Title
The Impact of Flipped Classrooms on Student Performance and Confidence in an Introductory Organic Chemistry Course

Author
Hoang, Jason

Publication Date
2021-10-11

Data Availability
The data associated with this publication are within the manuscript.

THE IMPACT OF FLIPPED CLASSROOMS ON STUDENT PERFORMANCE AND CONFIDENCE IN AN INTRODUCTORY ORGANIC CHEMISTRY COURSE

By

Jason Hoang

A capstone project submitted for Graduation with University Honors

May 6, 2021

University Honors

University of California, Riverside

APPROVED

Dr. Matthew Casselman

Department of Chemistry

Dr. Richard Cardullo, Howard H Hays Jr. Chair

University Honors

ABSTRACT

Organic chemistry has always been, and continues to be, a source of anxiety and frustration for undergraduate STEM students, with many students often dreading the courses before they even take them. Flipped classrooms, in which a significant portion of instruction is completed at home on the student's own time and classroom activities are used in place of typical lecture, are a recent addition to education that has shown promise in bolstering student understanding. In this study, calibration between student confidence and performance will be examined, as the relationship between the two is an interaction that has yet to be extensively studied in prior research. Students were given two assessments to gauge their performance and their confidence levels with topics taught by flipped and non-flipped methods. Answer correctness and confidence will be combined in order to calculate calibration. The original hypothesis was that topics taught in a flipped format will calibration values closer to zero, which would demonstrate closer correlation of answer correctness and confidence. However, overall results demonstrated study participants scored higher and were more confident on the non-flipped topics than the flipped topics. However, the results also suggest that several different factors, such as inherent topic difficulty, may have unknowingly impacted calibration calculations. Future studies could continue to investigate calibration while factoring these components into their methodology.

ACKNOWLEDGEMENTS

I am extremely grateful for the guidance and mentorship from Dr. Matthew Casselman throughout the entirety of this project. Apart from the Capstone, his positive influence has been a constant factor in my undergraduate career ever since the beginning of my second year of college. He helped me gain a deep appreciation for organic chemistry, and he also reignited my passion for teaching. Thank you so much for being an exceptional professor, leader, and mentor.

I would also like to thank Kiersten Marymee, my fellow Honors mentee under Dr. Casselman, for her help during this project. I greatly appreciated your assistance and diligence, which in turn helped me stay on track with my own Capstone.

Finally, I would like to express my gratitude for my University Honors Counselor, Dennis McIver. He has been a constant source of support and positivity during my whole time as an Honors student.

TABLE OF CONTENTS

Introduction: Rationale

Organic chemistry has always been considered a difficult subject to learn and master, as it requires students to think in novel and unusual ways due to many intrinsic and extrinsic factors (O'Dwyer & Childs, 2017). The subject has always been a source of anxiety and frustration for undergraduate STEM students, with many students often dreading the courses before they even take them. In my first year at University of California Riverside (UCR), the looming threat of organic chemistry plagued me. I recall having an easy time in general chemistry, thanks to an extensive background of high school coursework, such as Honors Chemistry and Advanced Placement Chemistry, that played a significant role in developing knowledge about general chemistry prior to university education. Many students do not obtain a foundation in organic chemistry prior to college, which puts most students on a level playing field. This challenging course with students of similar academic preparation makes it an optimal candidate to study pedagogy, since there is less chance of bias and differences in preparation influencing students' performance. In my case, I will investigate the effects that the flipped classroom method has on student performance and confidence in the context of an organic chemistry course.

Organic chemistry is a discipline that often serves as a prerequisite for more specialized STEM fields; therefore, a strong command of the subject is integral for future success in multiple STEM majors. Researchers should be looking for ways to ensure that the basic foundations are firmly grasped to better prepare the next generation of STEM researchers and workers. Researchers should search for ways to make the field more appealing to encourage more people to prepare for a STEM career. If education researchers and STEM instructors can make the subject more approachable and effective for novice learners, much of that struggle will be alleviated. My project will have potential benefits in the field of organic chemistry education, because it may help identify what features of a flipped classroom have the greatest benefits for students.

Introduction: Literature Review

The concept of a flipped classroom is not new; many studies have been done to determine their usefulness and effectiveness, and some instructors have even begun to incorporate flipped methods into their course plans. Research has shown that exam performance significantly improves due to the higher level of student interaction with the material (Gross *et al.*, 2015). In a course taught in a flipped mode, students are exposed to the instructional material outside of class, freeing up time in class for other activities to reinforce and practice the material learned at home. This is in contrast to a traditional classroom where students are presented with material in lecture and expected to practice and engage that material independently outside of class. A recent randomized-controlled study found that the online learning component of the flipped classroom accounted for most of the learning gains in a lesson on stereochemistry (Casselman *et al.*, 2019). These findings and others strongly suggest that flipped classrooms are beneficial for student learning as evidenced by greater exam scores and final grades and merit wider implementation in the future.

Student performance is the typical metric that is used to assess new instructional methods. It is clear why that is the case as it can easily be measured by analyzing numerical scores on assessments. However, there are other metrics that can be measured. For example, researcher Jeremy F. Strayer has conducted a study investigating the effects of flipped classroom methods on student cooperation, innovation, and task orientation (Strayer, 2012). He found that students in flipped classrooms were more willing to cooperate with one another and engage in different activities in the classroom. Student confidence and its connection with performance will be probed as the relationship between performance and confidence is an interaction that still has yet to be extensively studied. One study that has addressed this metric serves as inspiration for this project was one completed within an introductory geology course (Jones *et al.*, 2019). In this study, the authors implemented a partially flipped classroom and used active learning strategies in the actual in-class time. They compared student performance and confidence across different semesters, and they found out that active learning strategies are more integral to student success than simply just flipping the classroom. They generally observed increases in both performance and confidence in throughout the different experimental groups, except for a slight but statistically significant drop in confidence in the group that had concepts that were only shown in pre-class videos with no in-class reinforcement. This result is reinforced by another study where a meta-analysis was done on active learning, and the authors discovered that "active learning increases examination performance by just under half a SD" (Freeman *et al.*, 2014).

A common limitation of both studies is that they do not investigate the correlation between performance and confidence; they just look at the independent effects that a flipped classroom has on those factors. In another study, this relationship was examined, and researchers developed a coefficient, known as the rho coefficient, to quantify the calibration between performance and confidence (Dinsmore & Parkinson, 2013). They found that the rho coefficient scatter plots were able to successfully give some indication on how those two factors were linked. Higher rho coefficients tended to yield plots that showed a stronger linear relation between confidence and performance, which indicates that the subjects were more highly calibrated. Priscilla Bell and David Volckmann took a slightly different approach; they mainly focused on performance and confidence in the context of the Dunning-Kruger effect (Bell & Volckmann, 2011). Their study suggested that students who do better on examinations estimate their performance accurately, while lower-performing students are overconfident. Because of this, they believe that high correlations between confidence and performance are valid reflections of the students' actual knowledge. However, the study was done on general chemistry courses, so there is still room to explore the calibration in organic chemistry.

Introduction: Setting

The study was conducted in CHEM 008A, an established course at the University of California, Riverside. CHEM 008A is the first part in a three-part sequence of organic chemistry; as such, it has the extremely important role of establishing the framework for the two courses that succeed it. Despite the introductory level of the subject, the course can prove to be a source of trouble for many students due to its fast pace and abundant quantity of material. Over the ten weeks of instruction, CHEM 008A covers topics including: principles of bonding in organic compounds, functional groups and IR spectroscopy, acid-base chemistry, conformational analysis, stereochemistry, nucleophilic reactions, elimination reactions, retrosynthesis, addition reactions, and radical reactions. CHEM 008A and the rest of the organic chemistry series are required by many of the undergraduate majors in the UCR College of Natural and Agricultural Sciences (CNAS). A passing grade of "C-" or better is needed to advance to CHEM 008B. This requirement for many majors points to the importance of organic chemistry in many of the student's upper-division courses in CNAS.

CHEM 008A is offered in Fall, Winter, and Summer quarters at UCR. In any given academic quarter where CHEM 008A is offered, the specific offerings of the course may differ slightly from each other due to the changing availabilities of professors in the Department of

Chemistry. One of the professors who frequently teaches the course is Dr. Matthew Casselman, an Assistant Professor of Teaching in the Department of Chemistry at UCR. While the majority of the organic chemistry professors adopt a more traditional lecture approach in their classrooms, Dr. Casselman utilizes a significant amount of flipped classroom methods in his instruction, where roughly one-third of all class periods is a lecture tutorial activity. He has also conducted extensive research in the pedagogy of organic chemistry with an emphasis on increasing student engagement in the classroom. As a former student of Dr. Casselman's CHEM 008A course during the Fall 2018 academic quarter, I have firsthand experience of the structure and intricacies of the class. After realizing that my success in the CHEM 008A and the following organic chemistry courses was largely due to the effective usage of flipped classroom methods, I became impassioned about conducting research to make organic chemistry even more accessible and engaging for the next generation of students.

Methods and Materials

Due to the unexpected COVID-19 pandemic and resulting instruction being taught in a remote mode throughout the 2020-2021 academic year, this study was adapted to be conducted online. This study was reviewed and approved by the UCR Institutional Review Board Socio-Behavioral as Exempt under protocol HS 20-182. Assessments were administered as online quizzes via Google Forms. The participants were presented 30 content questions (split over two sessions) in a multiple-choice format. These questions consisted of different topics covered in Dr. Casselman's CHEM 008A course. His class is structured to contain both traditional lecture and flipped classroom activities, so some questions were from topics taught using a flipped method while others were taught in a more traditional lecture format. The multiple-choice assessments were administered at the beginning of two optional review sessions during the eighth and ninth weeks of the academic quarter. Students provided consent to participate in this study by electronically signing a Docusign form. In addition to consenting to have researchers analyze their assessment data, it also granted access to educational records, including final class grades. All responses to the assessments remained confidential, and Dr. Casselman, the faculty mentor, did not have access to any identifying data that could link an assessment response to the actual student who completed the assessment until after final course grades were assigned. Furthermore, although the multiple-choice assessments were scored using the function within Google Forms, the scores had no impact on the students' grades in CHEM 008A itself.

The quiz given in session #1 consisted of 14 questions, while the quiz given in session #2 consisted of 16 questions. Each question had two parts: "A" and "B". Part "A" tested the participants' knowledge of class material in a traditional multiple-choice format, offering four choices for the participants to choose from (Figure 1). They were instructed to choose the best answer out of the four available choices and not to leave any questions blank.

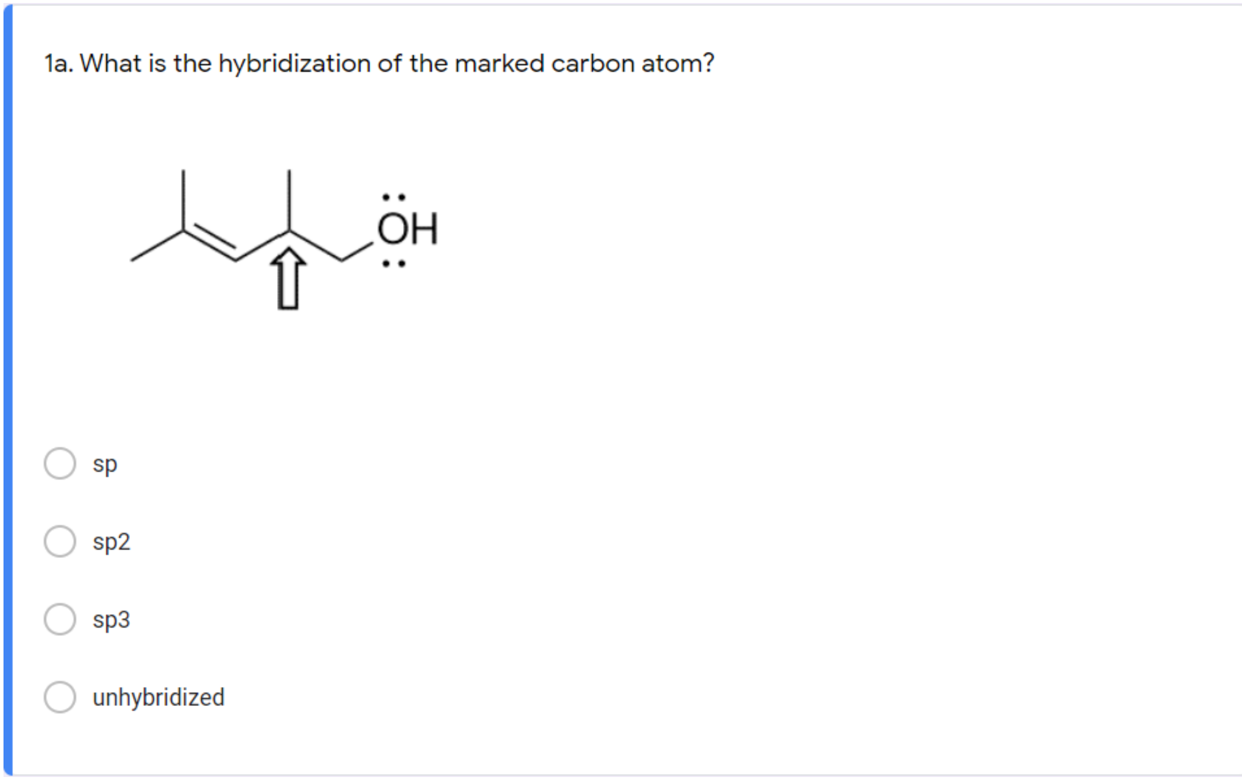

Figure 1. *An example of a question from the assessments that were administered to the participants. This question is the first question in the quiz that was given during Session #1.*

Part “B” consisted of a scale where students were to self-report their confidence in their answer (Figure 2). Students were asked to use integer values between 0 and 10 to gauge their overall confidence in their answer. Participants selecting “0” had indicated that they were not confident at all in their answer and that it was no better than a random guess, while participants selecting “10” indicated that they were 100% confident in the correctness of their answer.

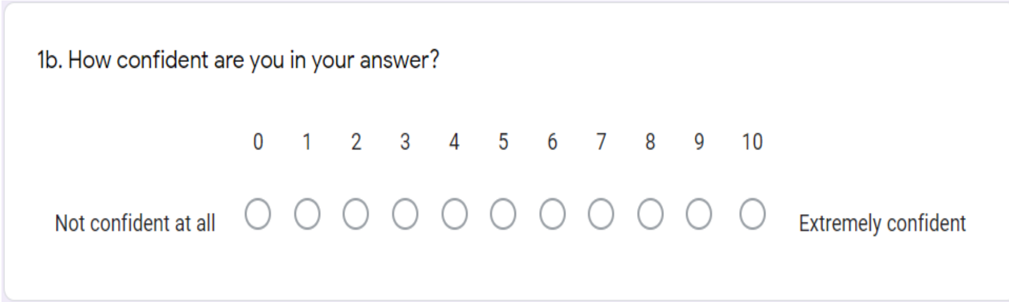

Figure 2. *Part "B" of the same question shown in Figure 1. This part asks students to rate their confidence in their answer.*

As previously mentioned, these assessments were administered online during Weeks 8 and 9 of the Fall 2020 academic quarter. At 6:50 PM on the days of the review sessions, a Zoom meeting was created and open for participants to join. A quick debriefing was provided to the students by the researchers, detailing the specific details about how to complete the assessment. At 7:00 PM, the researchers released the link to the Google Form assessment into the Zoom chat and allowed participants to begin the assessment. Thirty minutes were allotted for the completion of the quiz, and reminders were given at regular intervals towards the end of the allotted time. At 7:30 PM, the assessment was locked by the researchers, preventing further submissions past this point. After completion of the assessment, the study then transitioned to a regular review session standard for Dr. Casselman's CHEM 008A class. The professor reviewed the answers to the assessment questions and went into detail about the topics covered.

Data and Results

Results were initially analyzed from both sessions where flipped and non-flipped items were compared. The mean results for the flipped items are detailed in Table 1, and individual item level correctness and confidence is shown in Figure 3. One clear discrepancy as noted in Figure 3 is the omissions of items S1Q7 and S1Q11, both of which are flipped-topic questions. While review the answers to the assessments during the review session, we discovered that the correct answers to these questions were not one of the possible choices in the quiz. Due to the oversight during creation of the Session 1 assessment, items S1Q7 and S1Q11 were removed from the final results. The flipped items showed a moderate level of mean correctness (51.43%) and mean confidence (0.4576).

Table 1*: Flipped Questions, Both Sessions Combined*

Flipped Questions	13 questions
Mean correctness	51.43% = 0.5143
Mean confidence*	0.4576
Calibration	-5.67%

* Mean confidence was converted from a 0-10 scale into a 0-1 decimal value.

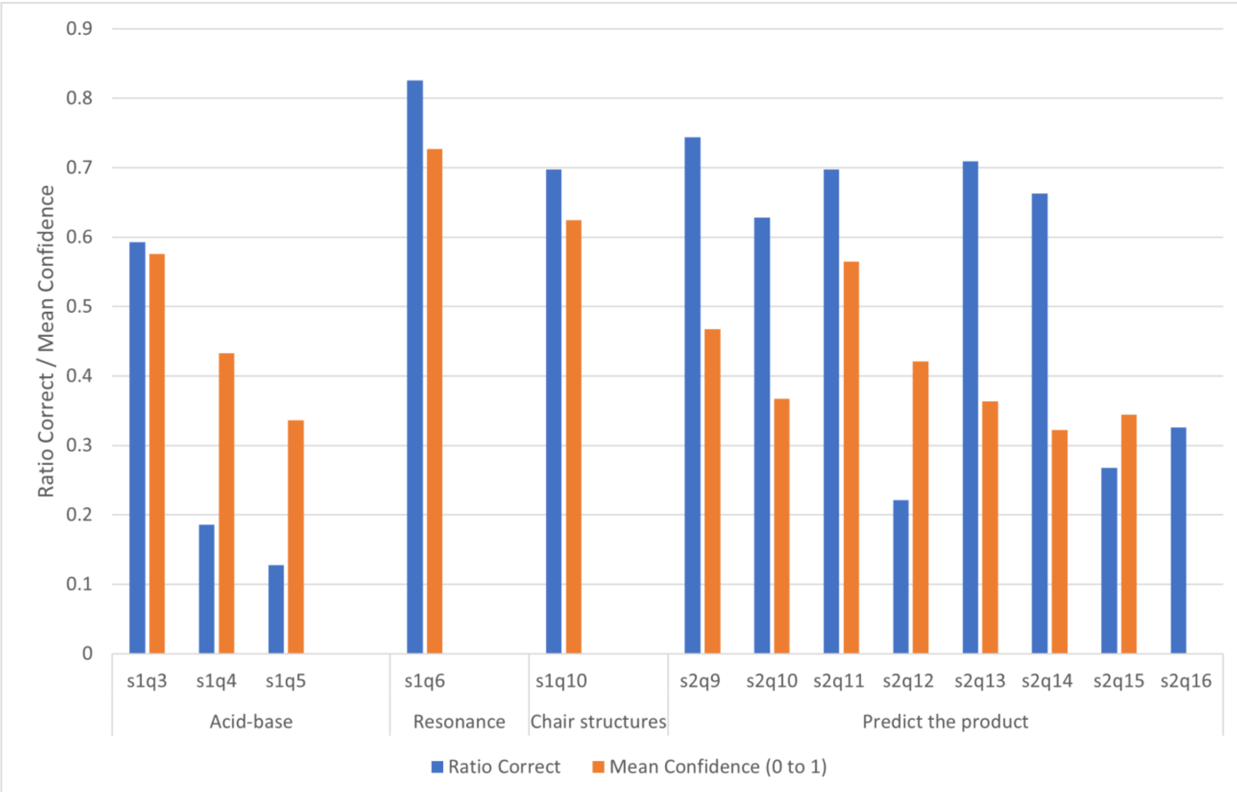

Figure 3*: Correctness and confidence on items taught in by a flipped method.*

Calibration was calculated by taking the mean confidence (converted to a scale from 0 to 1) and subtracting the mean correctness for each question; question level calibration is shown in Figure 4. This value is then multiplied by 100%. For example, for item S1Q3, a question about acid-base reactions, the calibration was calculated:

$$(0.5756) – (0.5930) = -0.0174 \text{ x } 100\% = 1.74\%$$

A negative calibration value represents a participant's under-confidence in their answer, meaning that a student felt their answer was incorrect when in fact it was correct. For example, for item S1Q3, students were about 1.74% under-confident in their answers, implying that they performed better than they expected. Conversely, a positive calibration value shows that a participant is over-confident in their answer, meaning that they believed their answer to be correct when in fact it was not.

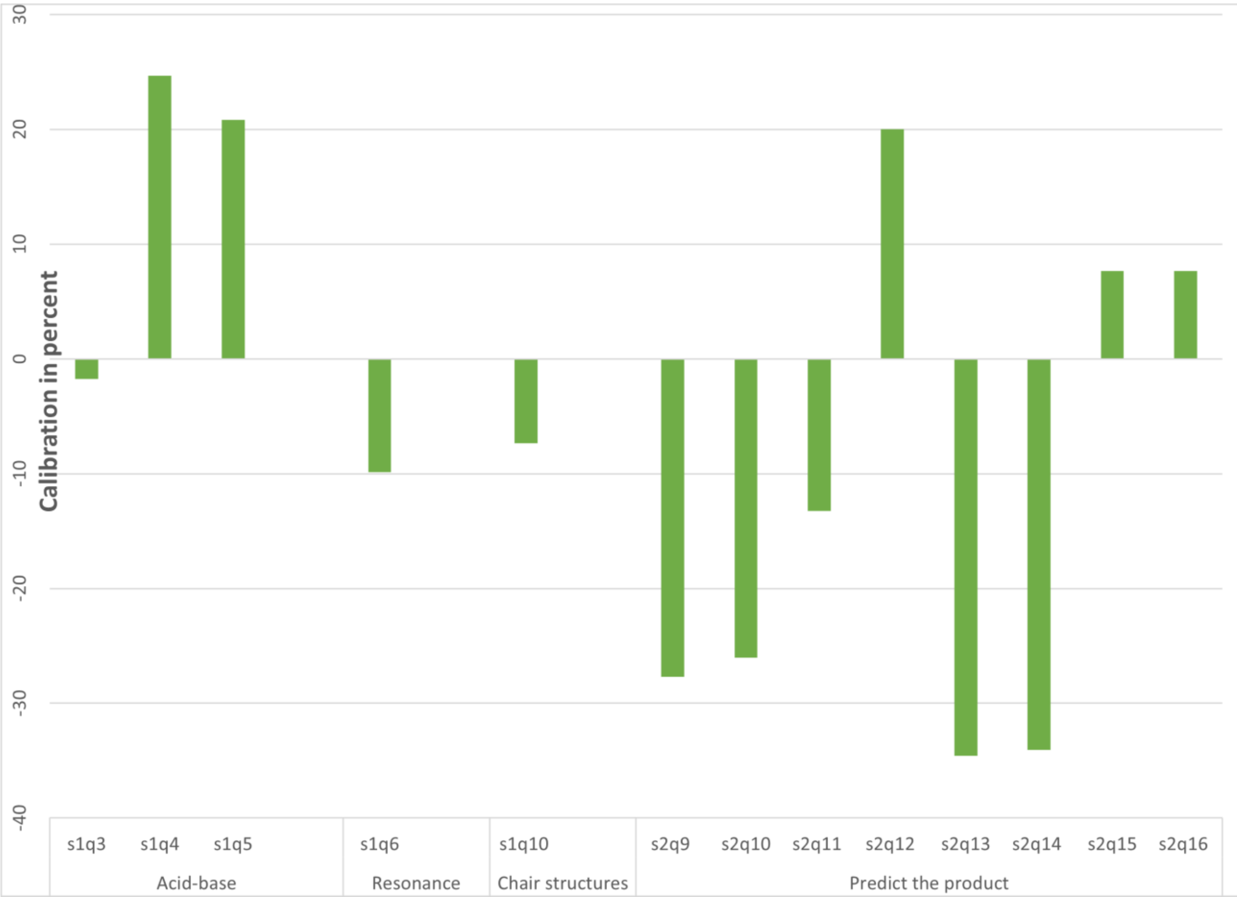

***Figure 4**: Calibration for Flipped Questions, Both Sessions Combined. Positive calibration values (bars above the x-axis) indicate that a student is overconfident in their answer, while negative calibration values (bars below the x-axis) indicate that a student is underconfident in their answer.*

When examining Table 1, we can gain an overall view of the flipped-topic questions. On average, the participants scored about 51.43% correct on the flipped questions. When converted to a scale from 0 to 1, with 0 being "not confident at all" and 1 being "extremely confident", the participants reported an average confidence of about 0.46. After calculating the calibration for questions related to flipped topics, the participants were observed to be 5.67% underconfident in their answers when compared to their actual performance.

However, when examined closer within the different question topics, some distinct findings can be observed. For two of the acid-base questions, participants were extremely overconfident in their answers (>20%). As a whole, the participants were generally underconfident in their answers for the "predict the product" questions, with the notable exception of item S1Q12.

Mean results for non-flipped questions are presented in Table 2. On average, the participants scored about 68.14% correct on the non-flipped questions. When converted to a confidence scale of 0 to 1, they reported an average confidence of about 0.6847. The calibration shows that the participants were nearly perfectly calibrated with only 0.33% overconfidence in their answers when compared to their actual performance. This suggests that the participants were reasonably accurate while gauging their level of knowledge during the non-flipped questions.

Table 2*: Non-flipped Questions. Both Sessions Combined*

Non-flipped Questions	15 questions
Mean correctness	68.14% = 0.6814
Mean confidence *	0.6847
Calibration	0.33%

* Mean confidence was converted from a 0-10 scale into a 0-1 decimal value.

As before, the question level correctness, confidence and calibration showed significant variation (Figure 5 and 6). Examining the breakdown of different topics may elucidate different patterns. Since there were no errors with the development of the non-flipped questions, every category of question topic has at least members in the group. On the “hybridization” and “stereochemistry” questions, participants were generally overconfident in their answers. Two of the “stereochemistry” questions, S1Q13 and S1Q14, interestingly had the most positive calibration values in the entire study. This implies that the participants were very overconfident in their answers for these questions. In contrast, participants were entirely underconfident in their answers for the “reaction roles”, “mechanistic steps”, and “reaction energy diagram” questions. For both of the “reaction energy diagram” questions, participants were remarkably underconfident in their answers (>20%).

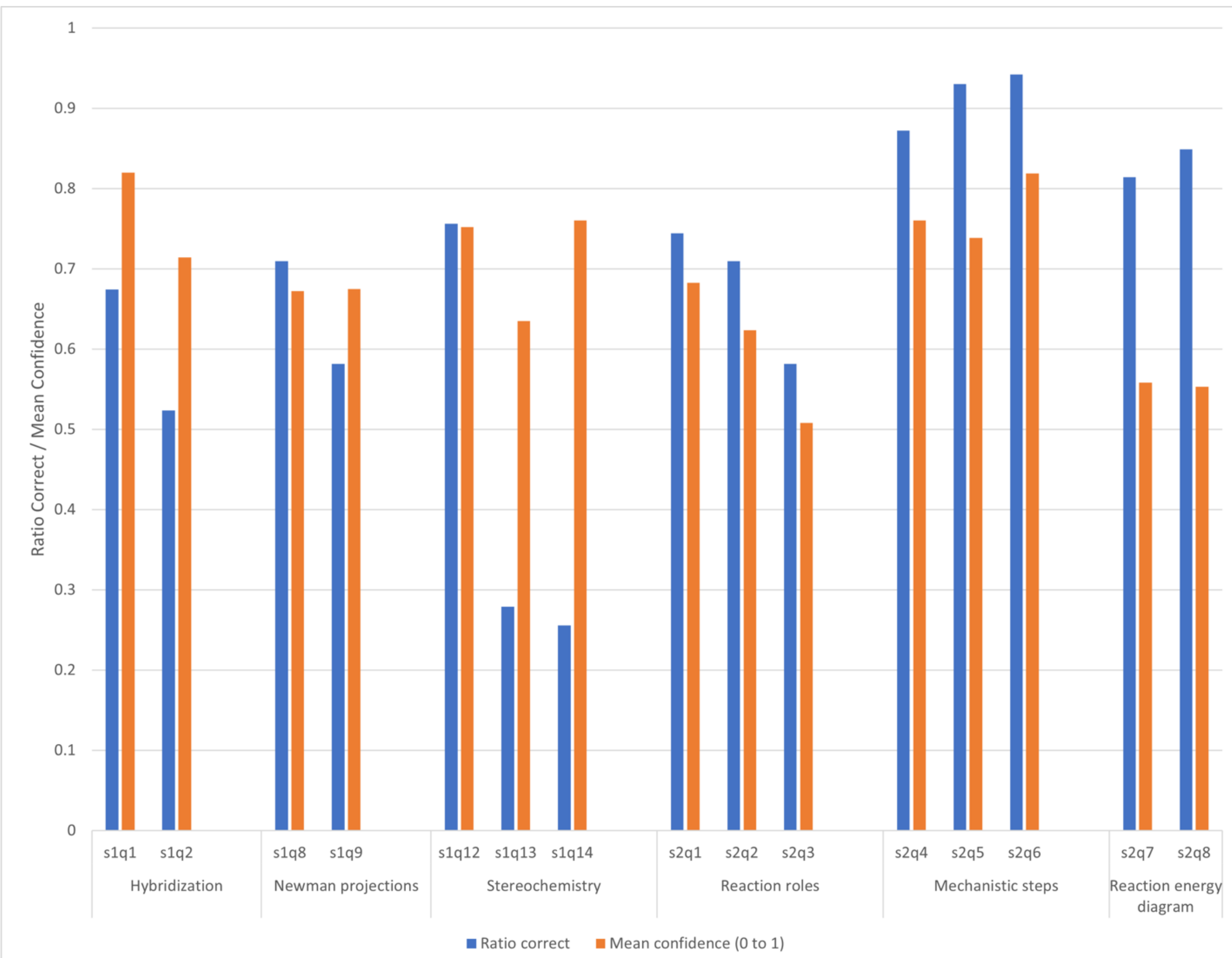

Figure 5*: Non-flipped Questions, Both Sessions Combined*

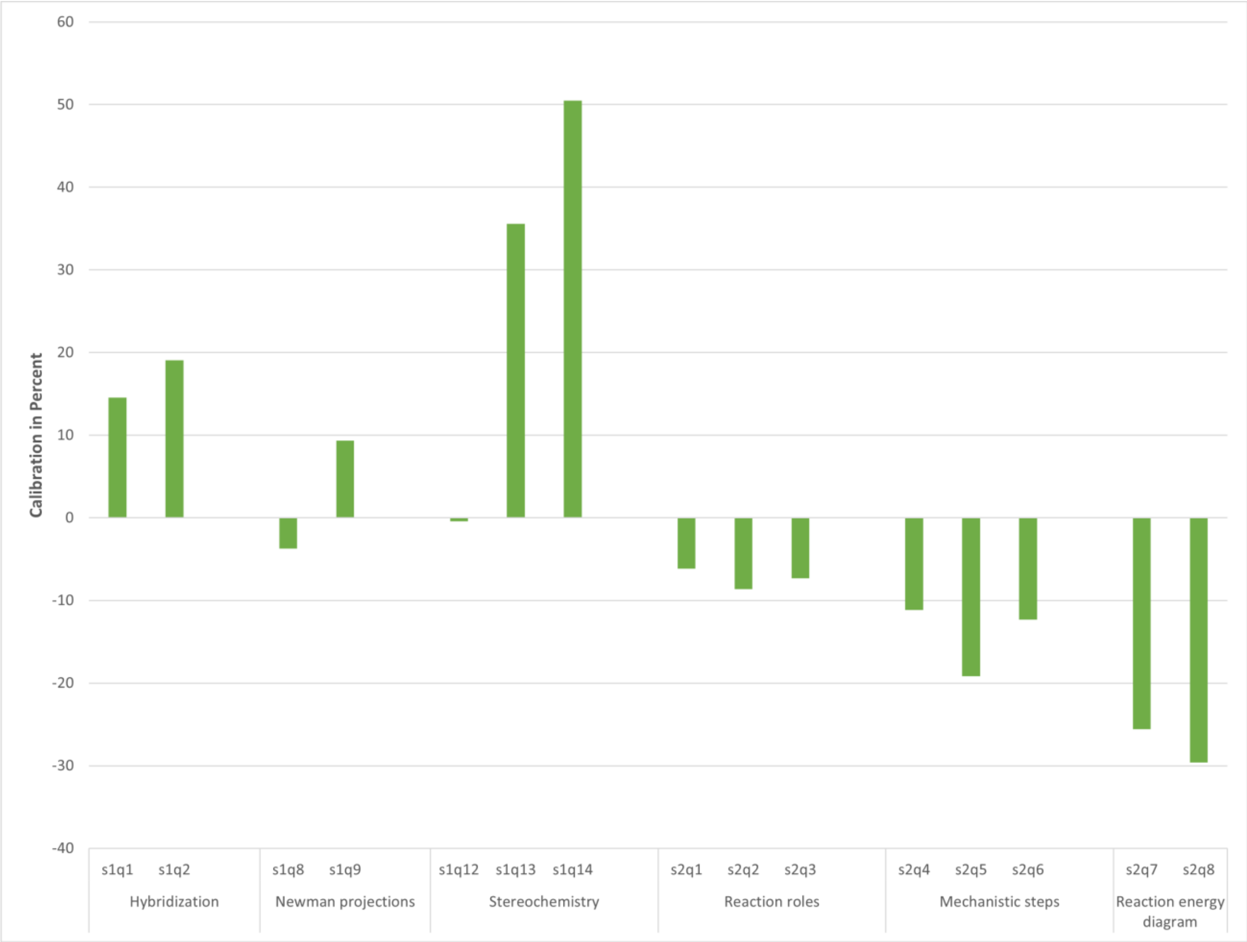

***Figure 6**: Calibration for Non-flipped Questions, Both Sessions Combined*

Another way to examine the data is to group it in a chronological manner, instead of simply looking it at in terms of flipped and non-flipped questions. As seen in Table 3, the participants scored about 51.7% correct on Session 1 questions. The calibration calculation shows that the participants were 12.6% overconfident in their answers when compared to their actual performance. Conversely, Table 3 also shows that participants scored about 66.9% correct on Session 2 questions. They were about 13.8% underconfident in their answers.

Table 3*: Comparing Session 1 Questions and Session 2 Questions*

	Session 1 Questions (12 total)	Session 2 Questions (16 total)
Mean correctness	51.74% = 0.5174	66.86% = 0.6686
Mean confidence*	0.6436	0.5310
Calibration	12.62%	-13.76%

* Mean confidence was converted from a 0-10 scale into a 0-1 decimal value.

The participant and general class populations were analyzed by calculating the mean course GPA for each group by considering final letter grades. The study participants (N=86) had a mean course GPA of 3.55, while the course as a whole (N=614) had a mean course GPA of 2.78. Both were reported on a 4-point scale.

Discussion

When comparing flipped to non-flipped topics, calibration was not widely different between these two sub-sets of data. Student calibration on flipped topics was -5.6% while it was +0.3% on non-flipped topics (Table 1 and 2). This is a surprising difference as the hypothesis originally stated that the flipped questions should have a more accurate calibration (value closer to zero). This may have been affected by a number of different factors related to how the study was completed. Within the overarching designation of “flipped” or “non-flipped” questions, disparities can be found where students performed poorly on some topics, such as “stereochemistry”, while doing well in others. This had an impact on the calibration factors calculated, and this comparing flipped and non-flipped averages may result in missing key differences in calibration. This may be due to the inherent difficulty of a specific topic. It is important to realize that not all topics in organic chemistry are created equally; some concepts may simply more difficult to grasp than others. The difference in inherent difficulty may overpower the influence of flipped classroom instruction.

Even within the individual topics, differences in performance, confidence, and calibration can be found. This is most clearly observed in the "predict the product" questions from Session #2, which consist of 8 flipped questions. Calibration on this topic ranges from -35% to +20%, with a mean calibration of -12.5%. This suggests that this topic is potentially quite challenging for students (and in my experience as a former student in this class). In terms of calibration, the students ranged from remarkably underconfident to remarkably overconfident in their answers in this topic which may suggest that some assessment items may not be accurately measuring student understanding or confidence. In terms of their raw performance, participants scored a high of 74% on S2Q9 and a low of 22% on S2Q12. It is possible that students were simply more adept at memorizing a specific reagent or mechanism than another one. Even though the scope of the questions in these assessments were limited to elimination and substitution reactions, it can be difficult to keep track of the many different reagents that are covered in the course.

The study may be limited by the number of questions asked for each topic on the assessments. It may have been advantageous to limit the number of topics in each assessment so that more accurate means could be calculated. However, because the assessments had to be designed to fit within the constraints of a typical class review session, the freedom to design questions was slightly limited. Furthermore, as previously mentioned, the data for the flipped questions was fragmented due to some errors while designing the questions, which resulted in topics related to resonance and chair structures being limited to one item each. These factors make it difficult to assess whether the inherent difficulty of a certain topic affected the students' performance and confidence, outside of the independent variables tested in the experiment.

Instead, if the data is analyzed by chronology rather than instructional mode, a different trend becomes apparent (Table 3). Session #1 of the experiment was administered during Week #8 of the academic quarter, and all of its topics were covered much earlier in the course. Session #2 was administered one week after Session #1, and its topics were covered more recently in the course. Because of the length of time since the material was first introduced, students may have forgotten most of the material for Session #1, which resulted in a lower overall correctness score. However, because the material was likely perceived to be basic or fundamental, student confidence in the material remained high. This resulted in a higher calculated calibration for Session #1 at +12.62% (over-confidence). The material for Session #2 was more recently covered in class, so the material may have been fresh in their minds. This may explain why students scored higher in terms of correctness on Session #2 material. However, this material being newer and necessarily more complex might have been perceived as difficult, resulting in a lower confidence score and a lower calibration at -13.76% (under-confidence).

In order to determine if these results are generalizable, we compared the participant group to the class as a whole. The average participant final grade earned was a 3.55 (A-/B+), and the average class final grade was calculated to be a 2.78 (B-). A high percentage of the participants (47.7%) achieved a grade of A or A+ in the class by the end of the quarter. This is uncharacteristic for an introductory organic chemistry course where only 21.8% of the overall class achieved this final grade in fall quarter 2020. As a result, the participant population may not be necessarily representative of the class as a whole. In the future, it may be beneficial to conduct this research within the class itself to ensure a more representative population or to analyze groups organized by final letter grade to determine if any trends in correctness, confidence and calibration can be observed. As previously mentioned, the Dunning-Kruger effect suggests that high-achieving students will estimate their performance more accurately than low-achieving students (Bell & Volckmann, 2011). If the group of participants from our study truly is more successful than the overall class, then the real calibration values may stray further away from the baseline than expected. From the entire pool of students taking the class for the quarter, the group of participants was narrowed down at a few different checkpoints. The participants had to be willing to invest extra time to attend the optional review sessions; simply choosing to take this step may suggest that they are generally willing to put more effort in their studies. In addition, some potential candidates for the study were eventually excluded from the results for not showing up to both sessions or not completing the required consent form. If the results were examined with a larger group, they might have become markedly different.

Conclusions

It was originally hypothesized that the participants' performance and confidence would be greater in the flipped topics than the non-flipped topics, but ultimately the results of this study suggest that the opposite may be true. Overall, participants scored better and were more confident on the non-flipped questions than the flipped questions. They were slightly overconfident on the non-flipped questions, and they were underconfident on the flipped questions. This ultimately impacted the calibration values determined for flipped and non-flipped items. These results may be complicated by several factors, such as differences in inherent topic difficulty, the writing of specific questions, or the level of familiarity with the material. Some of the variables are easier to pinpoint, such as the errors during preparation of the questions and limited representation of some course topics. Other variables are more difficult to discern; for example, the looming presence of the COVID-19 pandemic was a factor throughout the entirety of this study. The participants of this study, along with the rest of their peers at UCR, had to quickly adapt to learning in a purely online format. The distinction between flipped and non-flipped may have become unusually blurred and the benefits of a flipped classroom may not be as readily apparent when classes were remotely taught. However, this study does show that variation in calibration values does exist between flipped and non-flipped modes, as well as variation by subject matter. Ideally, future research can revisit the topic of the calibration between performance and confidence. For example, calibration can be studied within the context of other variables such as the inherent level of difficulty of a topic. By limiting a study to topics around the same level of difficulty, the effect of the difference in difficulty can be greatly reduced. With normal circumstances and a more structured experimental design, researchers can continue examining the effectiveness of flipped classroom methods.

REFERENCES

Bell, P., & Volckmann, D. (2011). Knowledge Surveys in General Chemistry: Confidence, Overconfidence, and Performance. *Journal of Chemical Education, 88*(11), 1469–1476. https://doi.org/10.1021/ed100328c

Casselman, M. D., Atit, K., Henbest, G., Guregyan, C., Mortezaei, K., & Eichler, J. F. (2019). Dissecting the Flipped Classroom: Using a Randomized Controlled Trial Experiment to Determine When Student Learning Occurs. *Journal of Chemical Education, 97*(1), 27-35. https://doi.org/10.1021/acs.jchemed.9b00767

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased Preclass Preparation Underlies Student Outcome Improvement in the Flipped Classroom. *CBE—Life Sciences Education, 14*(4), ar36. https://doi.org/10.1187/cbe.15-02-0040

Jones, J. P., McConnell, D. A., Wiggen, J. L., & Bedward, J. (2019). Effects of classroom "flipping" on content mastery and student confidence in an introductory physical geology course. *Journal of Geoscience Education, 67*(3), 195–210. https://doi.org/10.1080/10899995.2019.1568854

O'Dwyer, A., & Childs, P. E. (2017). Who says Organic Chemistry is Difficult? Exploring Perspectives and Perceptions. *EURASIA Journal of Mathematics, Science and Technology Education, 13*. https://doi.org/10.12973/eurasia.2017.00748a

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*, 171–193. https://doi.org/10.1007/s10984-012-9108-4

***Appendix A**: Questions with Type, Mean Confidence, Mean Correctness, and Calibration*

Question	**Instructional Mode** (Flipped or Non-Flipped)	**Mean Confidence** (converted from 0-10 scale to 0-1 scale)	**Mean Correctness** (as a percentage)	**Calibration**
S1Q1	Non-Flipped	0.8200	67.44%	14.56%
S1Q2	Non-Flipped	0.7140	52.33%	19.07%
S1Q3	Flipped	0.5756	59.30%	-1.74%
S1Q4	Flipped	0.4329	18.60%	24.69%
S1Q5	Flipped	0.3360	12.79%	20.81%
S1Q6	Flipped	0.7267	82.56%	-9.89%
S1Q8	Non-Flipped	0.6721	70.93%	-3.72%
S1Q9	Non-Flipped	0.6747	58.14%	9.33%
S1Q10	Flipped	0.6244	69.77%	-7.33%
S1Q12	Non-Flipped	0.7518	75.58%	-0.40%
S1Q13	Non-Flipped	0.6349	27.91%	35.58%
S1Q14	Non-Flipped	0.7605	25.58%	50.47%
S2Q1	Non-Flipped	0.6826	74.42%	-6.16%
S2Q2	Non-Flipped	0.6233	70.93%	-8.60%
S2Q3	Non-Flipped	0.5081	58.14%	-7.33%
S2Q4	Non-Flipped	0.7605	87.21%	-11.16%
S2Q5	Non-Flipped	0.7384	93.02%	-19.18%
S2Q6	Non-Flipped	0.8186	94.19%	-12.33%
S2Q7	Non-Flipped	0.5581	81.40%	-25.59%
S2Q8	Non-Flipped	0.5529	84.88%	-29.59%
S2Q9	Flipped	0.4674	74.42%	-27.68%
S2Q10	Flipped	0.3674	62.79%	-26.05%
S2Q11	Flipped	0.5651	69.77%	-13.19%
S2Q12	Flipped	0.4209	22.09%	20.00%
S2Q13	Flipped	0.3635	70.93%	-34.58%
S2Q14	Flipped	0.3224	66.28%	-34.04%
S2Q15	Flipped	0.3442	26.74%	7.68%
S2Q16	Flipped	0.4023	32.56%	7.67%