Chemistry Education Research and Practice

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the **Information for Authors**.

Chemistry Education Research and Practice

ARTICLE TYPE

Determining what our students need most: Exploring student perceptions and comparing difficulty ratings of students and faculty

Ozcan Gulacar*[a] and Charles R. Bowman[b]

If the goal of teaching is to help students understand a subject, teaching cannot begin until student difficulties with a subject are understood. In order to create a guide for assessing student difficulties with chemistry material, students were asked to rate exam questions on three factors: problem difficulty, familiarity, and self-confidence. These surveys were then compared to difficulty ratings of the same
10 questions as determined by chemistry professors. Students' ratings of problem difficulty, problem familiarity, and self-confidence correlated, as expected, with their success on exam problems. There was also some agreement between students and faculty on the difficulty of exam problems, though students were much more accurate of judging problem difficulty than were chemistry professors. Students were surveyed in two separate cohorts to test year-to-year reliability.

15 Introduction

Introductory engineering and science courses are notorious for their high attrition rates, which are usually attributed to deficiencies in students' knowledge (Moss & McMillen, 1980; Hundhausen *et al.*, 2011; Parker Siburt *et al.*, 2011). Chemistry,
20 specifically, is a challenging science because it requires good understanding of several concepts and the ability to transition between micro, macro, and symbolic representations (Hoffmann & Laszlo, 1991; Johnstone, 1993; Gabel, 1998; Pinarbasi & Canpolat, 2003; Johnstone, 2006). Studies, however, indicate that
25 students develop unconnected and compartmentalized knowledge, which makes learning chemistry hard for many students and, therefore, has received a lot of attention from educators (Lewis & Linn, 2003; Williamson *et al.*, 2004). The question arises, then, of how to best address these student
30 difficulties. Carter and Brickhouse suggest a way forward:

We may begin to understand student difficulties in chemistry if we understand the ways in which their perceptions of the context of our chemistry courses differ from our perceptions. Otherwise, students and faculty are living in different worlds and speaking
35 different languages (Carter & Brickhouse, 1989).

Carter and Brickhouse (1989) observed in their study that students and faculty did, indeed, live in two different worlds. Whereas students thought that doing homework and attending lecture were more important, faculty thought that student interest
40 in the subject was more important. Interestingly, students seemed to view mastery of chemistry as within their reach, while the faculty thought chemistry was just too difficult of a subject. Grove and Bretz (2007) created the CHEMX survey to measure expectations for learning chemistry (i.e., knowing how to learn
45 chemistry). Both students' and chemistry teachers' expectations about learning chemistry were measured; expectations between

general chemistry students and faculty were significantly different, but that difference largely disappeared by the students' third year.

50 Most of the studies looking at how accurately students assess their success on exams are performed in psychology or related courses (Sinkavich, 1995; Smith, 2002; de Carvalho Filho, 2009; Rosenthal *et al.*, 2010). While it is a popular topic for psychologists studying how students learn, rarely are the
55 students' perceptions compared with those of the faculty. The two studies above (Carter & Brickhouse, 1989; Grove & Bretz, 2007) are two of the few done on student perceptions in chemistry. However, neither was done on the question level of an exam, rather both studied overall perceptions of students; they were
60 unique in that they compared those perceptions with faculty perceptions.

Further studies by Symington and Kirkwood (1996) in general chemistry and Sözbilir (2004) in physical chemistry are the only other such studies known to the authors, though Symington and
65 Kirkwood did not compare the faculty perceptions to that of the students. Recently, a rubric has been developed to help chemistry faculty assess the complexity of their exams, which may help instructors in better predicting student success (Knaus *et al.*, 2011). These studies all point to discrepancies between teacher
70 and student perception of various aspects of chemistry. However, these studies did not analyse perception differences on specific chemistry topics, or the differences in perceived difficulty of those topics.

In general, students have been observed to have varying
75 abilities at predicting their own success. A number of variables may affect a students' ability to accurately gauge their own success. For example, when asked to predict their performance on a test, higher-performing students were more accurate at predicting their success; lower-performing students tended to be

overconfident in their success (Hacker *et al.*, 2000). Self-confidence in correct answers, too, has been shown to be correlated to success on exams (Smith, 2002).

Metacognition, too, is assumed to be related to student success in predicting success. Any evaluation one makes on whether or not they were successful is a form of metacognition (i.e., thinking about your thinking). It should not be surprising, then, that students with higher metacognitive skills, as measured by the Metacognitive Awareness Inventory (MAI)(Schraw & Dennison, 1994), were able to predict their success on an exam than were classmates with lower metacognitive awareness (de Carvalho Filho, 2009). Metacognition may even help students achieve more by compensating for lower aptitude (Cooper & Sandi-Urena, 2009). It has been shown that students in psychology courses can accurately predict their success on individual problems, but have a more difficult time predicting overall test scores on exams (Rosenthal *et al.*, 2010).

Given the lack of research on how students and teachers in chemistry perceive the difficulty of exams, the following research questions were investigated:

1. How accurately do students and teachers assess the difficulty of chemistry problems?
2. Who is more accurate in their assessment of chemistry problem difficulty – students or teachers?
3. Does metacognition affect a student's ability to assess the difficulty of chemistry problems?

Method

Study Cohort

This study was conducted at Texas State University – San Marcos in 2011. The students were volunteers from the summer general chemistry 1 course. One hundred sixteen students in total participated, though not all responded to each exam. The study was repeated in 2013 with a second cohort of students. This cohort was from a fall term, rather than a summer term, and had ninety-five students. Although fall enrolment is higher than summer enrolment, not all of the professors teaching general chemistry consented to including the study instrument with their exam, which resulted in the roughly equal number of fall and summer participants.

Three professors from the same university, with between 5 and 15 years of experience each, also participated in the study. None of these professors taught the general chemistry course at the time of the study, though all have at various times taught general chemistry, and none are authors of this paper. The study was approved by Texas State IRB.

To mitigate item-order bias, the tests given to the 2013 cohort were reordered. The test questions were identical to the tests given the previous cohort, but given in a different order within a section. The sections of each exam (fill-in-the-blank, multiple choice, and long answer) were not reordered. The survey instruments used were unchanged. The instructor for the 2013 cohort was not the same as that of the 2011 cohort.

Instrument

The data from this study are primarily taken from a survey that was filled out by the students and professors. The students were asked to fill out one survey after each of their four in-term exams;

the professors filled out corresponding surveys at a later date at their convenience. The students filled out the survey during the two-hour exam period. The survey could be completed while taking the exam or after completing the exam; the survey was handed in when the student turned the exam in for credit. Each student could choose when to fill out the survey during the two-hour exam period. Filling out the survey was not mandatory. Students who filled out the survey completely were given bonus points.

The surveys asked the student to rate each question of the exam they had just taken on the difficulty of that question, how confident they were on their answer, and how familiar they were with the problem. Each question was rated on a scale of 1 to 3, with 3 being most difficult. The scales and the descriptions given to the students can be seen in Table 1. For exams 1 and 2, a rating of 3 was also most familiar and most confident. However, because students expressed confusion about the self-confidence and familiarity scales, a rating of 3 was changed to least familiar and least confident; guessing was considered "very little confidence." Analysis of ratings before and after the switch showed no observable differences; the data from before and after the switch were both included in the study as a result. The ratings shown in Table 1 were used in all surveys for the 2013 cohort.

Table 1 Survey rating scales (after revision)

Difficulty (D)	Familiarity (F)	Self Confidence (SC)
1- Easy (solution requires remembering some basic definitions or facts)	1- Very familiar (have seen and done several similar examples)	1- Extremely confident (know how to check and it is correct)
2- Medium (solution requires formulas and some application)	2- Somehow familiar (have seen before but have very little experience)	2- Somehow confident (think I got it right but not sure and do not know how to check)
3- Difficult (solution requires linking among different concepts and a lot of calculations)	3- Not familiar (have not seen at all before)	3-Very little confidence (have no idea if I got it right)

The instrument asked students to rate each question on the exam with the scales shown in Table 1. 3-point scales were used to simplify the responses. Fewer response possibilities was thought to reduce the time needed to complete the survey. Students were also asked for any additional comments they had on each question and to identify the chemistry topic(s) they felt were being tested by that problem. A sample survey is available as appendix one. There were a total of 25 questions on exam 1, 23 questions on exam 2, 40 questions on exam 3, and 38 questions on exam 4. These exams were a mixture of fill-in-the-blank/short response, multiple-choice, and long answer/free response questions (Table 2). Students entered their responses on paper, which were hand-entered by student volunteers into a computer database for analysis. Professors were given copies of the exams and asked to rate the difficulty only for each problem and identify the topic(s) involved in each problem; they entered their responses directly into a computer.

Table 2 Count of problem types on each exam

	Short Response	Multiple Choice	Long Answer
Exam 1	5	14	6
Exam 2	4	15	4
Exam 3	0	40	0
Exam 4	0	33	5

Analysis

In addition to the exam data collected, each survey was identified with the student who responded so that their classroom performance data could be connected to their survey responses. Student success on each MC problem was collected (the data were binary in nature, i.e., right or wrong) and associated with each students' exam response. These data were used to create a "performance index" for each multiple choice problem on the four exams. The performance index was calculated by totalling the number of students who got the problem wrong and dividing that number by the total number of students who answered that question. The most difficult questions, then, have a performance index of 1 (i.e., all students answered this question wrong), while the least difficult have an index of 0. In all, three performance indices were created: one based on the 2011 cohort only, one for the 2013 cohort, and one that combined both cohorts' performance data. These indices were created to assess how accurate student and teacher assessments were; ratings of higher difficulty should correlate to higher performance indices.

Students in the 2013 cohort were also asked to fill out the MAI for bonus points on their own time not during the exams. Students filled out the MAI online using a 10-point scale, rather than the original 100-point scale. 55 students filled out the MAI. This number of surveys was not adequate for running a factor analysis. However, the original loadings were used to calculate the two MAI factors for each student (Schraw & Dennison, 1994); Cronbach's alpha for each of the factors was above .900.

Statistical analysis was done using SPSS 20. Most of the tests run were correlations calculated as Spearman's rho, r_s. Each test was calculated via bias-corrected accelerated (BCa) bootstrapping using 1000 samples (Preacher & Hayes, 2008; Singh & Xie, 2010). BCa bootstrapping is itself a non-parametric method (Efron, 1981, 1987); the 95% confidence intervals generated are in brackets. In the cases where the correlations were with the students' success on a problem, point bi-serial correlations were used. Positive correlations for the point bi-serial correlations in this paper indicate that the higher the rating (difficulty, self-confidence, or familiarity), the more likely that a student was correct on that problem; negative correlations indicate the reverse. For r_s, the effect size is equal to the coefficient (Field, 2009). To bring clarity to the analysis, the self-confidence and familiarity numbers were reversed prior to statistical analysis. The result was that a positive correlation between self-confidence and success on a multiple-choice problem meant that higher self-confidence correlated with higher success.

Results and Discussion

Student Perception

In order to answer the first research question comparing students'

perceptions about problem difficulty with reality, point bi-serial correlations were run between students' perceptions of each test question and their actual performance on said test questions (i.e., if the student answered the question correctly or incorrectly). Only the multiple-choice problems were included in this part of the study as the grading information for the other portions of the exams was not retained. It should also be noted that the number of multiple-choice problems varied by exam. This may have biased the analysis somewhat, as it has been noted that students are less accurate in accessing their success on multiple-choice problems (de Carvalho Filho, 2009). However, because the question of interest was not about specific chemistry topics, but overall student perception, this should only lower the effect size of any correlations rather than change the effect itself.

As can be seen in Table 3, there was a significant, negative correlation between students' difficulty rating on the problems and their success (i.e., correct or incorrect) on the problems ($r_s = -.200$, in 2011 and $r_s = -.237$ in 2013; $p < .001$ for both). This result indicates that students did have a somewhat accurate assessment of which problems were difficult (i.e., the higher the student difficulty rating, the less likely a student was to answer the question correctly). However, the accuracy was less than might be expected. A correlation coefficient of $-.200$ only accounts for about 4% of the variance in the data (i.e., $.040 = (-.200)^2$). The result was repeated in both cohorts with nearly the same effect size, even after applying bootstrapping methods, which suggests that this was not a one-time event.

In order to see if one exam was skewing the results, the correlations in Table 3 were run for each individual exam's questions. The results showed that the correlation remained largely consistent over each exam, with coefficients ranging from $-.110$ for exam 1 to $-.195$ for exam 4. The significance levels were likewise unchanged from the overall results, and year-to-year results remained unchanged. In general, the result was expected; students were able to judge the difficulty of chemistry problems with consistent accuracy. The surprise was that the accuracy of that judgment was much smaller than might be expected (i.e., only about 4% of the variance). In other words, one cannot simply assume that students will find problems that they are not likely to get right on an exam to be difficult.

Table 3 Correlations (Spearman's Rho) between student ratings (difficulty, familiarity, and self-confidence) of a chemistry exam problem and student success on those exam problems.

		Correlation Coefficient	Sig. (2 tailed)	N
Difficulty	2011	-.200 [-.219, -.181]	<.001	10368
	2013	-.237 [-.255, -.219]	<.001	9955
Familiarity	2011	.155 [.137, .175]	<.001	10368
	2013	.222 [.204, .240]	<.001	9955
Self-Confidence	2011	.186 [.166, .208]	<.001	10368
	2013	.249 [.230, .267]	<.001	9955

Of interest, too, was how students' feelings of familiarity with

a question were related to their success on a problem. The same was true of their confidence that their answer was correct. Here, too, were observed significant correlations between student perception and their performance (Table 3). These correlations were positively correlated, meaning that as students felt more familiar with a question, they were more likely to answer correctly (r_s = .155, in 2011 and r_s = .222 in 2013; p < .001 for both). The same was true of students' confidence in their answers; higher confidence was correlated with higher success (r_s = .186, in 2011 and r_s = .249 in 2013; p < .001 for both). Like the correlation with problem difficulty, these correlations were expected, but rather small in size, only accounting for 2-6% of the variance.

Again, correlations were also run for each individual exam. The results were much the same as for the difficulty ratings. None of the exams appeared to be exceptional with regards to correlations between student perceptions and student performance. It was good to note that the correlations here are opposite of those observed for the difficulty ratings. This means that students were more confident of and more familiar with problems they answered correctly.

With such consistent findings between the three ratings, it is reasonable to ask how closely related the ratings are. To determine this, correlations were run between each of the three student rating variables. As suspected from the above analysis, familiarity and self-confidence were found to be strongly, positively correlated (r_s = .707 in 2011 and r_s = .719 in 2013; p < .001 for both), which accounts for almost 50% of the variance between the two ratings. Furthermore, both familiarity and self-confidence were negatively correlated with difficulty (r_s = -.478 in 2011 and r_s = -.708 in 2013; p < .001 for a correlation between difficulty and familiarity; similar effect sizes were seen with self-confidence), which account for a large portion of the variance for each correlation.

As a result of the consistent pattern shown above and these strong correlations, a multiple linear regression was run to determine how interdependent the three variables were. (The data were unable to converge while running a logistic regression.) As expected, difficulty accounted for 5.6% of the model's variance (adjusted r^2 value). The addition of self-confidence accounted for an additional 1.8% of the model's variance, also a significant addition. However, the addition of familiarity does not account for any additional variance and further analyses suggest that they are co-linear (i.e., they describe the same thing).

Table 4 Linear model of predictors of success on a problem

	b	SE B	β	Sig.
Constant	0.26 [0.14, 0.37]	0.063		<.001
Difficulty Rating	-0.07 [-0.10, -0.05]	0.012	-.106	<.001
Self-Confidence Rating	0.11 [0.08, 0.13]	0.012	.166	<.001
Familiarity Rating	0.02 [-0.01, 0.04]	0.014	.027	.181
Knowledge of Cognition	0.04 [0.02, 0.06]	0.011	.081	.0012
Regulation of Cognition	-0.03 [-0.05, -0.01]	0.009	-.088	<.001

Metacognitive Awareness

To determine if metacognitive awareness had any influence on these observed correlations, the two MAI factors (knowledge of cognition and regulation of cognition) were added to the above linear regression (Table 4). While they did show a significant addition (p < .01) to the model, they only accounted for an additional 0.2% of the model's variance. They were not, however, co-linear with the other variables (difficulty, self-confidence, or familiarity). While this does agree with de Carvalho Filho's (2009) observation that metacognitive awareness increases a student's ability to predict their success on an exam question (research question 3), the effect size seems to be negligible.

Faculty Perception

The most important reason to gauge faculty perception of the difficulty of questions is to see how they match up with student perceptions. If teachers and students do not agree on which problems are most difficult, there is likely to be a disconnect between what the teacher emphasizes and what the students struggle to learn. In order to best measure the differences between student and faculty perceptions of difficulty, average (mean) student and average (mean) faculty ratings were created for each of the 126 questions. Unlike the correlations above, all questions, not just multiple-choice, were included, as no performance data were required to compare difficulty ratings.

In general, students and teachers agreed on the difficulty of problems. There was a large, positive correlation between students' and teachers' difficulty ratings (r_s = .468 [.303, .612] with 2011 cohort and r_s = .351 [.173, .519] with 2013, p < .001 for both). This accounted for roughly 21% of the variance between the faculty and students in 2011; 12% in 2013. In all, there was a low to moderate degree of agreement. However, this agreement varied by year. A much stronger agreement on difficulty was measured between the two cohorts of students (r_s = .765 [.668, .828], p < .001). Even with this higher degree of agreement, there does appear to be some variability in which problems students find difficult from year to year.

As such, it is important to know not just if the students and teachers agree, but who is generally more accurate in their assessment of difficulty (research question 2). Both student and faculty difficulty ratings were compared with each question's performance index (Table 5). Students in both cohorts had moderately strong, significant correlations between their average difficulty ratings and the performance index. As can be seen, the correlation was generally stronger when the cohort year matched with the performance index of that year. That is, students were better able to assess performance of themselves than another student cohort. The faculty, on the other hand, did not have significant correlations between their average difficulty rating of a problem and any of the performance indices. Overall, students, when averaged as a class, were much more accurate at determining the difficulty of the chemistry problem than were teachers.

There is a body of research that has aimed to increase the accuracy of both student and teacher perceptions of exam question difficulty for both general chemistry and organic chemistry (Knaus *et al.*, 2009; Knaus *et al.*, 2011; Raker *et al.*,

2013). These methods focus on cognitive complexity as a major factor in determining how difficult a given problem is; the more demand put on a student's cognitive processes (e.g., memory capacity), the more difficult a problem is. Thinking about the required cognitive load needed does help faculty more accurately gauge problem difficulty. The problem, as shown above, is that most chemistry faculty do *not* think in these terms. The faculty that participated in this study were not chemical education faculty; their selections of difficulty were based on whatever criteria they developed for themselves; criteria that were clearly not accurate. This difference between chemistry education faculty and other chemistry faculty has been observed before (Grove & Bretz, 2007).

Similarly, the students were not given a more rigorous framework. The comparisons in this study are between students' and faculty's perceptions of difficulty by whatever criteria they determine for themselves (within the rough framework shown in Table 1). Without any guidance, students appear to be much better at making judgements about chemistry problem difficulty.

Table 5 Correlations between question performance indices (2011, 2013, and combined) and average surveyed difficulty ratings for faculty, 2011 student cohort, and 2013 student cohort

	Performance Index 2011	Performance Index 2013	Performance Index Combined	N
Faculty	.178 [-.011, .363]	.191 [.002, .386]	.184 [-.004, .360]	102
Students 2011	.561[a] [.378, .702]	.441[a] [.259, .619]	.516[a] [.325, .669]	102
Students 2013	.341[a] [.129, .520]	.479[a] [.310, .631]	.414[a] [.224, .582]	102

[a] Correlation significant at 99.9% level ($p < .001$)

Conclusions

Summary

It has been shown that, at least for the students studied, chemistry students do have some idea of whether or not they are getting questions on exams correct; they do accurately perceive the difficulty of chemistry problems (research question 1). However, that degree of accuracy could be much higher. On an individual basis, the correlation coefficient between students' ratings of problem difficulty and their success on that problem was only around .200. However, when students' difficulty ratings were averaged and compared with the calculated performance index, the correlation raised to the range of .341-.561, accounting for approximately 12-30% of the total variance. What this means is that, individually, students have a low degree of success in predicting their success on a given problem. Metacognition was not shown to be a significant affective component for students predicting their own success (research question 3). However, students were collectively much better at predicting success/assessing problem difficulty. Assessment of the topics in which students are struggling, then, would be more accurate if determined collectively for the whole class.

It seems, too, that students and teachers agree somewhat on what is difficult (or easy) on chemistry exams, with an agreement that is in the range of 12-21% of the total variance. However, this does not mean that both teachers and students are equally good at assessing what is difficult. As was shown, students generally have a better idea than do teachers of what problems are most difficult (research question 2). This is likely because teachers, as experts in chemistry, have a very different view of the classroom than do students, who are novices in chemistry (Carter *et al.*, 1988). It seems that students and teachers still live in two different worlds (Carter & Brickhouse, 1989).

The primary purpose in surveying a second cohort was to verify that the correlations observed with the first cohort were not specific to that group. The repetition of the correlations seen in Table 3 supports the idea that these correlations are likely to be seen in various classrooms and were not isolated correlations. The differences in which problems students in each year found most difficult, however, suggest that the types of problems found most difficult will vary from year to year and teacher to teacher.

Implications for Chemistry Classrooms

As has been shown above, chemistry faculty are not very accurate at predicting the difficulty of chemistry problems. To make a real impact, it is important that chemistry faculty are made aware of this problem, and showing evidence-based studies, *e.g.* Grove and Bretz (2007) or this paper, may be the best way to convince them to re-evaluate how they determine what their students find difficult.

Since classrooms change from year to year, one survey of difficulties cannot determine the best course of action for every classroom. Students from the 2011 and 2013 cohorts did not rate the same problems as most difficult. There are a number of potential learning environments that will allow on-the-spot evaluation, such as clickers (King, 2011), flipped classrooms (Tucker, 2012; Johnson, 2013), or peer-led team learning (Woods, 2013), just to name a few. Because students generally have better idea of what is difficult for them, getting them involved is key, and these learning environments are designed to involve students. Surveying students with the survey used in this study for a sample of chemistry problems would be another way for teachers to find out what problems their students find difficult.

The strength of the correlations between the three student ratings (i.e., problem difficulty, familiarity, and confidence in a correct answer) of the exam questions suggests that improving one may help the other two. While reducing the difficulty of the questions – which would undoubtedly make students more confident – is not desirable, increasing self-confidence and familiarity with the problems may have a positive influence on student success. However, improving self-confidence will not directly improve a student's performance in chemistry; the interaction between the two is complex (Ajzen, 2002; Bauer, 2008; Bowman, 2012), and the correlations measured in this study cannot determine causation. In the end, improvement in the correlations observed will have a limit.

Although it would also be important to determine the areas in which there are the largest disagreements (as measured by the difference in average student and faculty rating), such an assessment cannot be made here. As seen above, the problems students find most difficult vary from year to year, though the statistical patterns persisted. Only a large, longitudinal study could determine what disagreements persist between years, if

any. Of course, the exact topics are likely to vary from classroom to classroom, as well. Therefore, it is much more important and useful for teachers to determine which topics are proving difficult in each individual class.

This study can, however, act as one potential method for evaluating current courses (they need not be chemistry) and their needed foci. Carter and Brickhouse (1989) believed their general findings to be transferable to other universities; the same is likely true here. Whether one chooses to measure the gap between faculty and student perception of difficulty by the method in this paper or by other means, the fundamental question, as always, is "*are we addressing students' needs?*" As this study showed, the best people to ask are the students.

Notes and references

a Sam Houston State University, Huntsville, Texas, USA.
Tel: (936) 294-1532; E-mail: ogulacar@shsu.edu
b Drexel University, Philadelphia, Pennsylvania, USA.
E-mail: bowmancr@drexel.edu

Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior. *J. Appl. Soc. Psychol., 32*(4), 665-683. doi: 10.1111/j.1559-1816.2002.tb00236.x

Bauer, C. F. (2008). Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts. *J. Chem. Educ., 85*(10), 1440-1445. doi: 10.1021/ed085p1440

Bowman, C. R. (2012). *Relationship Between Study Habits and Student Attitudes Towards Science and Technology.* (Ph.D. Doctoral Dissertation), Drexel University. Retrieved from http://hdl.handle.net/1860/3836

Carter, C. S., & Brickhouse, N. W. (1989). What makes chemistry difficult? Alternate perceptions. *J. Chem. Educ., 66*(3), 223-225. doi: 10.1021/ed066p223

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-Novice Differences in Perceiving and Processing Visual Classroom Information. *J. Teach. Educ., 39*(3), 25-31. doi: 10.1177/002248718803900306

Cooper, M. M., & Sandi-Urena, S. (2009). Design and Validation of an Instrument To Assess Metacognitive Skillfulness in Chemistry Problem Solving. *J. Chem. Educ., 86*(2), 240. doi: 10.1021/ed086p240

de Carvalho Filho, M. K. (2009). Confidence judgments in real classroom settings: monitoring performance in different types of tests. *Int. J. Psychol., 44*(2), 93-108. doi: 10.1080/00207590701436744

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*(3), 589-599. doi: 10.1093/biomet/68.3.589

Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc., 82*(397), 171-185. doi: 10.1080/01621459.1987.10478410

Field, A. (2009). *Discovering Statistics using SPSS.* London, England: Sage.

Gabel, D. (1998). The complexity of chemistry and implications for teaching. In B. J. Fraser & K. G. Tobin (Eds.), *International Handbook of Science Education* (pp. 233-248). London, UK: Kluwer Academic Publishers.

Grove, N., & Bretz, S. L. (2007). CHEMX: An Instrument To Assess Students' Cognitive Expectations for Learning Chemistry. *J. Chem. Educ., 84*(9), 1524. doi: 10.1021/ed084p1524

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *J. Educ. Psychol., 92*(1), 160-170. doi: 10.1037/0022-0663.92.1.160

Hoffmann, R., & Laszlo, R. (1991). Representation in chemistry. *Angew. Chem., Int. Ed. Engl., 30*(1), 1-16.

Hundhausen, C., Agarwal, P., Zollars, R., & Carter, A. (2011). Scaffolded Software for Improving Problem-Solving Skills. *J. Eng. Educ., 100*(3), 574-603.

Johnson, G. B. (2013). *Student Perceptions of the Flipped Classroom.* (Masters Thesis), University of British Columbia. Retrieved from http://hdl.handle.net/2429/44070

Johnstone, A. H. (1993). The development of chemistry teaching: A changing response to changing demand. *J. Chem. Educ., 70*(9), 701-705. doi: 10.1021/ed070p701

Johnstone, A. H. (2006). Chemical education research in Glasgow in perspective. *Chem. Educ. Res. Pract., 7*(2), 49-63. doi: 10.1039/B5RP90021B

King, D. B. (2011). Using Clickers To Identify the Muddiest Points in Large Chemistry Classes. *J. Chem. Educ., 88*(11), 1485-1488. doi: 10.1021/ed1004799

Knaus, K., Murphy, K., Blecking, A., & Holme, T. (2011). A Valid and Reliable Instrument for Cognitive Complexity Rating Assignment of Chemistry Exam Items. *J. Chem. Educ., 88*(5), 554-560. doi: 10.1021/ed900070y

Knaus, K. J., Murphy, K. L., & Holme, T. A. (2009). Designing Chemistry Practice Exams for Enhanced Benefits. An Instrument for Comparing Performance and Mental Effort Measures. *J. Chem. Educ., 86*(7), 827. doi: 10.1021/ed086p827

Lewis, E. L., & Linn, M. C. (2003). Heat Energy and Temperature Concepts of Adolescents, Adults, and Experts: Implications for Curricular Improvements. *J. Res. Sci. Teach., 40*(1), 155-175.

Moss, G. D., & McMillen, D. (1980). A strategy for developing problem-solving skills in large undergraduate classes. *Stud. High. Educ., 5*(2), 161-171. doi: 10.1080/03075078012331377196

Parker Siburt, C. J., Bissell, A. N., & MacPhail, R. A. (2011). Developing metacognitive and problem-solving skills through problem manipulation. *J. Chem. Educ., 88*(11), 1489-1495. doi: 10.1021/ed100891s

Pinarbasi, T., & Canpolat, N. (2003). Students' Understanding of Solution Chemistry Concepts. *J. Chem. Educ., 80*(11), 1328. doi: 10.1021/ed080p1328

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Beh. Res. Meth., 40*(3), 879-891. doi: 10.3758/BRM.40.3.879

Raker, J. R., Trate, J. M., Holme, T. A., & Murphy, K. (2013). Adaptation of an Instrument for Measuring the Cognitive Complexity of Organic Chemistry Exam Items. *J. Chem. Educ., 90*(10), 1290-1295. doi: 10.1021/ed400373c

Rosenthal, G. T., Soper, B., McKnight, R. R., Price, A. W., Boudreaux, M., & Rachal, K. C. (2010). Do Students Know If They Answered Particular Questions Correctly on a Psychology Exam? *J. Instr. Psychol., 37*(1), 57-62.

Schraw, G., & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemp. Educ. Psychol., 19*(4), 460-475. doi: 10.1006/ceps.1994.1033

Singh, K., & Xie, M. (2010). Bootstrap Method. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (Third ed., pp. 46-51): Elsevier.

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *J. Instr. Psychol., 22*(1), 77-88.

Smith, L. F. (2002). The Effects of Confidence and Perception of Test-taking Skills on Performance. *North Am. J. Psychol., 4*(1), 37-50.

Sözbilir, M. (2004). What Makes Physical Chemistry Difficult? Perceptions of Turkish Chemistry Undergraduates and Lecturers. *J. Chem. Educ., 81*(4), 573. doi: 10.1021/ed081p573

Symington, D., & Kirkwood, V. (1996). Lecturer Perceptions of Student Difficulties in a First-Year Chemistry Course. *J. Chem. Educ., 73*(4), 339-343. doi: 10.1021/ed073p339

Tucker, B. (2012). The flipped classroom. *Educ. Next, 12*(1), 82-83.

Williamson, V., Huffman, J., & Peck, L. (2004). Testing Students' Use of the Particulate Theory. *J. Chem. Educ., 81*(6), 891-896. doi: 10.1021/ed081p891

Woods, D. R. (2013). Problem-oriented learning, problem-based learning, problem-based synthesis, POGIL, PLTL, MEA and project-based learning: what's best for you? *Ind. Eng. Chem. Res.* doi: 10.1021/ie401202k

Appendix 1: Sample survey used to collect data for this study

SURVEY
GENERAL CHEMISTRY-I (1341)
TEST I

Name: _____

"Would you like to contribute to the development of science education and get bonus points?"
Then, please read the following section and provide as much information and insight as you can.

IMPORTANT NOTE:

- After you complete your test, please go back and check every question in the test and carefully provide the information required in this survey.

- There is no right/wrong answer for the questions in the survey. You will get the bonus points based on the *completeness, not the correctness*. However, it is very important that you share your very honest opinion for each question.

EXPLANATION FOR FILLING THE TABLES:

- First, for each **fill-in-the-blank, multiple-choice, & long answer question**, please assign *a number (1, 2, or 3) to each question* in the test *for each category;* Difficulty (D), Familiarity (F), and Self Confidence (SC) using the explanation below.

Difficulty (D)	Familiarity (F)	Self Confidence (SC)
1- Easy (solution requires remembering some basic definitions or facts)	*1- Very familiar (have seen and done several similar examples)*	*1- Extremely confident (know how to check and it is correct)*
2- Medium (solution requires formulas and some application)	*2- Somehow familiar (have seen before but have very little experience)*	*2- Somehow confident (think I got it right but not sure and do not know how to check)*
3- Difficult (solution requires linking among different concepts and a lot of calculations)	*3- Not familiar (have not seen at all before)*	*3-Very little confidence (have no idea if I got it right)*

- Then, in the next column entitled (**Topics/Concepts**) on the table below, write the numbers of each topic/concept from the list on the last page which you think are involved in each question. There is no limit to the number of topics you can list but please list the ones closely associated with the question.

- In the column at the far right, please share your insights, challenges you face with the question, general comments, etc.
    - o If you got stuck, please tell us where, why, and how.
    - o If you do not remember a specific piece of information, please tell us what it is.
    - o If you cannot do a certain operation, again please tell us and explain why.

Remember!!! The more specific information you provide the easier the analysis of data will be and the more efficiently the curriculum and teaching methods will be improved.

Qs *Fill in the Blanks*	D	F	SC	Topics/Concepts	Insights/ Comments
1					
2					
3					
4					
5					

Multiple Choice	D	F	S C	Topics/Concepts	Insights/ Comments
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

Long Answer	D	F	S C	Topics/Concepts	Insights/ Comments
1					
2					
3					
4					
5					
6					

**Please remember to complete each question carefully to get the bonus points added to your test score.**
We greatly appreciate your time and the information that you share which will lead to better instructional designs and developing more effective teaching materials.

THANKS!!!

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

TOPIC LIST

1. Classification of matter

2. Properties of matter

3. Measurement

4. Atomic structure

5. Periodic table

6. Formula calculations

7. Naming compounds

8. Chemical equations

9. Balancing equations

10. Formula weight

11. The mole

12. Mole/mass conversions

13. Mole/volume conversions

14. Mole/particle conversions

15. Mixed mole conversions

16. Empirical/Molecular formulae

17. Stoichiometric calculations

18. Limiting reactants

19. Theoretical yield