Appendix A: Supplementary Materials for 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), Karl Wirth (Macalester College), Steven Fleisher (California State University - Channel Islands), Christopher Cogan (Ventura College), Eric Gaze (Bowdoin College)

Contents

All References cited in this Appendix are listed in the References at the end of the main paper.

1. Supplement to Introduction: Influence of the Kruger-Dunning Type Graphical Convention

We provide this Appendix to illustrate the details of numeracy of self-assessment and to provide an understanding of why we did what we did in developing the main paper.

Successfully *reading and interpreting graphs* is one of the basic concepts of numeracy listed by Gaze et al. (2014). Reading and interpreting graphs of the Kruger-Dunning type (Fig. 1) also enlists the other concepts. Understanding how sorting of data determines the mean values of the quartiles requires *number sense*. So does understanding how percentiles represent data differently from reporting data as raw score percentages. Understanding how the value of the mean for each quartile affects the likelihood of members of that quartile underestimating or overestimating their abilities requires *awareness of probability*. Understanding whether quartiles that exhibit visual differences are significantly different from one another requires *awareness of statistics*. Being able to understand how the patterns that depict pure random noise aid in interpreting the nature of human self-assessment from patterns that present real data requires developed *reasoning*.

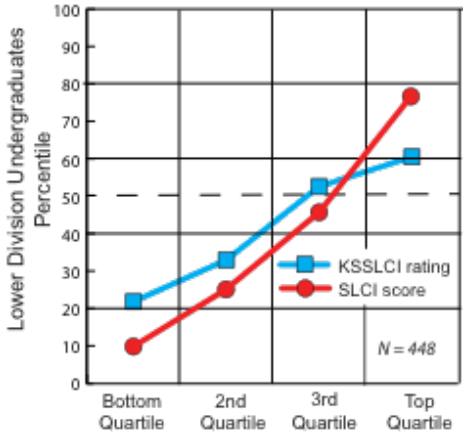

Figure A1-1. Classic Kruger-Dunning convention for depicting self-assessment accuracy. Line plots depict self-assessed competency (KSSLCI rating) and demonstrated competency (SLCI scores) from our 448 lower-division undergraduates. Interpreting this graph requires estimating self-assessed inaccuracy by the vertical distances between the two lines at each quartile mid-point.

To examine the Kruger-Dunning graphical convention, we employ a subset of our data that consists only of our lower-division undergraduate students (448 freshman and sophomore students). These lower-division undergraduates provide a population more similar to the homogeneous populations studied by most previous self-assessment researchers than does our entire database, which consists of a spectrum of experts through novices. Lower-division undergraduates constitute those anticipated to be novices in understanding science's way of knowing. Nuhfer et al. (2016b) documented the validity of this expectation on over 17,000 students, and we reconfirmed that for our studied populace here (Fig. A1-7).

We addressed the first three of the following six Kruger-Dunning graph complications in our earlier paper. We address these three briefly here, but we direct the reader to that earlier paper (Nuhfer et al. 2016a)[1] for illustrations and details.

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 (addressed in Nuhfer et al. 2016a, Figure 5).***

A Kruger-Dunning graphic can reveal the degree of random noise in paired self-assessment measures by how closely a best-fit line through the four quartiles of the self-assessed competency approximates horizontality (Nuhfer et al. 2016a, Figure 5, and this Appendix, Fig. A1-6). In Figure A1-1, that line is steeply inclined and shows a high signal-to-noise ratio. When best-fit lines through the four quartiles of the self-assessed competency are nearly horizontal, these indicate that the data acquired is mostly random noise and the data presented in the Kruger-Dunning type graph is unfit for interpretation.

2. ***The Kruger-Dunning type graphs present patterns that appear meaningful from datasets too small to offer reliability (addressed in Nuhfer et al. 2016a, Figure 6).***

Because the Dunning-Kruger graph splits data into four quartiles, each of the quartiles needs to have a population large enough to allow its mean to represent the character of self-assessment that each quartile expresses. We know that the character of self-assessment represented by random numbers is the mean calculated from random numbers between 0 and 100, which should theoretically be 50. The populations represented by numbers chosen at random from the range 0 to 100 must be large enough to produce a mean close to 50 in order to establish reproducibility. For good reproducibility, a population of about 400 participants with about 100 participants in each quartile is desirable (Nuhfer et al. 2016a, Figure 6).

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).***

The Kruger-Dunning type graphical convention (Fig. A1-1) plots the (y) and (x) aggregated by quartiles separately on the ordinate. The convention requires the reader to make the $(y - x)$ subtraction with a visual estimate to deduce the self-assessment accuracies from the vertical distances that separate the lines at each quartile. The graph simply portrays difference by distance rather than expressing that difference as a number. Thus, the Kruger-Dunning type graph is simply a variant of $(y - x)$ *vs.* (x) graphs. The differences seen between self-assessed competence and actual competence are largest in the bottom quartile because low scores of competence make large overestimates of competence more probable. In

[1] http://scholarcommons.usf.edu/numeracy/vol9/iss1/art4/

the top quartile, high scores designating high competence make large overestimates of competence impossible and underestimates more probable. We now proceed to examine additional complications not detailed in our earlier paper.

4. Sorting data pairs by one member of the pair invariably produces the "X-shaped" pattern and, sorting data by percentile rank renders all expressions of performance as norm-based rather than criterion-based.

The first part of Statement 4 observes that the sorting employed to construct the graphs appears to inject a self-fulfilling prophecy for paired self-assessment data to produce an X-shaped pattern showing the relatively unskilled as overestimating their abilities, and participants with the highest abilities as underestimating theirs. It seems impossible not to produce the X-shaped pattern with normal data taken from a populace that includes people that overestimate their abilities, some that underestimate and some that accurately estimate. The mean of competence in the top quartile, by definition containing the quarter of the database with the highest scores, is always higher than the mean of the accompanying array of unsorted self-assessment ratings included in that quartile. The reverse occurs in the bottom quartile; the mean of actual performance in the bottom quartile of sorted data always proves less than the mean of accompanying ratings of perceived performance.

Ackerman, Beier, and Bowman (2002) also observed that random numbers could produce graphs in the Kruger-Dunning convention that were similar to some results reported by Kruger and Dunning (1999). We found that graphing of pure random noise in the Kruger-Dunning convention, as simulated by the graphing of random number pairs, maximizes the differences between the quartiles (Nuhfer et al. 2016a, Fig. 5). Thus, random noise within self-assessment data seems to be a powerful contributor to the X-shaped pattern. The noisier the data, the greater the acute angle of intersection produced by plotting the paired measures in this graphical format.

The second part of Statement #4 notes that the act of sorting (ordering) data produces a norm-based expression of performance. Expressing data in percentiles is a method for ordering data. Percentiles quantify the ordered position of each score or rating where it rests along a scale bounded by 0 and 100 for all data pairs. When researchers present data as percentiles and sort the data by one of the measures, any skewed distributions of percentage scores become rearranged along the 0 to 100 scale toward a broader and normal distribution by converting the raw scores into percentiles. Expressing the data as percentiles obscures differences between two very different populations: one consisting of experts with mostly high scores of competence and another populace consisting of novices with mostly lower scores. It is easy to overlook that presenting data as norm-based or criterion-based can yield different interpretations and perceptions about the true character of the self-assessment abilities of a given populace.

The top quartile of lower-division undergraduates novices represents the "best of the unqualified" whereas the top quartile of a populace of professors represents the proverbial "cream of the crop" of qualified experts. While the true competencies of the people within these two different top quartiles are very different, Kruger-Dunning type graphs expressed as percentiles will not

communicate such differences. To see this criterion-based difference requires us to examine a populace that contains known experts and novices and how they differ in actual competence and self-assessed competence. As discussed in the main body of our paper under the subheading "*Using Categorical Data: Comparing Experts with Novices,*" numerical distinctions between novices and experts require us to view the distributions of the categories or quartiles, and Kruger-Dunning type graphs reveal the means, not the distributions. To see this requires a different kind of graphic, such as that we used under that subheading in the main paper. We employ that kind of graphic here in Figure A1-2 to overcome the limitation addressed in point #5 that follows.

5. Kruger - Dunning graphs fail to show the distributions of varied self-assessment skills in a populace.

We begin our explanation of this point by comparing Figures A1-1 and A1-2. The two figures present the same data, but Figure A1-2 discloses information that the Kruger-Dunning convention omits. Figure A1-2 shows the confidence intervals of the means of each quartile, the significance of differences in these means, and the ranges in distributions of values of participants within each quartile.

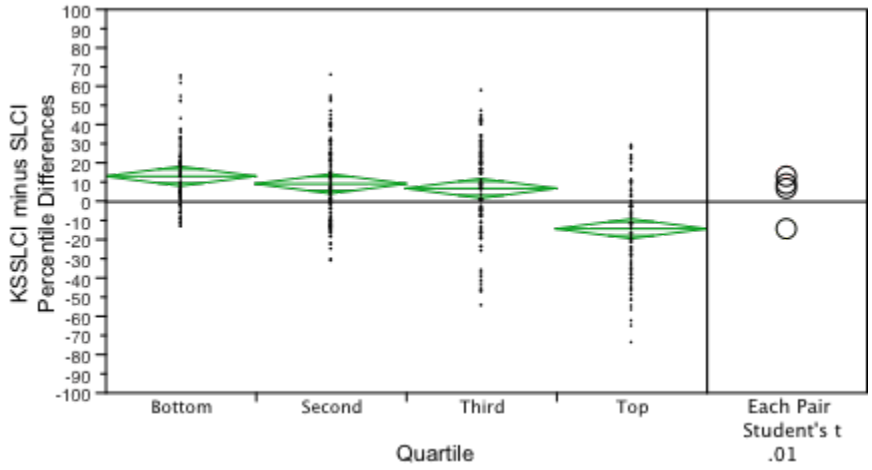

Figure A1-2. Self-assessment accuracy of 448 lower-division undergraduates and their distributions by quartiles. Accuracy is expressed in percentile differences. The height of the green diamonds reflects the bounds of the 99% confidence level. The black dots reveal the ranges of participants' scores in each quartile. Box to the right depicts the significant differences between quartiles as expressed by t-testing. Diameters of the circles are the bounds of the 99% confidence interval. Complete separation of circles shows that their means differ significantly. Overlapping of circles reflects a lack of significant differences between the quartiles. Graph produced by SAS Institute's JMP 11.2 software.

In both figures A1-1 and A1-2, the abscissas display the actual competence derived from the quartiles' mean SLCI scores as positioned by the mean middle values for each quartile (12.5, 37.5, 62.5, 87.5). The ordinates display the aggregated means of self-assessed competence minus demonstrated competence or ($KSSLCI - SLCI$) for each quartile.

The cross-plotting of ($KSSLCI - SLCI$) *vs.* ($SLCI$) in Figure A1-2 is easy to recognize as a ($y - x$) vs. (x) graphical convention. We documented all ($y - x$) vs. (x) conventions as troublesome sources of strong ceiling and floor effects that

graphs in the literature come from small studies that lack a critical mass of participants. The seminal study by Kruger and Dunning (1999) presented four graphs from four case studies. Two of their graphs reveal nearly horizontal self-assessment lines, and all four of their case studies come from populations too small to achieve good reproducibility.

2. Supplement to Results

Validation of the "Novice" and "Expert" Categories Employed in this Paper through Demonstrated Competence

The SLCI challenged our participants to demonstrate cognitive competence in the ability to recognize and understand science as an evidence-based way of knowing. Knowing factual content did not advantage participants in this challenge. Results from over 17,000 participants verified measurable distinctions between experts (professors) and novices (lower-division undergraduates). In that study, professors outperformed every undergraduate rank at the 99.9% confidence level (Nuhfer et al. 2016b). Our 1149 paired measures of participants of known rank also affirmed highly significant differences in demonstrated competence between experts and novices (Figure A1-7).

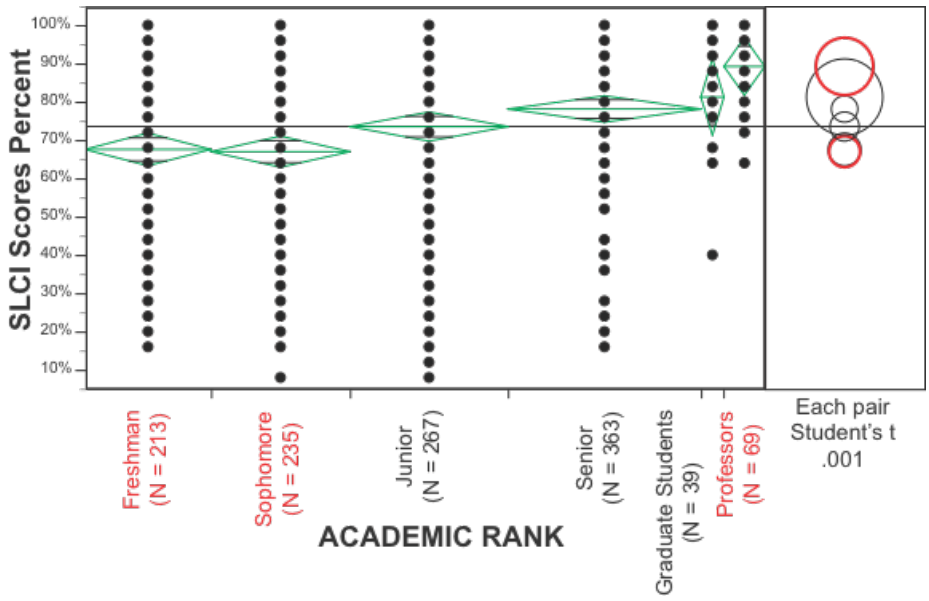

Figure A1-7. Participants' % (*N* = 1149) mean scores on the SLCI arranged along the abscissa by academic rank together with their distributions of responses (black dots). Green diamonds show the means and confidence boundaries at the 99.9% confidence level. Red font shows ranks we designated as experts (professors) and novices (freshmen + sophomores) that are significantly different at P < .0001. Diameters of the circles are the bounds of the 99.9% confidence interval. Horizontal line marks grand mean of 73.6%. Panel to the right depicts the significant differences between ranks as expressed by t-testing. Diameters of the circles are the bounds of the 99.9% confidence interval. Clear separation of circles depicts statistically significant differences. Graph created with SAS Institute's JMP 11.2 software.

Figure A1-7 confirms that lower-division undergraduates (freshmen and sophomores) are indeed novices by comparison to professors in demonstrating

proficiency in understanding science's way of knowing. The two groups' means of demonstrated competence differed at high levels of statistical significance. On the other hand, we saw in the main paper (Figs. 2 and 3) that experts' and novices' group means in self-assessment proficiencies differed little.

3. Supplement to Discussion

Validating the Classification Scale for Self-Assessment Accuracy

As noted in the main paper, which introduced the self-assessment classification scale in Figures 7 and 8, the criterion-based performance of the categories of experts can inform our setting of boundaries that define the self-assessment categories of "Good," "Adequate," "Marginal," "Inadequate" and "Extreme." In this paper, these terms are not arbitrary value judgments, but definitions. Here we detail our deducing of the boundaries that provide the definitions through using the distributions expressed by standard deviations of performance abilities of experts (professors) in Figure A1-8.

Definitions in the main paper's Figure 7 come from picking bounds at the intervals of 10 ppts *closest to the boundaries set by the distributions of experts' self-assessment accuracy*. Because self-assessment accuracies derive from two measures that are reliable but imperfect (Nuhfer et al. 2016a), selecting rigid percentage point values at the exact boundaries of the standard deviation cells established from a single study is not likely justifiable. We opted instead to set more general boundaries that facilitated intuitive understanding at 10 ppt intervals that more easily permit comparisons with self-assessment results on different topics obtained with different measuring instruments.

Figure A1-8 displays the distributions of experts' self-assessment accuracies and shows the bounds of standard deviations and the ppt boundaries deduced from them. Self-assessments defined as "Good" fall within the first two standard deviations of the experts' distribution. Adequate to marginal self-assessments include all of the third standard deviation and small amounts of the second and fourth standard deviation. Inadequate self-assessments lie within the fourth standard deviation, and extremely inaccurate self-assessments fall in the fifth standard deviation and beyond.

Thereafter, we applied these same bin sizes to our study populace of participants who self-identified their academic rank (N = 1149). Extreme errors in self-assessment accuracy seen in the populace of experts (Fig. A1-8) were about one-third of that seen in the general populace (Fig. A1-9).

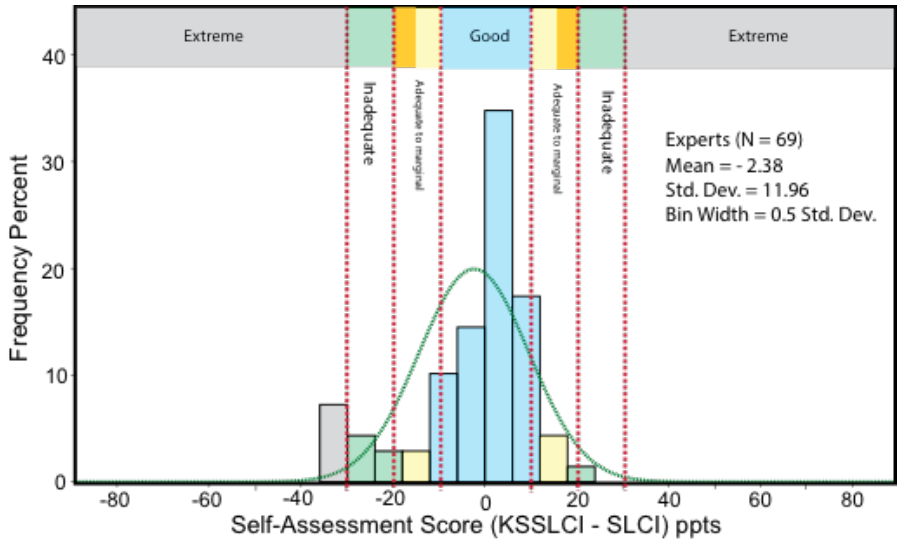

Figure A1-8. Basis for the classification scale in main paper's Figures 7 and 8 based on the distributions of self-assessment accuracy of experts (professors). Bin widths are 0.5 standard deviations of approximately 6 ppts. Definitions in main paper Figure 7 come from picking units of 10 ppts closest to the boundaries set by the distributions of experts' self-assessment accuracy. Red vertical lines show the relationships of these classification scheme standard deviations to the category breaks as defined by intervals of ten percentage points. Color-coding in band at top follows the terminology definitions of the percentage point classification scale detailed in our main paper Figures 7 and 8.

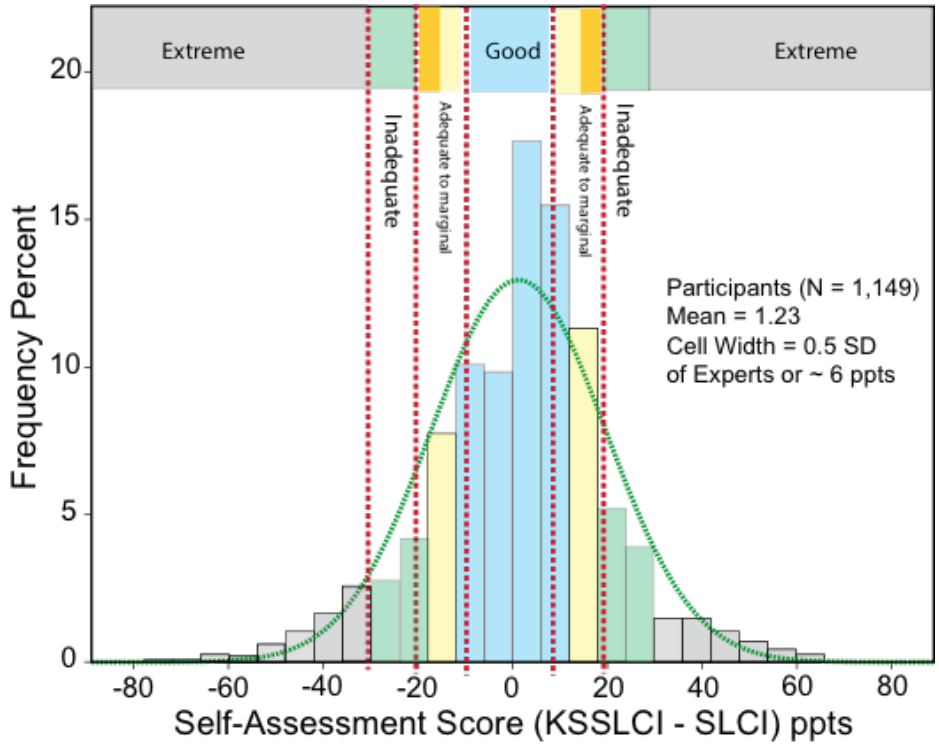

Figure A1-9. Classification scale based on standard deviations of experts' performance and applied to the whole study populace. The scale is guided by the bin widths of 0.5 standard deviations of experts' self-assessment accuracies. Perfect self-assessment is defined at $(KSSLCI - SLCI) = 0$ ppts. Bin widths are approximately 6 ppts. Red vertical dashed lines show the relationships of the standard deviations to the category definitions in the classification scale as defined by percentage points. Color-codings and labels in band at top follow the terminology definitions of the classification scale in our main paper Figures 7 and 8.

As seen in Figure A1-9, a classification scheme, whether based rigidly on standard deviation boundaries or more generally at convenient intervals of 10 ppts, yields similar results and enables meaningful dialog with language that communicates quantitative understanding.

These same cell bins serve to show profound differences in the distributions of self-assessment abilities between novices and experts (Fig. A1-10).

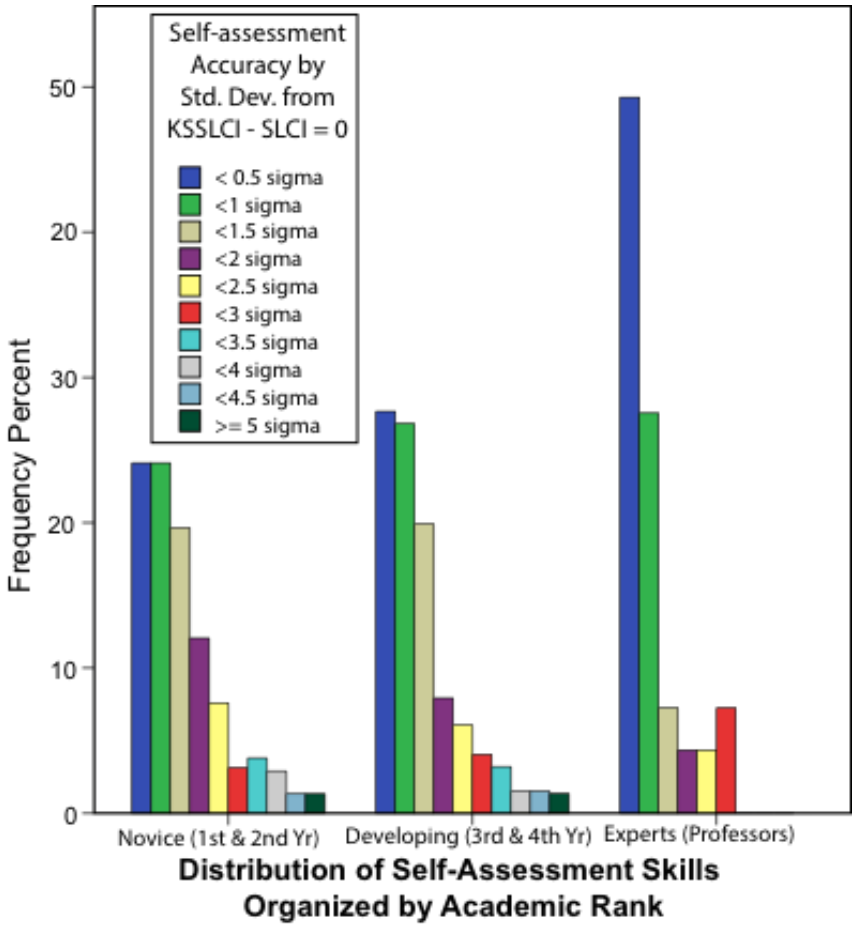

Figure A1-10. Distributions of categorical self-assessment proficiency across academic ranks by standard deviations from perfect self-assessment defined by $(KSSLCI - SLCI) = 0$ ppts. Novices are defined as lower-division undergraduates, developing experts as upper-division undergraduates, and experts as professors.

Figure A1-10 conveys information compatible with that conveyed by Table A1-2 where the categories expressed follow the definitions established in our classification scale. Table A1-2 reveals that different groups' proficiencies can be better characterized according to distributions of the members' self-assessment abilities in each group than by the sparse information provided in the quartiles of Kruger-Dunning type graphs.

Table A1-2.
Distribution of self-assessment accuracy of novices and experts by categories of self-assessment skill.*

Category of Self-Assessment Skill	Novices: Lower-Division Undergraduates *N* = 448	Experts: Professors *N* = 69	Overall Study Population *N* = 1149
Good	43.40%	73.90%	48.30%
Adequate	17.60%	7.30%	17.80%
Marginal	11.60%	5.80%	10.50%
Inadequate	15.10%	10.10%	12.50%
Extreme	12.30%	2.90%	10.90%

*Chi-square testing revealed that experts differ significantly (P < .0001) from novices and from the overall study population who self-identified their class rank. Novices are not significantly different from the overall study population.

The distributions reported in Table A1-2 indicate that people's self-assessed competence generally accords with their actual competence. About three-quarters of professors (experts) accurately self-assessed the abilities (within ±10 ppts) that they later demonstrated. The percentage of lower division undergraduates (novices) who could do this was 43.4%.

We can also apply the plots that employ the cell bins of standard deviations to look at the apparent differences in relative expertise across sorted data aggregated by quartiles. For this, we examine only the undergraduates in Figure A1-11, which shows clear differences between the quartiles.

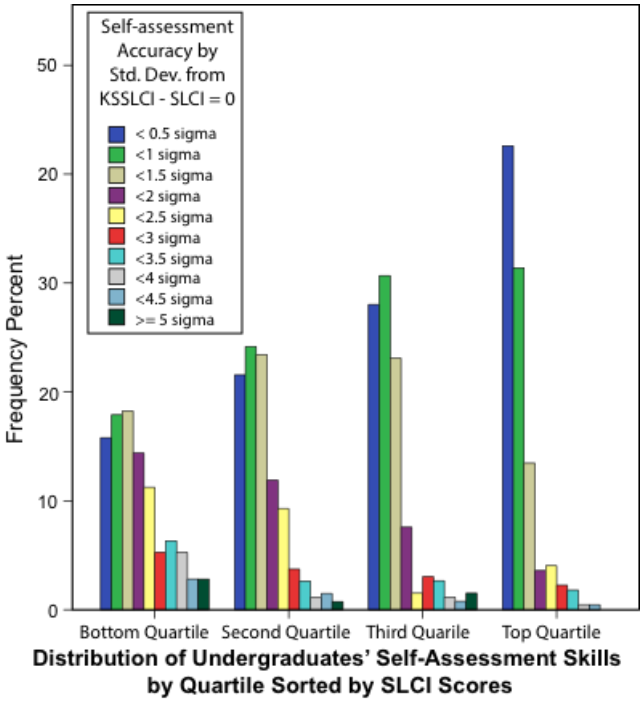

Figure A1-11. Distributions of relative self-assessment proficiency across sorted data aggregated into quartiles. Data is depicted as standard deviation from perfect self-assessment defined by $(KSSLCI - SLCI) =$ 0 ppts.

However, the trends of the categories of the undergraduates' academic ranks do not reflect the trends of the aggregates of the undergraduates' quartiles. Table A1-3 shows the distributions of these ranks according to the five defined categories of self-assessment skill. Chi-squared testing failed to confirm a statistically significant difference between the academic ranks in *self-assessment skills*, whereas Figure A1-7 confirms highly significant differences in *content proficiency* (SLCI scores) between lower-division and upper-division undergraduate ranks.

Table A1-3.
Distribution of self-assessment skills of undergraduate students from our database by academic rank.

Category of Self-Assessment Skill*	Freshmen** *N* = 213	Sophomores *N* = 235	Juniors *N* = 267	Seniors *N* = 326
Good	45.1%	42.1%	46.7%	52.5%
Adequate	16.4%	18.7%	21.7%	17.5%
Marginal	12.2%	10.6%	10.1%	9.8%
Inadequate	14.6%	15.7%	11.6%	9.8%
Extreme	11.7%	12.8%	10.5%	10.4%

*Definitions of the categories follow criteria from Figure 2.
**Chi-squared testing confirms no significant difference between any two undergraduate ranks.

We suspect that metacognitive self-assessment skill is a significant characteristic of intellectual development (Perry 1999), which occurs slowly. Most colleges do not produce significant advances toward higher Perry stages in their undergraduates, but educating through planned curricula focused on advancing thinking can produce measurable progress in intellectual development in undergraduates (Alverno College Faculty 2000).

The purposeful teaching of metacognitive self-assessment skill might similarly accelerate its development and produce measurable gains. Instructors' employment of knowledge surveys can give students frequent practice in self-assessment and provide meaningful assessment data about student learning.