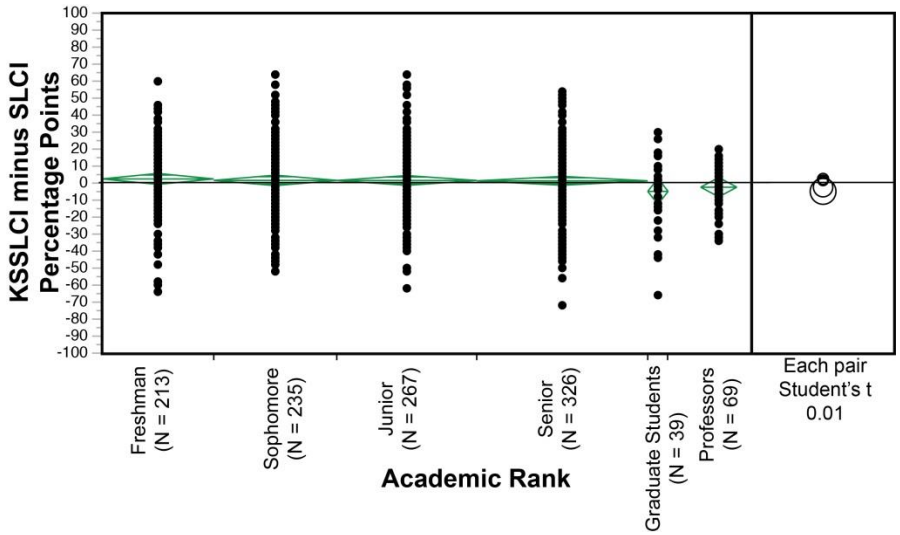

Figure 3. Degrees of self-assessment accuracy plotted as percentage points. The abscissa is dimensionless and simply displays self-assessment accuracies by academic rank. The black dots show the distributions of responses in ppts in each rank. The height of the green diamonds reflects the bounds of the 99% confidence level of the mean; width of diamonds reflects the numbers in each rank category. Panel to the right depicts the significant differences between ranks as expressed by t-testing. Diameters of the circles in the right-hand panel are the bounds of the 99% confidence interval. Overlapping of these circles reflects a lack of significant differences between means. Graph produced by SAS Institute's JMP 11.2 software.

The means of the different categories likewise seem unsuitable for distinguishing differences in self-assessment skills between categories (Fig. 3). On average, novices (N = 448) overestimated their competence on the SLCI by 2.1 ppts, and experts (N = 69) underestimated theirs by 2.4 ppts. The influence of ceiling and floor effects could contribute to these small differences. Although the presence of such effects does not completely rule out the possibility of tendencies for novices to overestimate and experts to underestimate, our particular dataset indicates that such tendencies, if they exist, are weak.

The compositions of the bottom and top quartiles in Figure1B do reflect a systematic distribution of novices and experts. In Figure 1B, the bottom quartile contains 61.3% novices, 38.7% developing experts, and 0% experts. In contrast, the top quartile contains 17.4% novices, 64.8% developing experts, and 17.8% experts. Of those in the expert category (professors), 74% of them ended up in the top quartile, whereas only 11% of novices reached the top quartile.

Whereas Figure 1B indicates clear differences in mean self-assessment accuracies between low competence and high-competence quartiles, Figure 2

indicates lesser, marginally significant differences in mean self-assessment accuracies between novices and experts. With the data expressed simply as raw percentage points (Fig. 3), these differences become smaller still and lose significance.

Both Figures 2 and 3 are technically $(y - x)$ vs. (*categories*) type instead of the $(y - x)$ *vs.* (x) type graphs that we (Nuhfer et al. 2016a) showed as significant generators of ceiling effects. However, Figures 2 and 3 may carry some ceiling and floor effects (although less so than those shown in $(y - x)$ *vs.* (x) type graphs) because the high scores achieved by participants in the expert category (professors in Figs. 2 and 3) leave a limited potential for overestimation.

Figure 3 allows us to begin to see a difference between experts and novices in the vertical spreads (variances) of the data points furnished by the populations within each category. The spreads are less evident in Figure 2 because converting raw scores into percentiles orders the data, and this ordering redistributes any skewed distributions of scores toward normal distributions (Fig. 4).

Figure 4 provides detailed comparisons of the spreads by rank when expressing the data either as percentiles (Fig. 2) or as percentage points (Fig. 3). The graphing as percentage points discloses that experts exhibit smaller spreads in their scores than do novices, and experts' self-assessment accuracies cluster more tightly around perfect self-assessment (Fig. 4).

In our first paper, we showed the importance of recognizing the patterns of randomness in various graphical formats (Nuhfer et al. 2016a). Here, Figure 5 displays the pattern of randomness across the categories as rendered by the graphical convention that produced Figures 2 and 3 from actual measurements. The number of participants in each rank in Figure 3 determined the size of the random number array that we employed to simulate each rank in Figure 5.

Figure 5, like its real data counterpart (Fig. 3), displays all ranks as having mean self-assessment accuracies close to that of the perfect self-assessment score of 0, with no significant differences in means between ranks. In our actual data (Figs. 2 and 3) and simulated data (Fig. 5), members of all categories tend to overestimate and underestimate to about the same degree. This tendency produces mean self-assessed competencies across all academic ranks at close to the perfect self-assessment value of zero.

Whereas the categories' mean accuracies are all close to zero in Figure 5, the sorting of random number data by competence scores and aggregating it into quartiles produces quartiles whose mean (*KSSLCI rating - SLCI score*) values differ greatly and systematically from one another. The comparisons of Figures 5 and Appendix Figure A1-5 show the power of random noise to influence the graphical patterns produced by sorted data. The convention employed in Figures A1-2, A1-3 and A1-5 and the Kruger-Dunning-type convention both yield patterns that are particularly prone to the influences of noise and sorting.

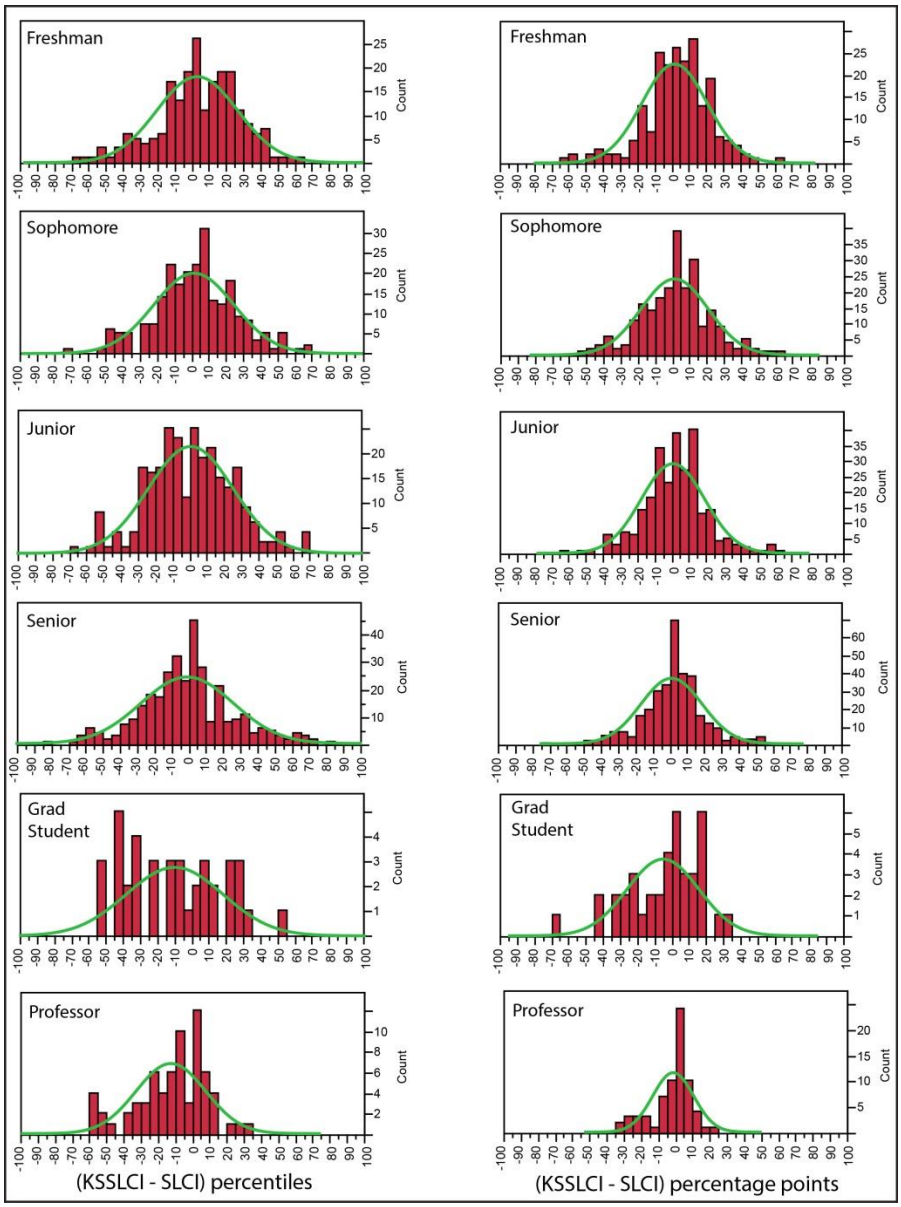

Figure 4. Distributions of participants within each academic rank as expressed in percentiles from Figure 2 and as percentage points from Figure 3. Left column details the spreads in Figure 2; right column details the spreads in Figures 3. Raw data in percentage points shows a general tightening of spreads from novice to experts, whereas data normalized when expressed as percentiles obscure this trend. The category of graduate students contains too few participants to yield a good representation and contains a much higher percentage of non-science majors than do the other categories. Appendix A Part 3 further details how good data does reveal ways to distinguish differences between experts and novices in self-assessment skills.

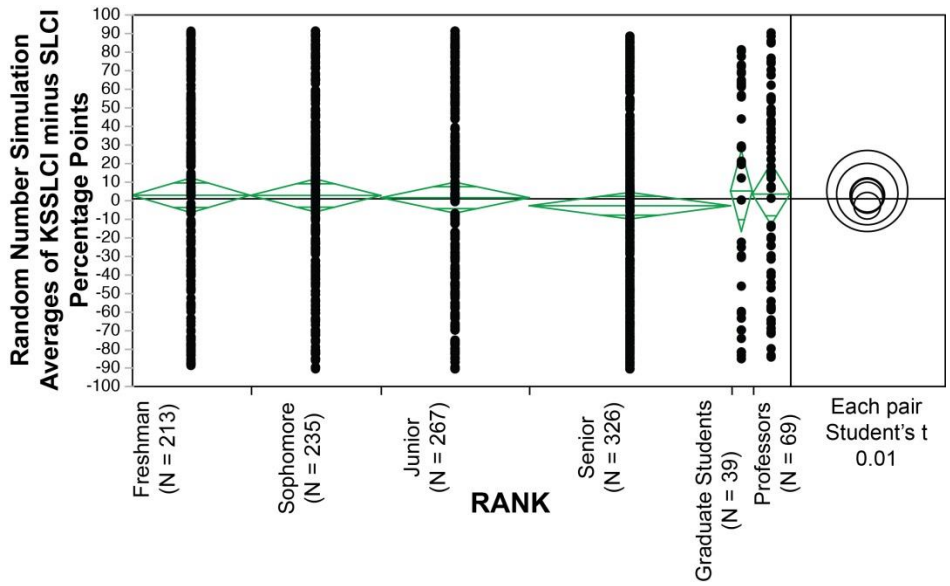

Figure 5. Random number simulation of self-assessment accuracy and the distributions of responses by academic rank. The height of the green diamonds reflects the bounds of the 99% confidence level; width of the diamonds reflects the numbers in each rank category. Diameters of the circles in the right panel are the bounds of the 99% confidence interval. Overlapping of circles reflects no significant differences between means by t-testing. This pattern produced by the aggregation of data by categories differs greatly from the pattern yielded by a similar simulation of sorted data aggregated by quartiles (see Appendix Fig. A1-5). Graph generated using SAS Institute's JMP 11.2 software.

As we noted in Nuhfer et al. (2016a), the graphical convention that seems least troublesome for a straightforward presentation of self-assessment data is the (y) vs. (x) scatter plot with a line fit. We show our comparisons between experts and novices through this convention in Figure 6.

Taken alone, correlation coefficients of self-assessed competence *versus* demonstrated competence revealed little difference between experts and novices (Fig. 6). Both r-values are highly significant at p <.0001 but not much different from each other or from the correlation established from the entire population studied (r = .60; N = 1154; Nuhfer et al. 2016a). This substantiates the assertion of Ackerman and Wolman (2007, p. 58):

> Thus, although the mean correlations between self-estimates of ability and objective ability measures are modest in magnitude, it appears that substantial gains in correspondence can be obtained when specific measurement conditions are met.

The significant positive correlations indicate that people as a whole, whether experts or novices, tend to self-assess their competence to the degree that is generally correct. We stress that the ability to perceive this relationship rests in

collecting a critical mass of reliable data from instruments that are well aligned (Nuhfer 2015; Nuhfer et al. 2016a).

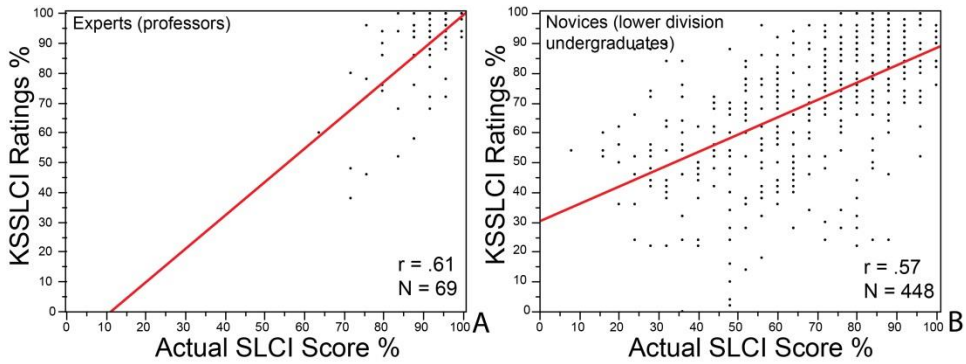

Figure 6. Comparisons of correlations between experts (A) and novices (B) in our study populace. Correlation coefficients are surprisingly similar.

In summary, our results show that a few novices tend to score as highly as experts on tests of competence (SLCI). Those who do will end up in the top quartile together with most of the experts in a norm-referenced study. However, experts' self-assessments show less variation than those of novices and are more consistently closer to perfect accuracy than are those of novices. Because novices do differ from experts in both competency and self-assessment accuracy, the top quartile in a norm-referenced study is not synonymous with the expert category in a criterion-referenced study. The categorical criterion-referenced study detailed here appeared to provide better information about the characteristics of self-assessment than did the norm-referenced study detailed in the Appendix.

Results from Demographic Data

In our study, we looked at other demographic data beyond class rank. We conclude this section by summarizing our findings in the groups of students with respect to 1) English as a first language; 2) status as a first generation student; 3) status as a science major or expressed interest to major in science and 4) gender. Nuhfer et al. (2016b) reported the results of the demonstrated competency (SLCI scores) from over 17,000 undergraduate students across these same four categories. The study verified significant differences in mean competence at the 99.9% confidence levels within the first three categories and no significant difference between men and women.

Here, we focus solely on the self-assessment characteristics of our undergraduate participants, which consist of 664 women and 371 men distributed as 213 freshmen, 235 sophomores, 267 juniors and 326 seniors. This population had the demographic distributions of 432 (41.5%) first-generation students, 712

(68.4%) students majoring in or considering majoring in science, and 162 (15.6%) students whose native language was not English.

Table 1.
Four Mean Self-Assessed Competency Rating Measures and One Demonstrated Competency Score Measure by Demographic Category*

	PRE-KSSLCI GLOBAL Rating (%)	KSSLCI Rating (%)	POST-KSSLCI GLOBAL Rating (%)	SLCI Score (%)	**GLOBAL POST-SLCI Rating (%)	MEAN (KSSLCI-SLCI) (ppts)
First Generation Student?						
No (*n* = 603)	77.1%	77.7%	76.7%	75.2%	78.9%	2.5
Yes (*n* = 432)	72.0%	68.6%	67.1%	68.0%	70.1%	0.6
Science major Commitment?						
No (*n* = 329)	71.0%	63.9%	65.8%	63.9%	64.7%	0.0
Yes (*n* = 712)	76.8%	78.6%	76.0%	76.1%	77.0%	2.5
English as First Language?						
No (*n* = 160)	70.1%	61.7%	61.9%	63.5%	66.4%	-1.8
Yes (*n* = 879)	75.9%	76.1%	74.7%	73.8%	76.5%	2.3
Gender						
Women (*n* = 664)	72.9%	70.3%	69.4%	70.6%	71.6%	-0.3
Men (*n* = 371)	78.9%	80.3%	78.6%	75.4%	80.3%	5.0

* Mean ratings and scores (in percent) from different self-assessment measures employed in the self-assessment studies reported by demographic categories. The differences within every category are significant at or above the 95% confidence level. We express self-assessment accuracy as the difference ($KSSLCI - SLCI$) calculated as the means of all students in each category. Perfect accuracy is expressed by $KSSLCI - SLCI = 0$.
**Our adding the Post-SLCI Global self-assessment query later in the study caused us to collect fewer responses.

Table 1 displays the results of measures across the different demographic categories in the order in which the participants responded to the four self-assessed competency ratings that follow.

1. **Pre-KSSLCI Global Rating:** "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?"

2. **KSSLCI Knowledge Survey**: This granular self-assessment value derives from the cumulative rating in % derived from all 25 items in the KSSLCI.

3. **Post-KSSLCI Global:** "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?"
4. **Post-SLCI Global** "Now that you have completed taking the Inventory, what score in percent (between 0% and 100%) do you think you actually obtained?"

The first is a predicted self-assessment; the knowledge survey is a granular self-assessment, and the third and fourth items are postdicted global self-assessments. See Appendix Figure A1-6 and its discussion for more details on the relationships between these self-assessed competency ratings and the demonstrated competency score relative to the Kruger-Dunning graphic.

Table 1 reveals a slight "reverse Dunning-Kruger Effect." The groups who are advantaged by having a major interest in science, a college-educated parent and English as a native language do have higher mean competency scores (see also Nuhfer et al. 2016b), but these advantaged subgroups tend toward being slightly less accurate in self-assessment than their disadvantaged counterparts. The differences between first-generation students and those who were not first-generation proved significant at only the 95% confidence level. The differences exhibited in mean confidence ratings within all other demographic categories were significant at the 99% confidence level.

One aberration in Table 1 was the finding of significant differences in the SLCI scores between men and women in this dataset at the 99% level of confidence. The larger 17,000-participant dataset that validated the SLCI (Nuhfer et al. 2016b) confirmed that the SLCI is a gender-neutral instrument. That study revealed that when the difference between men's and women's SLCI scores proves significant in a population, the difference was not produced by an inherent gender characteristic. Instead, the differences arose because of the unequal distribution between genders of the socioeconomic factors that diminish the mean scores on the SLCI. Socioeconomic factors that reduce mean SLCI scores of a populace are (a) status as a first-generation student, (b) a low interest in majoring in science, and (c) having English as a non-native language (see Nuhfer et al. 2016b).

In the dataset used for Table 1, the percentages of undergraduate women (N = 664) who are first-generation/nonscience-commitment/English-as-non-native-language are 45.2%/35.5%/17.6%. By comparison, undergraduate men (N = 371) in this dataset have only 35.0%/24.3%/11.3% membership in these respective categories. These socioeconomic differences in the composition of each gender populace substantially elevate the men's mean score above the women's mean score in our studied population of undergraduates.

Although men and women do not significantly differ in their science literacy competence as measured by the SLCI (Nuhfer et al. 2016b), men and women do seem to differ significantly in mean self-assessment accuracy. In this study, the

group mean of undergraduate women underestimated their performance by only 0.3 ppts. The group of undergraduate men overestimated their actual performance by a mean of about 5 ppts (Table 1). This difference in means is highly significant at the 99.9% level of confidence.

Kruger and Dunning (1999, p. 1123) considered gender differences in self-assessment skill and reported: "Gender failed to qualify any results in this or any of the studies reported in this article…." However, subsequent studies (Hargittai and Shafer 2006; Pazicni and Bauer 2013; Bolívar-Cruz, Verano-Tacoronte and González-Betancor 2015) report gender differences in self-assessment abilities that are consistent with ours. Our data showed that, on average, women self-assess their competence more accurately than do men. We consider the other demographic differences listed in Table 1 as too small and tentative to try to interpret, but the gender difference in self-assessment ability appears substantial.

Some scholars suggest that women's underconfidence in science (relative to men's) may be discouraging women to major in science (Beyer, Rynes and Haller 2004; Cech, Rubineau, Silbey and Seron 2011), and they recommend taking action to boost women's confidence to that of men's. However, those studies did not consider self-assessment accuracy, and self-assessment accuracy probably has more value than overconfidence. Men appear to be in greater need of training in metacognitive self-assessment than women.

Summary of Results

Categorical data enables criterion-referenced examination of the nature of human self-assessment in ways that normative-based analyses cannot. The means of *demonstrated competence* (Appendix A Fig. A1-7) clearly do reflect the immense differences between experts and novices. However, the means of *self-assessment accuracies* clearly do not distinguish the self-assessment skills of novices from experts (Figs. 2, 3 and 5). Correlations between self-assessed competence and actual competence do not serve as a key to distinguish experts from novices (Fig. 6), but they indicate that people, in general, are more often correct than not in estimating their competencies.

Kruger-Dunning-type graphs (Fig. 1) rely on sorted data for calculating the means of self-assessed competence and demonstrated competence for each of the competency quartiles. Researchers then use differences between the paired measures displayed on graphical patterns to make conclusions about the self-assessment abilities of low-competence performers and high-competence performers. These conclusions support the second hypothesis. Random noise present in all self-assessment data, combined with ceiling and floor effects, also offer graphical patterns anticipated by the second hypothesis. These latter patterns have no origins in human behavior, but they seduce researchers into interpreting them as such.

DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

The clearest distinction between the self-assessment skills of experts and novices seems to lie in their different distributions of self-assessment accuracy (Fig. 4), but the self-assessment literature rarely employs graphical conventions that can display distributions. We next move to discuss ways in which researchers might use the paired measures of self-assessed competence to illuminate the nature of human self-assessment.

Discussion

Improving the Discourse about Self-Assessment Skill

Since 1999, showing the patterns from Kruger-Dunning-type graphics and related $(y - x)$ *vs.* (x) type graphs (Nuhfer et al. 2016a) remained the default for communicating the nature of self-assessment. While the information this graphic provides is both limited and probably distorted, such graphics remain a cornerstone for statements such as "*People are typically overly optimistic …,*" and "*In particular, poor performers grossly overestimate…*" (Ehrlinger et al. 2008, p. 98).

The grand mean SLCI score of our 1154 participants is 73.6%, and the grand mean KSSLCI rating is 74.8%. Given the imperfect reliability of both instruments, the apparent overconfidence of 1.2 ppts is too small to invoke as support for any hypothesis that asserts that people have a marked propensity to overestimate their abilities. Handel and Fritzsche (2016, p. 233) also found only a slight overall inaccuracy in their studied populace but as a small underestimate rather than an overestimate.

As established above in our discussion of Kruger-Dunning-type graphs, the numeracy traditionally employed to support claims of gross overestimation seems insufficient. Such graphs (Fig. 1 A and B) are incapable of imparting meaning to discussions that employ descriptions such as "overly optimistic" or "grossly" because such descriptors lack quantitative meaning. The self-assessment literature's neglect to furnish the language needed for better discourse furnishes a barrier to the most basic discussions—even about "good" or "poor" self-assessment accuracy.

Supplying the minimal language needed to advance discourse requires answering two essential, quantitative questions. The first question speaks to the value of measuring self-assessment.

1. What magnitude of self-assessment error is permissible for a person who is "skilled" in self-assessment?

To address this first question, we can look to the magnitudes of self-assessment error that typify a population of experts. The second question directly addresses

whether data obtained from a general populace better supports the second or the third hypothesis.

2. What is the frequency of occurrence of varied degrees of self-assessment errors (expressed as a percentage) across a large population?

For education, answering both questions enables discussion about acceptable levels of self-assessment skill and achieving some consensus on when a level of skill is so deficient as to merit efforts for remediation. To furnish the required language, we employ the same data that produced Figure 1B to generate a classification scale (Fig. 7A) that enables characterizing our study populace (Fig. 7B) with categories defined by quantitative bounds. Using our data in this way addresses both questions.

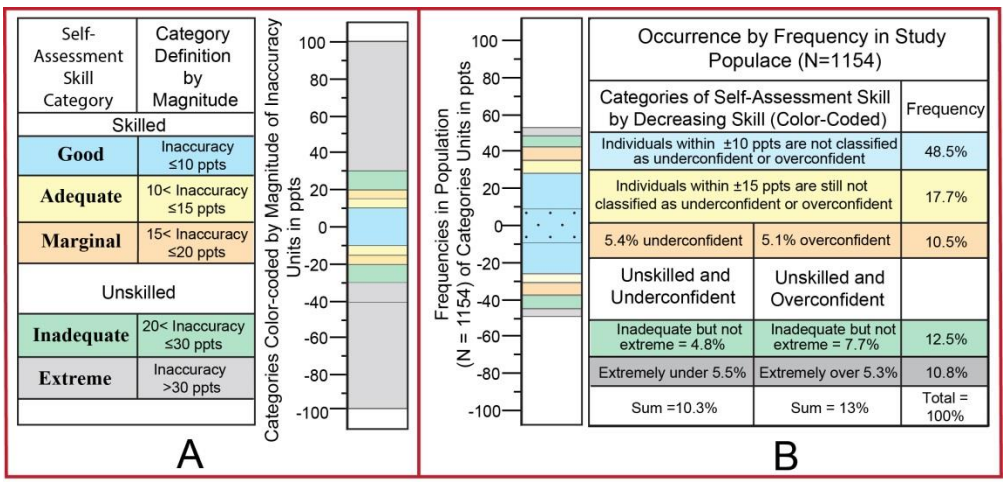

Figure 7. A classification scale (A) and its application to our study populace (B). Magnitudes of self-assessment inaccuracy ($KSSLCI\ rating - SLCI\ score$) expressed in percentage points (ppts) define the classification categories (A) The frequencies of the occurrences of these categories in our study population appear in B. The panels depict results by both tables and graphics. The blue shaded area with dots in B expresses our recognizing (Nuhfer et al. 2016a) that random guessing by *all* participants could contribute up to about 18% within the "good" range of ±10 ppts. The chances of guessing influencing the "Extreme" category are very small.

As detailed in Nuhfer et al. (2016a) the limit imposed by the instrument that yields the least reliable measures in paired data (in this case the SLCI's *R* of .84) limits the strength of correlation possible between the measures. It also limits the precision with which we can expect to define boundaries between the different skill categories in Figure 7A. While the boundaries are set at convenient intervals of 10 ppts, 20 ppts, etc., they are not arbitrary. The criterion-referenced performance of known groups of experts and novices in our study populace served to set these boundaries (see Appendix A, Part 3).

We earlier defined "good self-assessment skill" as demonstrating self-assessed competency within ±10 percentage points (ppts) of demonstrated proficiency, based on our discerning that over three-quarters of known experts could self-assess at this level of proficiency (Nuhfer et al. 2016a, p 19). Of our 1154 participants who range from novices to experts, 615 or 48.5% of those participants met the criteria for having good self-assessment skill (Fig. 7B). About 80% of experts self-assess within the bounds of ±15 ppts defined as "adequate self-assessment skill." This zone (Fig. 7A) accounts for 66.2% of our participants who demonstrated adequate or better self-assessment skills (Fig. 7B).

The distinction between adequate and inadequate self-assessment is an important one because scores that cross the boundary into "inadequate" can trigger investments in remediation efforts. Given this initial effort at a proposed classification scale and the realization that our instruments are reliable but imperfect, we sought not to set a dogmatic boundary between the two. Instead, we designated a ±5 ppt band between skilled and unskilled (between ±15 and ±20 ppts) self-assessments as "Marginal" (Fig. 7A). This choice allows users flexibility to make an informed evaluation of the state of the self-assessment skills of their own students.

Based on our work to date, we inform students that self-assessments in which error exceeds ± 20 ppts can indicate a need for efforts at developing better self-assessment skill. Participants with marginal self-assessment skills constituted 10.5% of our study populace. Errors of overconfidence or underconfidence that exceeded "marginal" (± 20 ppts) occurred in 23.3% of our participants. Of these (Fig. 7A), 13% overestimated and 10.3% underestimated (Fig. 7B).

The extreme categories (defined by inaccuracy exceeding 30 ppts) constituted only 10.8 % of our studied population (Fig. 7B). Less than half of them (5.3%) were extremely overconfident and constituted a group that could merit the label coined by Kruger and Dunning (1999), "unskilled and unaware of it." Figure 8 details the distributions of our populace across the defined categories and adds clarity to information conveyed by Figure 1B.

In histograms like Figure 8, random guessing has about one hundred times the influence near the center of the histogram, where ($KSSLCI\ rating - SLCI\ score$) is zero, than it has on the sides where self-assessments are "Extreme" (see Nuhfer et al. 2016a, Fig. 13 for detailed explanation). If *all* 1154 participants were randomly guessing, that would have placed over 200 scores in the "good" (blue) zone of Figure 8. Fortunately, the study of over 17,000 students who took the SLCI (Nuhfer et al. 2016b, Fig. 1) shows that the numbers of participants who engage in random guessing on the SLCI contributes much less than 18% of "Good" ratings in Figure 8, and almost nothing in the "Extreme" zones. While some guessing doubtless occurs in our dataset, its influence on our Figures 7B and 8 appears minor.

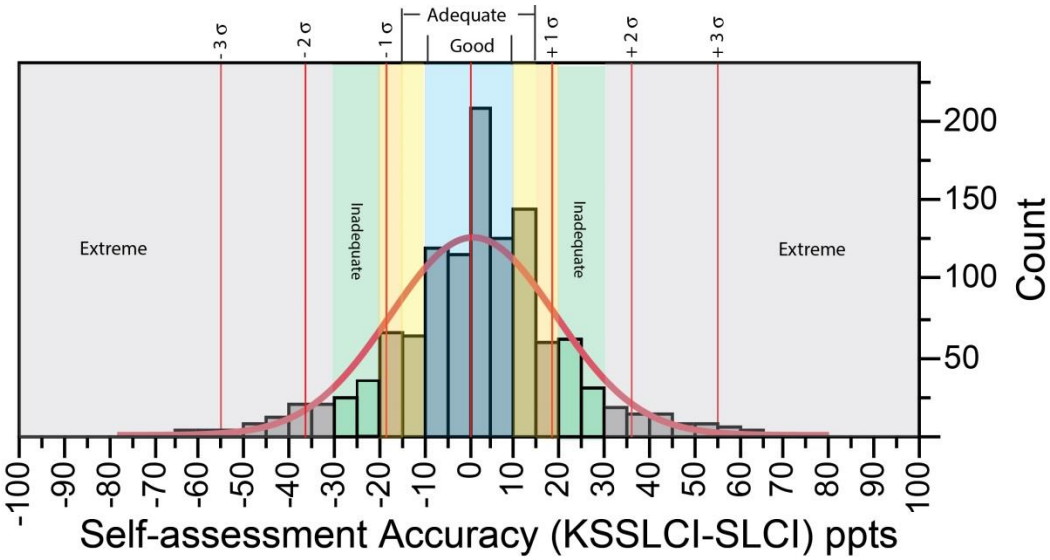

Figure 8. Distributions of the categories of self-assessment accuracy based on the differences in percentage points (ppts) between scores received from 1154 participants who took the 25-item Science Literacy Concept Inventory (SLCI) and their self-assessed ratings of competence in understanding science as a way of knowing as registered by the 25-item knowledge survey of the Inventory (KSSLCI). Standard deviation (sigma) = 18.4 ppts. Color codings of categories are the same as in Figure 7 with "Extreme" inaccuracies covering the entire gray area.

In Appendix A, we explain our process for setting the boundaries in Figures 7 and 8 by using the standard deviations of self-assessment inaccuracies (*KSSLCI rating - SLCI score*) deduced from the distributions produced by the population of experts. The use of standard deviations alone rather than inaccuracy in ppts provides a basis for an alternate classification scale. We chose to feature a scale based on percentage points here because doing so offers immediate use to readers who measure self-assessment accuracies of their students as percentages and have neither a large enough dataset from which to create their own scale nor a population of known experts with which to calibrate their measures.

To our knowledge, Figure 7 represents the first effort to construct a criterion-referenced self-assessment scale. We recognize that our self-assessment results and categories defined in this first effort could be contextual to the topic that we investigated, the instruments that we used, and the populace that we examined. Future studies may alter the boundary cut-offs, but conversations about where the boundaries might be better set cannot occur without establishing the language needed to enable such discourse. In addition, our study allows others to use our instruments as a convenient way to calibrate their populations' self-assessment characteristics and to compare self-assessed abilities in their study populace as measured by their instruments with ours.

Implications for Teaching, Learning and Assessment

Self-assessment appears to be a teachable metacognitive skill (Kruger and Dunning 1999) that is meaningful and measurable. It may be one of the most beneficial skills of all for students to develop (Rivers 2001; Pintrich 2002).

The obvious way to promote skill in metacognitive self-assessment is to design lessons that require students to practice it. Informal ways of doing so include adding requirements that students self-assess the scores that they believe they are going to obtain on each submitted assignment. Each quiz or test that starts with a predicted assessment of an estimated score on the coming evaluation and ends with a postdicted assessment of the score anticipated after completing each test or quiz offers an opportunity for practice.

This research employed a knowledge survey (KSSLCI). Instructors often credit knowledge surveys as sources of information for promoting effective learning and for improved course design (Nuhfer 1996; Nuhfer and Knipp 2003; Nicolaysen and Ritterbush 2005; Wirth and Perkins 2005; Wirth, Perkins and Nuhfer 2005; Clauss and Geedey 2010; Goodson, Slater and Zubovic 2015). Knowledge surveys promote good class planning (Nuhfer and Knipp 2003), particularly through aiding employment of tight instructional alignment (Cohen 1987).

In assessments, most scholars report that data obtained from knowledge surveys prove useful for "closing the loop" and informing future class modifications to support student learning (Nuhfer et al. 2010; Bell and Volckmann 2011; Favazzo, Willford and Watson 2014). Others used numerical arguments to reject knowledge surveys as a useful measure of assessment (Bowers, Brandon and Hill 2005; Ebert-May and Weber 2006) and offered views that differed little from those that consigned self-assessed learning measures to random noise (Porter 2012, 2013).

To employ numerical analyses to resolve the disagreement about whether knowledge surveys offer valid assessments for measures of student learning required a study that furnished a critical mass of data obtained from closely aligned instruments of documented reliability. The database employed in this paper, which is that used in Nuhfer et al. 2016a and Nuhfer 2015, meets that requirement.

Pre-course knowledge surveys provide a record of predicted self-assessments about content that participants do not yet fully understand. Post-course knowledge surveys provide a record of postdicted self-assessed competence about content on which participants are now better informed. The results shown in this paper indicate that collective self-assessments offer a valid measure that is significantly related to the true competencies of the populace as a whole. When people understand the challenge to which they self-assess their competence, these self-

assessments are usually valid estimates of performance that they can demonstrate. Designing course materials that improve learners' metacognitive abilities may be one of the most productive ways to use the content of any discipline to promote adult intellectual development.

Conclusions

We tested three competing hypotheses regarding self-assessment by analyzing a large dataset (N = 1154) that registered reliable paired self-assessed competence ratings and demonstrated competence proficiency scores. The first hypothesis, which proposes self-assessed estimates of proficiency to be random noise, proved untenable.

Our results contradicted the generally accepted second hypothesis, which proposes: (a) peoples' self-assessed competence ratings show a pronounced bias toward overestimations of their actual abilities and (b) low-proficiency performers are those most prone to egregious overestimations. The prevalent acceptance of this second hypothesis rests largely on the interpretation of patterns yielded by the Kruger-Dunning-type graphical format. Our analyses revealed that these patterns invite misinterpretations of data traceable to overlooked aspects of numeracy. By studying categorical data from known experts and novices, we confirmed that qualified experts are indeed more skillful in self-assessment than are novices. However, our study refuted two tenets of the second hypothesis by showing that (a) no strong propensity exists toward overconfidence in self-assessment ratings and (b) few people (about 5%) merit their being characterized as "unskilled and unaware of it."

Our study permitted creating a quantitative classification scale for self-assessment skills and making a detailed characterization of the skills of a population sampled from higher education. Our results supported the third hypotheses by confirming that (a) peoples' self-assessed competence generally accords with their demonstrated proficiency and (b) peoples' frequencies of self-assessed underestimation of their competence are similar to their frequencies of overestimation. Both qualities held true for novices and experts, and our data from undergraduate college students indicated that, on average, women seem significantly better at self-assessment than do men.

Metacognitive self-assessment is a quality that is measurable and meaningful. However, deprecating self-assessment by deeming it as noise or meaningless nonsense is partly responsible for why teaching self-assessment and tracking gains acquired by practice remains widely neglected in higher education.

In much of the peer-reviewed self-assessment literature, we believe we have found key weaknesses in the numeracy employed during nearly two decades of collecting, presenting, and interpreting self-assessment data. Because of

DOI: http://dx.doi.org/10.5038/1936-4660.10.1.4

insufficient attention to numeracy, current prevalent explanations of the nature of human self-assessment seem to rest on a tenuous foundation.

Acknowledgments

The authors thank the editors and anonymous reviewers of *Numeracy* for their in-depth suggestions, criticisms and encouragement. The peer-review process at *Numeracy* proved to be one of the best among our collective experiences in prompting us to probe deeper into both the study and its presentation. This enabled producing a needed contribution to a challenging area of study. We wish to thank Lauren Scharff, U.S. Air Force Academy, Paul Walter of St Edwards University and Mary Dalles of U-WI at Platteville for valuable feedback on draft manuscripts. This work was performed in accord with approval on human subjects research by IRB-105122 from 2010–2013 at CSU Channel Islands and IRB-13-019 from 2013–2017 at Humboldt State University to comply with all relevant federal guidelines and policies.

References (including references cited in Appendix A)

Ackerman P. L. and S. D. Wolman. 2007. "Determinants and Validity of Self-Estimates of Abilities and Self-Concept Measures." *Journal of Experimental Psychology: Applied*, 13 (2): 57–78. http://dx.doi.org/10.1037/a0026556.

Ackerman P. L., M. E. Beier and K. R. Bowen. 2002. "What We Really Know About Our Abilities and Our Knowledge." *Personality and Individual Differences* 33: 587–605. http://dx.doi.org/10.1016/S0191-8869(01)00174-X.

Alverno College Faculty. 2000. *Self Assessment at Alverno College*. G. Loacker, (Ed.) Milwaukee, WI: Alverno College.

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.

Beyer, S., L. Rynes and S. Haller. 2004. "Deterrents to Women Taking Computer Science Courses." *Technology and Society Magazine*, IEEE 23 (1): 21–28. http://dx.doi.org/10.1109/MTAS.2004.1273468.

Bolívar-Cruz, A., D. Verano-Tacoronte and S. M. González-Betancor. 2015. "Is University Students' Self-Assessment Accurate?" In *Sustainable Learning in Higher Education, Innovation, Technology, and Knowledge Management* ed. M. Peris-Ortiz, and J. M. Merigó Lindahl, Chapter 2, 21–34. Switzerland: Springer.

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.

Buratti, S. and C. Allwood. 2012. "The Accuracy of Meta-Metacognitive Judgments: Regulating the Realism of Confidence." *Cognitive Processing*, 13: 243–253. http://dx.doi.org/10.1007/s10339-012-0440-5.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). "Skilled or Unskilled, but Still Unaware of It: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons." *Journal of Personality and Social Psychology* 90: 6077. http://dx.doi.org/10.1037/0022-3514.90.1.60.

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.

Cech, E., B. Rubineau, S. Silbey and C. Seron. 2011. "Professional Role Confidence and Gendered Persistence in Engineering." *American Sociological Review* 76 (5): 641–66. http://dx.doi.org/10.1177/0003122411420815.

Clauss, J. and K. Geedey. 2010. "Knowledge Surveys: Students Ability to Self-Assess." *Journal of the Scholarship of Teaching and Learning* 10: 14–24.

Cohen, S. A. 1987. "Instructional Alignment: Searching for a Magic Bullet." *Educational Researcher* 16 (8): 16–20. http://dx.doi.org/10.3102/0013189X016008016.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt.

Dunlosky, D. and A. R. Lipko. 2007. "Metacomprehension: A Brief History and How to Improve Its Accuracy." *Current Directions in Psychological Science* 16 (4): 228–232. http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x.

Dunlosky, D. and K. A. Rawson. 2012. "Overconfidence Produces Underachievement: Inaccurate Self-evaluations Undermine Students' Learning and Retention." *Learning and Instruction* 22 (4): 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J. and J. Metcalfe. 2009. *Metacognition*. Los Angeles: Sage Publications.

Dunlosky, J., M. J. Serra, G. Matvey and K. A. Rawson. 2005. "Second-Order Judgments about Judgments of Learning." *Journal of General Psychology* 132: 335–346. http://dx.doi.org/10.3200/GENP.132.4.335-346.

Dwan, K., D. G. Altman, J. A. Arnaiz, J. Bloom, A. W. Chan, E. Cronin, E and P. R. Williamson. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS ONE* 3 (8): 8e3081. http://dx.doi.org/10.1371/journal.pone.0003081.

Ebert-May, D. and E. P. Weber. 2006. "RESPONSE: Re: The Use of a Knowledge Survey as an Indicator of Student Learning in an Introductory Biology Course." *CBE Life Science Education* 5 (4): 315–316. Retrieved July 9, 2015, from http://www.lifescied.org/content/5/4/315.2.full.

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.

Ertmer, P. A. and T. J. Newby. 1996. "The Expert Learner: Strategic, Self-regulated, and Reflective." *Instructional Science* 24: 1–24. http://dx.doi.org/10.1007/BF00156001.

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. Accessed October 18, 2015. http://dx.doi.org/10.1128/jmbe.v15i2.693.

Flavell, J. H. 1976. "Metacognitive Aspects of Problem Solving." In *The Nature of Intelligence*, ed. L. B. Resnick, 231–235. Hillsdale, NJ: Erlbaum.

Gaze, E., 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. Accessed October 18, 2015. http://dx.doi.org/10.5038/1936-4660.7.2.4.

Goodson, L., A., D. Slater and Y. Zubovic. 2015. Adding Confidence to Knowledge. *Journal of the Scholarship of Teaching and Learning* 15 (1): 20–37. http://dx.doi.org/10.14434/josotl.v15i1.12761.

Handel, M. and E. Fritzsche. 2016. "Unskilled but Subjectively Aware: Metacognitive Monitoring Ability and Respective Awareness in Low-Performing Students." *Memory and Metacognition* 44 (2): 229–241. http://dx.doi.org/10.3758/s13421-015-0552-0.

Hargittai E. and S. Shafer. 2006. "Differences in Actual and Perceived Online Skills: The Role of Gender." *Social Science Quarterly* 87 (2): 432–448. http://dx.doi.org/10.1111/j.1540-6237.2006.00389.x.

Hartwig, M. and J. Dunlosky. 2014. "The Contribution of Judgment Scale to the Unskilled-and-Unaware Phenomenon: How Evaluating Others Can Exaggerate Over- (and Under-) Confidence." *Memory & Cognition* 42: 164–173. http://dx.doi.org/10.3758/s13421-013-0351-4.

Isaacson, R. M. and F. Fujita. 2006. "Metacognitive Knowledge Monitoring and Self-regulated Learning: Academic Success and Reflections on Learning." *Journal of the Scholarship of Teaching and Learning* 6 (1): 39–55.

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.

McMillan, J. H. and J. Hearn. 2008. "Student Self-assessment: The Key to Stronger Student Motivation and Higher Achievement." *Educational Horizons*, 87 (1): 40–49.

Miller, T. M. and L. Geraci. 2011. "Unskilled but Aware: Reinterpreting Overconfidence in Low-Performing Students." *Journal of Experimental Psychology*: *Learning, Memory, and Cognition* 37: 502–506. http://dx.doi:10.1037/a0021802.

Nicolaysen, K. P. and Ritterbush, L. W. 2005. "Critical Thinking in Geology and Archaeology: Interpreting Scanning Electron Microscope Images of a Lithic Tool." *Journal of Geoscience Education* 53: 166–172. https:/doi.org/10.5408/1089-9995-53.2.166.

Nuhfer, E. B. 1996. "The Place of Formative Evaluations in Assessment and Ways to

Reap Their Benefits." *Journal of Geoscience Education* 44: 385–394. https:/doi.org/10.5408/1089-9995-44.4.385.

———. 2015. "Clarification to Points in "Correlating Student Knowledge and Confidence Using a Graded Knowledge Survey to Assess Student Learning in a General Microbiology Classroom"." *Journal of Microbiology & Biology Education* 16: (2). http://www.asmscience.org/content/journal/jmbe/10.1128/jmbe.v16i2.986#backarticlefulltext.

——— 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. Accessed December 20, 2015. http://dx.doi.org/10.1187/cbe.06-05-0166.

Nuhfer, E. B. and others. 2010. Knowledge Surveys. *MERLOT ELIXR Case Story*: Learning Object: Accessible Digital Case Lessons. CSU Center for Distributed Learning. Apr 1, 2010. Accessed Feb 14, 2016. http://www.merlot.org/merlot/viewMaterial.htm?id= 437918.

Nuhfer, E. B., C. Cogan, S. Fleisher, E. Gaze and K. Wirth. (Nuhfer et al. 2016a). "Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency." *Numeracy* 9 (1): Article 4. http://dx.doi.org/10.5038/1936-4660.9.1.4.

Nuhfer, E. B., C. Cogan, A. Goodman, C. Kloock, C. Wheeler, G. Wood and N. Zayas. 2016 (Nuhfer et al. 2016b). "Using a Concept Inventory to Assess the Reasoning Component of Citizen-Level Science Literacy: Results from a 17,000-Student Study." *Journal of Microbiology & Biology Education* 17(1):143–155. http://dx.doi.org/10.1128/jmbe.v17i1.1036.

Pazicni, S. and C. F. Bauer. 2013. "Characterizing Illusions of Competence in Introductory Chemistry Students." *Chemistry Education Research and Practice* 15: 24–34. Accessed October 18, 2015. http://dx.doi.org/10.1039/C3RP00106G.

Perry, W. G., Jr. 1999. "*Forms of Intellectual and Ethical Development in the College Years*." Reprint of the original 1968 1st edition with introduction by L. Knefelkamp. Jossey-Bass, San Francisco, CA

Phan, K. L., T. D. Wager, S. F. Taylor and I. Liberzon. 2004. "Functional Neuroimaging Studies of Human Emotions." *CNS Spectrums* 9 (4): 258–66. https:/doi.org/10.1017/S1092852900009196.

Pintrich, P. R. 2002. "The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing." *Theory Into Practice* 41: (4), 219–225. http://dx.doi.org/10.1207/s15430421tip4104_3.

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. Accessed June 4, 2016. http://www.hcmstrategists.com/contextforsuccess/papers/PORTER_PAPER.pdf.

———. 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.

Rivers, W. P. 2001. "Autonomy at All Costs: An Ethnography of Metacognitive Self-Assessment and Self-Management among Experienced Language Learners." *The Modern Language Journal* 85: 279–290. http://dx.doi.org/10.1111/0026-7902.00109.

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. Accessed March 20, 2016.

Vasilev. M. R., 2013. "Negative Results in European Psychology Journals." *Europe's Journal of Psychology* 9 (4). 717–730. http://dx.doi:10.5964/ejop.v9i4.590.

Wirth, K. R., and D. Perkins. 2005. "Knowledge Surveys: An Indispensable Course Design and Assessment Tool." Presented at the Innovations in the Scholarship of Teaching and Learning at Liberal Arts Colleges, St. Olaf, MN. Retrieved May 25, 2015, from http://www.macalester.edu/geology/wirth/CourseMaterials.html.

——— and E. Nuhfer. 2005. "Knowledge Surveys: A Tool for Assessing Learning, Courses, and Programs." *Geological Society of America Annual Meetings Program with Abstracts*. 37 (7): 119. Accessed February 14, 2016. https://gsa.confex.com/gsa/2005AM/finalprogram/abstract_97119.htm.

Wittmann, M. K. N. Kolling, N.S. Faber, J. Scholl, N. Nelissen, and M. F. S. Rushworth. 2016. "Self-Other Mergence in the Frontal Cortex During Cooperation and Competition." *Neuron*, 91 (2): 482–493. http://dx.doi.org/10.1016/j.neuron.2016.06.022.

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.