USING CONFIDENCE AS FEEDBACK IN MULTI-SIZED LEARNING ENVIRONMENTS

Thomas L. Hench
Delaware County Community College, Media, Pennsylvania, USA
thench@dccc.edu

ABSTRACT
This paper describes the use of existing confidence and performance data to provide feedback by first demonstrating the data's fit to a simple linear model. The paper continues by showing how the model's use as a benchmark provides feedback to allow current or future students to infer either the difficulty or the degree of under or over confidence associated with a specific question. Next, the paper introduces Confidence/Performance Indicators as graphical representations of this feedback and concludes with an evaluation of s trial use in an online setting. Findings support the efficacy of using the Indicators to provide feedback to encourage students in multi-sized learning environments to reflect upon and rethink their choices, with future work focusing on the effectiveness of Indicator use on performance.

INTRODUCTION
Confidence provides a means to assess the metacognitive knowledge students have about their performance – in essence, do students know what they know and what they don't know. Darwin Hunt (2003), one of the early researchers in the role of confidence, stated that the importance of having this knowledge is critical, for being misinformed is "much worse than being uninformed". Traditionally, one-dimensional assessment (performance only) supplies very little information about what students know and what they don't know. However, the addition of confidence as a second dimension provides important additional information in assessing students' knowledge of their performance (Adams and Ewen, 2009) while also promoting a potentially deeper level of reflection and self-regulation. Work by Bruno (1993), another early investigator, to measure knowledge quality led to the development of a two-dimensional assessment process which attempts to measure both correctness and confidence by a single quantity. Employed with success in training situations, this methodology, however, involves extensive calculations to implement which limits its potential use in middle to large scale learning situations containing hundreds or thousands of students.
Another approach is the confidence (or certainty) based marking (CBM) scheme developed by Gardner-Medwin and Gahan (2003) which assumes a linear relationship between the confidence (here referred to as certainty) and the mark expected by the students as shown by the left-hand figure in Figure 1.

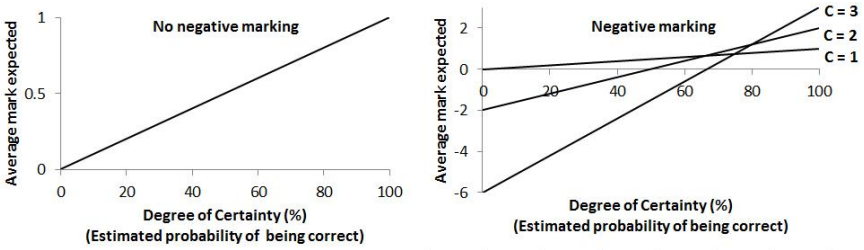

Figure. 1. Confidence-based marking schemes

Building upon this assumed relationship, Gardner-Medwin and Gahan proposed the use of a negative marking scheme (right-hand figure), where students are penalized for under or over estimating confidence and rewarded for reflection and deeper thought before answering. In this scheme, students receive points of 3, 2, or 1 for correct responses and 0, -2, or -6 for incorrect responses, depending on their estimated probability (confidence) of being correct. In essence, the scheme uses confidence as a motivating factor. While results (Schoendorfer and Emmett, 2012) obtained from the use of CBM, primarily in medical school education, yielded positive results in terms of improved performance, the method does not focus specifically on obtaining quantifiable confidence levels. Other research into incorporating confidence into grading utilized methods such as a Problem Solving Inventory (Larson, et. al, 1998) and the calculation of a "confidence score" (Petr, 2000) as ways to achieve what Paul (2007) calls "scoring systems which encourage honesty" and thus reliable measures of confidence. Additionally, recent research describes the use of the difference between confidence and accuracy as part of a "bias score" component of a mark (Michailova and Katter, 2013) and as a measure of a "metacognitive gap" (XXXX, 2012). An important part of these approaches is their use of a quantifiable measure of confidence as a second dimension of assessment in multi-sized (i.e. small, medium, or large) learning environments. However,

the use of confidence as this additional dimension requires knowing the relationship, if any, between confidence and performance. If confidence has no correlation with performance, then its use in assessment becomes unclear. Thus, the research question addressed in this paper is as follows – "What relationship, if any, exists between confidence and performance?" The answer to this question determines whether or not the use of confidence as a second dimension of assessment along with performance is possible.

METHOD

The experimental data gathered to investigate the research question comes from student responses in the author's online Astronomy course over a period of six semesters (September 2010 to December 2013) using the commercially available SurveyMonkey© software linked to the course syllabus. As part of the coursework, each new group of students answered the same baseline set of fifty six multiple choice questions each semester and then indicated either a low, medium, or high confidence level in their answers. Class size varied from 30 to 50 students per semester with the total number of responses per baseline question ranging from N = 170 to 288. Figure 2 shows the overall confidence level distribution and performance for a typical question presented each semester over the period of the study.

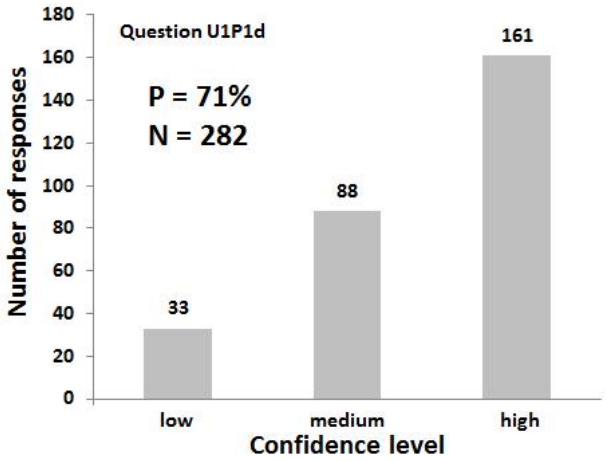

Figure. 2. Survey software output

In addition, Bloom's revised taxonomy (Krathwohl, 2002) permitted critical thinking levels to be assigned to each question. As shown in Figure 3, questions designated as Level I require factual and conceptual knowledge resulting from remembering and understanding to complete, whereas Level II questions need procedural knowledge obtained through the processes of applying and analyzing.

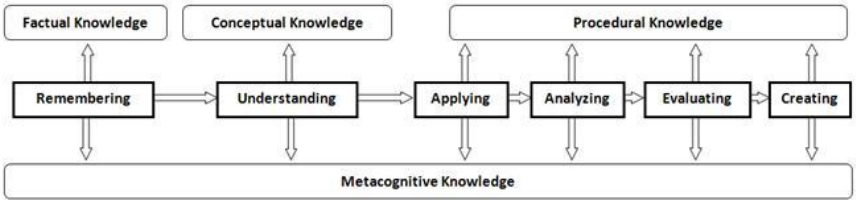

Figure. 3. Bloom's revised taxonomy

The determination of a quantifiable confidence level from student responses employed a physical analogy. The left-hand side of Figure 4 illustrates a confidence level distribution similar to that shown in Figure 2 and displayed as a bar chart with the magnitude of the total low, medium, and high confidence level responses are indicated by l, m, and h.

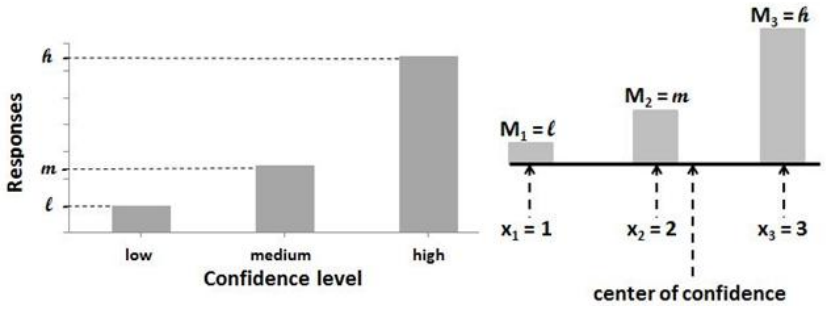

Figure. 4. Bar chart of typical confidence and center of confidence

As noted in the figure, the distribution of the confidence level responses provides an approximate description of the overall level for this particular question, in this case between medium and high. The right-hand side of Figure 4 illustrates another way of viewing this same information. Here the confidence magnitudes (l, m, and h) correspond to masses $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$ distributed at distances along a horizontal axis of $\mathbf{x_1}$ = 1 (low confidence), $\mathbf{x_2}$ = 2 (medium confidence), and $\mathbf{x_3}$ = 3 (high confidence), with the center of mass of this system given by the familiar expression

$$\text{center of mass} = M_i x_i / M_i. \qquad (1)$$

Substituting confidence magnitudes for masses and confidence levels for distances yields an analogous quantity called the center of confidence denoted algebraically as **C**, or

$$\mathbf{C} = (l + 2m + 3h)/(l + m + h). \qquad (2)$$

Applying equation (2) to the data shown in Figure 2 yields a center of confidence value of 2.45, in agreement with a visual estimate of the center of mass of an analogous physical system. Closer inspection of equation (2) reveals this result also corresponds to the expression used to determine the weighted average of the confidence magnitudes shown in the bar graph. Therefore, in addition to providing a visual representation, the center of confidence also provides the confidence level expected for a particular question. Stated in another way, each question has associated with it a center of confidence specific to that question. This result suggests an interpretation of the meaning of confidence based not upon the response given by an individual student after answering a specific question but to the expected response to that specific question before it is answered. It is this latter interpretation which is used as the meaning of confidence in this paper.

Before investigating the relationship between the confidence associated with a question (as represented by the center of confidence) and the performance on that question, the meaning of the latter needs further clarification. For each question, **P** represents the percentage of students who answered a particular question correctly as indicated in Figure 2. Conversely, this percentage also represents the expected or probable performance associated with that specific question. Thus, similarly to the treatment of confidence, each question has associated with it an expected or probable performance. Consequently, the meaning of performance here becomes the expected or probable outcome for a specific question rather than the outcome resulting from the answer given by an individual student to a specific question. This paper employs the probabilistic interpretation for the meaning of performance with the quantity **P** now denoted as the performance probability and expressed as a percentage. In view of the previous discussion, the research question is restated as "What relationship, if any, exists between the center of confidence **C** and performance probability **P**?"

RESULTS

Gardner-Medwin and Gahan's assumed linear "no negative marking" case shown in Figure 1 suggests a possible model for the relationship between **C** and **P**. Specifically, as the confidence level of increases the probability of answering correctly increases in direct proportion. The model as adapted here assumes that if all students answer a question correctly (**P** = 100%), they all would response at the highest confidence level thus yielding a center of confidence of **C** = 3. Similarly, if all students answer incorrectly (**P** = 0%), they do so at the lowest center of confidence level giving a center of confidence of **C** = 1. For the case of **P** = 50%, half of the students answer correctly and select the highest confidence level and the other half answers incorrectly and choses the lowest confidence level, thereby yielding a center of confidence of **C** = 2. A plot of these points results in the modeled performance probability $\mathbf{P_m}$ as a function of the center of confidence **C** (the dashed line and equation in Figure 5).

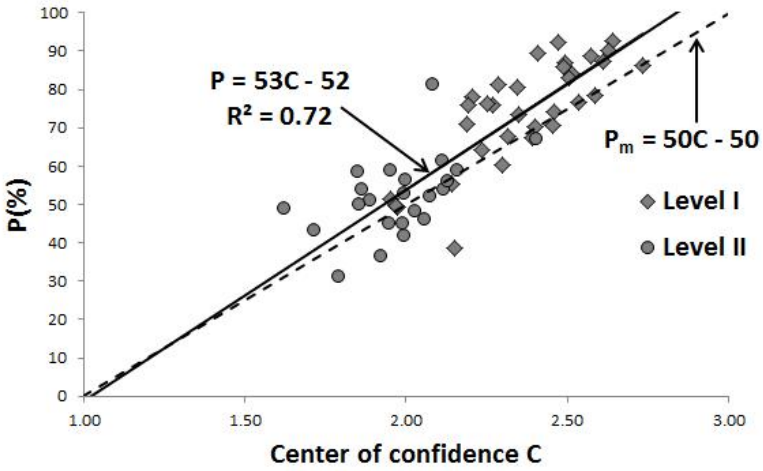

Figure. 5. Performance probability versus center of confidence for all responses

Also included in the figure are the experimental values of **P** and **C** as determined from the data for each baseline question, the accompanying linear regression line (solid line) and best fit equation to these points, the R-squared value, and the question critical thinking level for each question. Included in this plot is the point (71, 2.45) corresponding to the question referenced in Figure 2.

On first inspection, the data appears to be a reasonable fit to the model. To test the validity of the linearity of the model, the four assumptions shown in Table 1 regarding the use of a linear regression to describe the relationship need further examination. The violation of any of these assumptions as indicated by the validity tests calls into question the use of a linear model.

Table 1. Assumptions and validity tests for linear regressions.

Assumption	Validity Test
Linearity – the independent and dependent variables are linearly related to one another	No discernible pattern in the distribution of points about a horizontal line in a standardized residual versus predicted value plot
Homoscedasticity - the variance of values of the dependent variable from the regression line is constant	Approximately constant spread of points about a horizontal line in a standardized residual versus predicted value plot
Independence – the random errors associated with the dependent variable are unrelated to one another	Durbin-Watson statistic of ~ 2.0 with an acceptable range of 1.75 to 2.25
Normality – the residual errors associated with the dependent variable are randomly distributed	Presence of a diagonal line resulting from normal probability plot

Figure 6 shows, on the left, the plot of the standard residual versus predicted performance probability P obtained by an Excel analysis of the data. The apparent linear relationship from Figure 5 and the discernment of no pattern associated with the points in Figure 6 both support the validity of the linearity assumption. In addition, the spread of points above and below the zero line is approximately equal therefore supporting the homoscedasticity of the data. (Possible outliers seen in Figure 6 will be addressed later in the paper). A value for the Durbin-Watson statistic of 1.93, calculated using Excel, supports the independence assumption and Figure 6 shows the diagonal line obtained from the normal probability plot, again obtained via Excel, again lending support of the normality assumption.

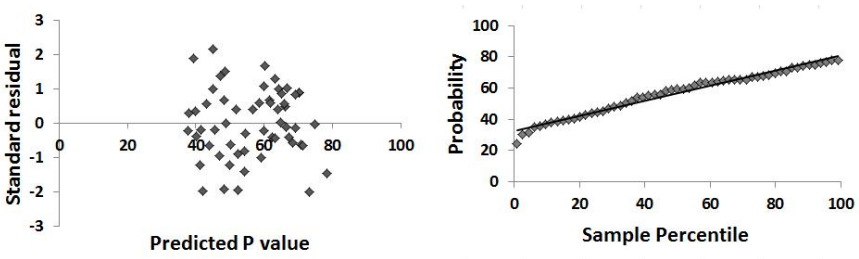

Figure. 6. Standard residual/predicted P plot and normal probability plot

Furthermore, in a normal distribution, 63% of the total residual points fall within plus and minus one standard deviation and 95% between plus and minus two standard deviations. These conditions are also met by the data in Figure 6.

In summary, the validation of the four assumptions stated in Table 1 supports the use of linear relationship to model the behavior between the experimentally determined centers of confidence and the performance probabilities and, as such, provides an answer to the research question posed in the paper.

DISCUSSION

The closeness of agreement between the experimental line and the model line shown in Figure 5 suggests the use of the latter as a benchmark for comparing and interpreting the experimentally determined values of **C** and **P**. To examine this possibility, Figure 7 shows the previously plotted data with only the model line shown.

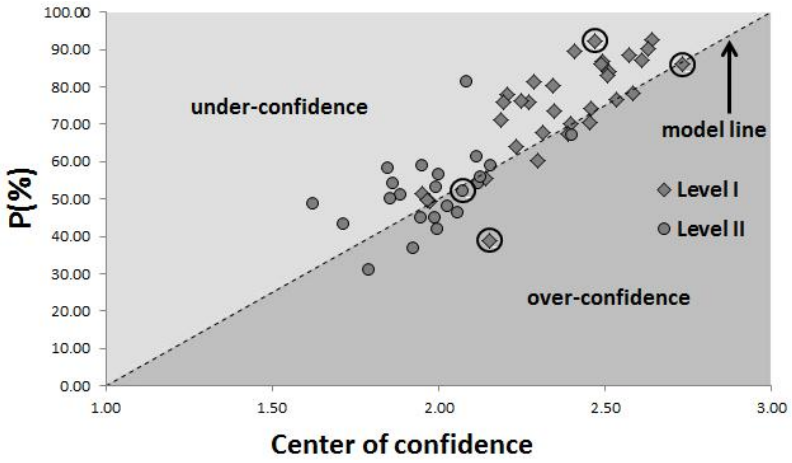

Figure. 7. Confidence regions

Specifically, points lying in the region either above or below the model line indicate situations of under or over confidence For example, the encircled point in the lower center of Figure 7 corresponds to a center of confidence **C** of 2.2 and an actual performance probability **P** of 39%. At this confidence level, the model predicts an expected performance probability of 60%. Thus, this question has associated with it an overestimation of the confidence in performance. For the point in the upper right of the figure given by **C** = 2.5 and **P** = 92%, the model predicts a performance probability of 75%. For this question, an under confidence in performance for that question is expected. For the two other encircled points lying on or close to the model line, **C** = 2.1, **P** = 52% and **C** = 2.7 **P** = 86%, the expected performance probabilities are 55% and 85% respectively. In these two cases, the performance predicted by the centers of confidence is in close agreement with the experimental performance probability. In this case, each question is considered calibrated, the difference between these two calibrated questions possibly attributable to the degree of difficulty of one question compared to the other. Thus, the use of the model line as a benchmark for comparing actual centers of confidence and performance probabilities allows for the identification of relative problem difficulty and the degree of under or over confidence associated with a question. Furthermore, the concept of miscalibration (Klayman, et.al., 1999) offers an explanation for the variation in confidence seen in the Figure 7 by describing how judgment errors result in over confidence on difficult problems and under confidence on less difficult ones. This interpretation is also consistent with the

"hard/easy effect" (Murad, 2014) found in non-incentivized self-reporting of confidence as found in this study. The predominance of Level I questions in the under confidence region and the over confidence associated with some Level II questions supports this explanation.

While this interpretation does not presume the absence of errors in the data which may account for some of the differences shown, it nevertheless offers an alternative explanation for deviations from the model. Indeed, points lying at large distances from the model line possibly result from content or structure differences in questions, with outliers (under and over confidence points) indicating issues as to how the questions were phrased and resulting in a possible misinterpretation of the question and subsequent misplaced confidence. In any case, the deviation from the benchmark model line reveals differences in questions, whether intended or not.

The Confidence/Performance Indicators shown in Figure 8 graphically represent the information previously discussed for the four examples taken from Figure 7. Importantly, the indicators allow for both confidence and performance to be combined in a straightforward manner. In each indicator, the benchmark performance probability predicted by the model for a given center of confidence (top circle) is shown by the position of the arrow on the performance scale. The lower circle indicates the actual performance probability as found from the data.

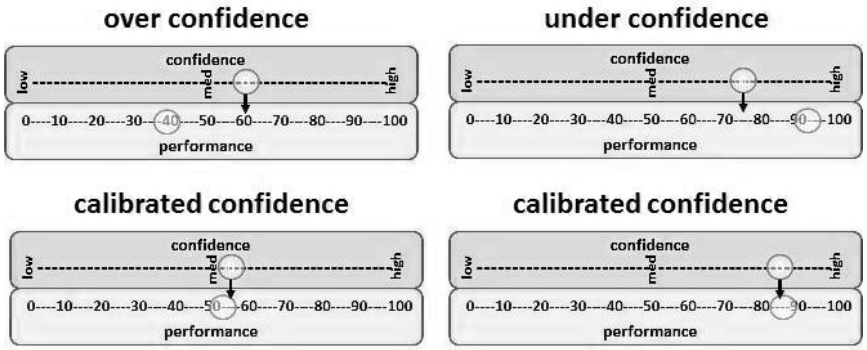

Figure. 8. Confidence/Performance Indicators

When included as part of a question, the indicators provide information which allows students to gauge the relative difficulty of a question as well as checking for any degree of under or over confidence associated with the question. In this sense, the Confidence/Performance Indicators provide a mechanism to deliver feedback by addressing what Glasson (2008) notes as "what has been done well in relation to the success criteria", "what still needs to be done in order to achieve the success criteria", and "advice on how to achieve that improvement". Specifically, the use Confidence/Performance Indicators suggests a means to encourage reflection and rethinking on the part of the student without using negative grading.

The author conducted a trial of the indicators to determine the efficacy to encourage rethinking and reflection on the part of students. Specifically, the indicators, embedded into nineteen of the fifty six baseline questions over the course of eight weeks provided students in an online Introduction to Astronomy class with the option of referring to them as part of the determining an answer. Prior to this, all students completed a tutorial on the concept of the indicators and their use in identifying the relative difficulty of questions and cases of under or over confidence. After answering the questions, students then completed a survey to determine the number who had or had not chosen to use the indicators, their reasons for using or not using them, and their level of helpfulness for those who had used the indicators. The two areas previously mentioned, question difficulty and under/over confidence, and two additional questions regarding rethinking and reflection and the overall helpfulness comprised the four survey questions given to those students who chose to use the indicators. Table 2 shows these questions, along with the rating scale employed. In addition, two open ended questions asked the students to comment about why they did or did not use the indicators. As only those who used the indicators responded to the survey, a forced-choice format provided the possible responses to survey questions. Research (Rasinski, et.al., 1994; Smith, et.al, 2006) which suggests that people who answer forced-choice questions spend more time and invoke deeper processing when answering supports this choice.

Table 2. Survey questions for students using the Confidence/Performance Indicators

Scale → very unhelpful unhelpful helpful very helpful Survey questions ↓	
Question Difficulty	How would you rate the Confidence/Performance Indicators in helping you judge the difficulty of the questions?
Under/Over Confidence	How would you rate the Confidence/Performance Indicators in alerting you to under or over confidence issues with the questions?
Reflect/Rethink	How would you rate the Confidence/Performance Indicators in making you rethink or reflect on your answers?
Overall	Overall, how would you rate the Confidence/Performance Indicators in helping you to answer the follow-up questions?

Of the 47 students answering the nineteen baseline questions containing the indicators, 87% (41) indicated that they referred to the Confidence/Performance Indicator when answering and thus completed the survey questions. Figure 9 shows the distribution of responses to the four survey questions and Table 3 provides an analysis of the three most common areas mentioned in the open-ended questions answered by all students. (Note: Cases of greater than 100% result from rounding errors.)

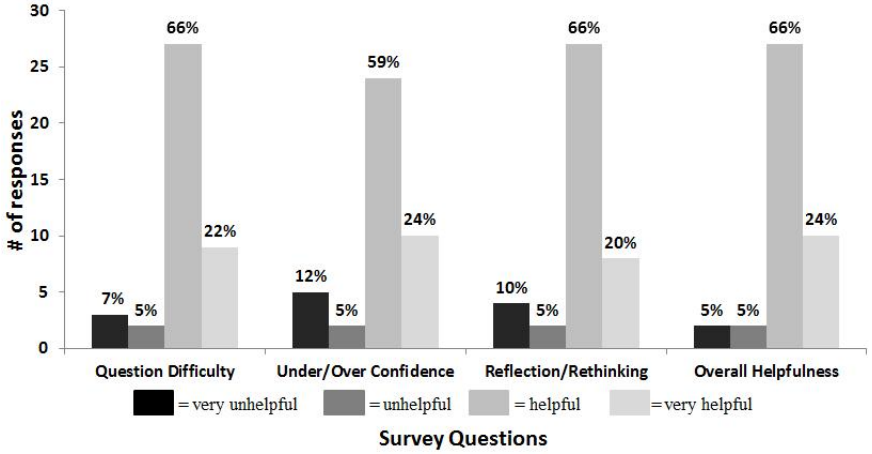

Figure. 9. Survey results of feedback areas

Table 3. Open-ended survey questions and response areas

Students using the indicators: "In the space below, enter any comments (pro or con) about the use of the Confidence/Performance Indicators in answering the follow-up question." Three most common response areas: 1) Rethink, review, recheck, or reflect: 10 occurrences 2) Comparisons: 6 occurrences 3) Usability issues: 5 occurrences
Students not using the indicators: "Briefly list below the reason(s) why you did not use the Confidence/Performance Indicators when answering the follow-up questions." Three most common response areas: 1) Already possess sufficient confidence in answer = 4 occurrences 2) Negative effect on answers (lower confidence) = 2 occurrences 3) No real reason = 2 occurrences

As illustrated in Figure 9, the majority of students (equal to or greater than 83%) selected with either helpful or very helpful responses when responding to the questions shown in Table 2. Additionally, analysis of the open-ended questions indicates that students felt the indicators encouraged rethinking and comparison, a result consistent with the survey results. A sample of comments regarding the use of the indicators include "definitely

helps me rethink and recheck my answer", "make you think before answering questions", "offer a view of how other students are looking at problems and the level of difficulty", "helped me gauge how accurate my questions were and gave me more confidence for each answer I submitted", and "they let me know that the reason I was taking so long to answer was that it was a more difficult question." Important comments regarding the usability of the indicators such as "a better understanding on how to use the indicator, when answering questions will be helpful" and "Only con is it takes some getting used to but once you understand it its useful" suggest that those students finding the indicators unhelpful or very unhelpful need better preparation. Indeed, one student's comment that "I can't see how past students answers can help me, because they could be wrong or right" suggests a lack of understanding of what information the indicators provide.

Comments from those students not using the indicators such as "I wanted to see how much I really new about the questions without using the performance indicators" and "I am not really sure why I do not use them, I just do not" again suggest incomplete knowledge of indicators' function and their use at providing feedback.

In view of the survey results and open-ended responses, the results of the trial use of the Indicators support their efficacy as a feedback mechanism to encourage rethinking and reflection. To address the usability concerns identified in Table 3, the Confidence/Performance Indicator tutorial requires a revision to include more examples and situations of their application with students. Furthermore, the Indicators will be employed in all baseline questions in an online section of ninety to one hundred students. Having demonstrated here their ability to foster rethinking and reflection, the author plans to pursue further research to determine the effectiveness of the use of Confidence/Performance Indicators on student performance.

In summary, using existing data of student responses to a set of fifty-six baseline questions gathered over a period of six consecutive semesters, analysis showed that the calculated centers of confidence and corresponding performance probabilities followed a linear model. This model, in turn, provided a benchmark for interpreting the experimental data which resulted in feedback regarding question difficulty and the degree of under or over confidence associated with a question. The introduction, demonstration, and subsequent positive evaluation of Confidence/Performance Indicators as a graphical means of displaying feedback suggests their continued use as an efficacious method of providing this feedback to encourage rethinking and reflection on the part of students. More specifically, once created and implemented the indicators require no interaction with an instructor and function in small, medium, or large learning situations. Furthermore, generating the data necessary to establish the indicators requires only the addition of low, medium, or high confidence response options as part of formative or summative assessments with data collection and analysis performed electronically. Thus, as a feedback mechanism, Confidence/Performance Indicators provide a quantifiable second dimension to assessment which is adaptable to multi-sized learning environments.

REFERENCES

Adams, T., Ewen, G. (2009). The Importance of Confidence in Improving Educational Outcomes, *Proceedings 25th Annual Conference on Distance Teaching & Learning*, Board of Regents of the University of Wisconsin System.

Bruno, J. (1993). Using Testing to Provide Feedback to Support Instruction: A Reexamination of the Role of Assessment in Educational Organizations**,** Item Banking: *Interactive Testing and Self-Assessment***,** Editors: Leclercq, D. and Bruno, J. NATO ASI Series Computer and Systems Sciences Volume F112 Springer-Verlag Berlin Heidelberg GmbH**,** 190-209.

Gardner-Medwin, A., Gahan, M. (2003). Formative and Summative Confidence-Based Assessment, *Proceedings 7th International Computer-Aided Assessment Conference*, Loughborough, UK, July 2003, 147-155.

Glasson, Y. (2008). Improving Student Achievement: A Practical Guide to Assessment for Learning, *Education Services Australia*, 78.

XXXX. (2012). Assessing Metacognition Via An Online Survey Tool, *Proceedings of the 10th International Conference on Computer-Based Learning in Science*, Barcelona, Spain, June 26 – June 29.

Hunt, D. (2003). The concept of knowledge and how to measure it, *Journal of Intellectual Capital*, 4(1), 100-113.

Klayman, J., Soll, J., Gonzalez-Vallejo,C., Barlas, S. (1999). Overconfidence: It Depends on How, What, and Whom You Ask, *Organizational Behavior and Human Decision Processes*, 79(3), 216–247.

Krathwohl, D.(2002). A Revision of Bloom's Taxonomy: An Overview, *Theory Into Practice*, 41(4), copyright 2002 College of Education, The Ohio State University, Autumn 2002.

Larson, D., Scott, D., Neville, M. Knodel, B. (1998). Measuring Student's Confidence with Problem Solving in the Engineering Design Classroom, *Proceedings of the Annual Conference American Society for Engineering Education*, June 28-July 1 Seattle, Washington.

Michailova, J., Katter, J. (2013). Thoughts on quantifying overconfidence in economic experiments, *MPRA Paper No. 53112*, Helmut-Schmidt University & York University.

Murad, Z., Sefton, M, and Starmer, C. (2014) How do risk attitudes affect measured confidence? *CeDEx Discussion Paper Series*, Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham.

Paul, J. (2007). Improving educational assessment by incorporating confidence measurement, analysis of self – awareness, and performance evaluation: *The computer-based alternative assessment (CBAA) Project*. Retrieved from http://www.jodypaul.com/ASSESS/

Petr, D. (2000). Measuring (and Enhancing?) Student Confidence with Confidence Scores, *Proceedings of the 30th ASEE/IEEE Frontiers in Education Conference*, Kansas City, MO October 18-21.

Rasinski, K., Mingay, D., Bradburn, N. (1994). Do respondents really mark All That Apply on self-administered questions, *Public Opinion Quarterly*, 58(3), 400-408.

Schoendorfer, N., Emmett, D. (2012) Use of certainty-based marking in a second-year medical student cohort: a pilot study, *Advances in Medical Education and Practice*, Dove Medical Press Ltd., 3, 139-143.

Smyth, J., Dillman, D., Christian, L., Stern, M. (2006) Comparing Check-All and Forced-Choice Question Formats in Web Surveys, *Public Opinion Quarterly*, 70(1), 66-77.