Chemistry Education Research and Practice

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c6rp00195e

Impact of instructional decisions on the effectiveness of cooperative learning in chemistry through meta-analysis

Andrew Apugliese and Scott E. Lewis*

Received 20th September 2016,
Accepted 5th December 2016

DOI: 10.1039/c6rp00195e

www.rsc.org/cerp

Meta-analysis can provide a robust description of the impact of educational reforms and also offer an opportunity to explore the conditions where such reforms are more or less effective. This article describes a meta-analysis on the impact of cooperative learning on students' chemistry understanding. Modifiers in the meta-analysis are purposefully chosen to model instructors' decisions in implementing cooperative learning. Modifiers investigated include: using cooperative learning periodically or in every class period; setting a maximum group size at four or smaller *versus* five or larger; using closed-ended or open-ended assessments; and assessing a single topic or assessing the cumulative topics in the course. The results showed cooperative learning's effectiveness is robust across a wide range of instructional decisions except no evidence of effectiveness was found with cumulative assessments. The overall results from the meta-analysis provide a benchmark for evaluating future efforts to evaluate pedagogical interventions in chemistry.

Introduction

Meta-analyses offer a systemic means to quantitatively review the research literature. Briefly described, a meta-analysis identifies a collection of prior research articles that fit a set of common criteria, and treats the results in each article as a data point. Conducting a meta-analysis offers benefits beyond what can be accomplished in an individual study. By analyzing multiple data sets, reproducibility of effects can be investigated. Similarly, by treating each study as a unique data point, meta-analysis can provide a ready claim to the independence of observations assumption that underlies inferential statistics. Additionally, meta-analysis can facilitate an investigation into which conditions a particular treatment is effective by comparing sub-groups of studies within the analysis. This study describes a meta-analysis that seeks to identify the effectiveness of cooperative learning (CL) under a variety of conditions determined by instructor discretion.

Prior work on cooperative learning effectiveness

CL, or structured group work, is a pedagogical technique that facilitates students actively engaging with the content and communicating the content to their fellow peers. Johnson and Johnson (2009) have written extensively about the essential characteristics and effectiveness of CL. They describe the essential characteristics as: positive interdependence, individual accountability, promotive or encouraging interactions and the use of social and process skills (Johnson and Johnson, 2009). In a meta-analysis, Johnson and Johnson report on the effectiveness of CL over individualistic instruction, such as lecture, and found an average effect size of 0.53 among individuals 18 years and over (Johnson *et al.*, 1998).

Specific to science, technology, engineering and math (STEM) education, Bowen conducted a meta-analysis of 37 research studies, published between 1980 and 1996, on the effectiveness of CL in undergraduate STEM courses (Bowen, 2000). Across the 37 studies, 49 effect sizes were calculated and averaged to find an effect size of 0.51. The treatment of multiple effect sizes from a single study (*e.g.* data from multiple tests were reported) as separate data points provides greater weight to those studies relative to the studies with only one effect size reported. The decision to treat each effect size as a separate data point can be problematic as studies with multiple effect sizes still represent only one independent sample.

Recently, Warfa conducted a meta-analysis on the impact of CL in chemistry classes (Warfa, 2016). The analysis began by reviewing the literature from 2001 through 2015 using the search terms "cooperative learning" and "chemistry" and either "treatment group" or "control group". The resulting articles were screened for the following characteristics: studies authored in English, occurred within a face-to-face chemistry

Department of Chemistry, Center for the Improvement of Teaching & Research in Undergraduate STEM Education, University of South Florida, Tampa, FL 33620, USA. E-mail: slewis@usf.edu

classroom setting, had an outcome measure of chemistry achievement, used experimental/quasiexperimental research design that compared CL to a control pedagogy and provided sufficient statistical information to enable analysis (Warfa, 2016). These criteria identified 25 articles from the research literature. Multiple data points from one sample within a study were combined for a single effect size. One study (Acar and Tarhan, 2008) was removed from the analysis as an outlier and another study contributed two data points as it listed two independent samples, one composed of General Chemistry students and another Organic Chemistry students (Chase *et al.*, 2013).

Each of the effect sizes of the 25 independent samples were converted to Hedges's g values to remove a positive bias associated with Cohen's d when sample sizes are small (Lipsey and Wilson, 2001). Next, each effect size was weighted using a random effects model and the average weighted effect size reported was a g value of 0.68. Warfa also examined the data for evidence of publication bias and the presence of moderator effects. For publication bias, the funnel plot method suggested a bias in favor of small sample sizes. Follow-up trim-fill investigations suggested the bias would not appreciably change the reported results. For moderators, Warfa explored the role of class size and geographic location on the effectiveness. The results suggest that CL was most effective in non-US-based locations though this could be attributed to the grade level where non-US-based studies were disproportionately at the high school level. The results also suggested that CL was most effective with small class sizes (less than 50 students) though the author cautions that too few classes were present with medium and large class sizes to make a definitive conclusion.

Warfa's analysis shows that overall CL is more effective than traditional instruction and the moderators chosen indicate geographic location and grade level where CL is used are related to the observed effectiveness. Warfa's analysis also revealed that potentially other modifiers could explain the heterogeneity observed among the articles analyzed. This study seeks to explore the role of additional potential modifiers among Warfa's data set with the intent of further defining where CL is effective. In particular, modifiers are identified related to instructor decisions regarding the implementation and assessment of CL. Examples of instructor decisions are: the type of assessments used, the extent cooperative learning is incorporated into the class and the maximum group size permitted. The focus on instructor decisions is purposeful in that it can inform instructional practices regarding either how to enact cooperative learning or under which classroom conditions cooperative learning would have an expected benefit. For a hypothetical example, if CL is found to not be effective with closed-ended exams then instructors who rely entirely on closed-ended exams may elect to either not employ CL or change the nature of their exams.

Meta-analysis has also been described as a means for establishing benchmarks for effect size that is more directed toward a particular area of study. As Lipsey *et al.* (2012) argue, a meta-analysis can provide a better description then Cohen's benchmarks of small, medium and large effect sizes, as meta-analysis presents empirical evidence of the effect sizes found in the relevant research literature. Thus a meta-analysis on the use of cooperative learning to impact students' chemistry achievement can provide a description for expected effect sizes for pedagogical interventions in chemistry. Finally, meta-analyses provide an overview of a considerable body of research literature for instructors and researchers; such overviews are becoming more necessary given the recent increase in research article production in chemistry education (Ye *et al.*, 2015).

Research questions

This research is directed by the following research questions:

(1) Under which conditions, as determined by instructor discretion, has CL been found to be effective for chemistry instruction?

(2) What is the range of effect sizes observed for CL in chemistry that can establish small, medium and large effect sizes for pedagogical interventions in chemistry?

Methods

Article categorization

This research study began with the data set of 25 samples from 24 research articles established by Warfa. The criteria for inclusion in the data set are described above and are further described in the original article (Warfa, 2016). Within the 24 research articles: 16 describe a college level setting and eight describe high school level; 14 were conducted in the U.S. and 10 were conducted outside the U.S. and 14 articles had class sizes smaller than fifty students. Each article was reviewed based on four constructs and for each construct placed into one of the dichotomous categories as described below.

Assessment type. Assessments that feature multiple-choice exams were termed closed-ended assessments. Assessments that feature alternative assessment techniques including short answer questions, essay questions or free response, were termed non-closed. Originally, articles were to be classified as employing either closed-ended, open-ended or a mix of both assessment techniques, however very few articles described using entirely open-ended assessments. As a result, the categories were established as an assessment using only multiple-choice (labeled closed) or employing at least one alternative type question (non-closed).

Assessment coverage. Assessments were categorized as either cumulative, measuring the content for an entire semester or academic year, or single topic measuring student performance across a defined set of topics. For example, articles described in-term tests given throughout the semester (single) or a final exam that covered all topics in the course (cumulative).

CL usage. In some articles the use of CL was in place for a portion of the class time (periodic), such as two meetings a week for traditional instruction and one meeting a week engaged in CL. Alternatively, other articles described an implementation of CL in every class period (consistent). Of the sixteen studies

Table 1 Article coding

Authors (Year)	Assessment type	Assessment coverage	CL usage	Group size	Hedges's *g*
Acar and Tarhan (2008)	Non-closed	Single	Consistent		2.70
Adesoji and Ibraheem (2009)	Non-closed	Single	Consistent	≥5	0.125
Anderson *et al.* (2005)	Non-closed	Single	Periodic	≥5	1.696
Barthlow (2011)	Closed	Single	Consistent	≤4	0.713
Bilgin (2006)	Closed	Single	Consistent	≤4	1.28
Bilgin and Geban (2006)	Closed	Single	Consistent	≤4	2.035
Bradley *et al.* (2002)	Non-closed	Cumulative	Consistent		−0.0207
Cetin *et al.* (2009)	Closed	Single	Consistent		2.08
Chase *et al.* [General Chemistry] (2013)	*Non-closed*[a]	*Split*[c]	Periodic	≤4	*0.0608*
Chase *et al.* [Organic Chemistry] (2013)	*Non-closed*[b]	*Split*[c]	Periodic	≤4	*−0.0597*
Demircioglu *et al.* (2005)	Closed	Single	Consistent	≥5	0.958
Doymus (2007)	Closed	Single	Consistent	≥5	0.802
Doymus (2008)	Non-closed	Single	Consistent	≤4	1.00
Eaton (2006)	*Closed*	*Cumulative*	Consistent	≥5	−0.169
Goeden *et al.* (2015)	Closed	Cumulative	Periodic	≤4	−0.497
Hein (2012)	Closed	Cumulative	Consistent	≤4	0.176
Hemraj-Benny and Beckford (2014)	Split[c]	Single	Periodic	≤4	0.656
Kirik and Boz (2012)	Closed	Single	Periodic	≥5	1.44
Lyon and Lagowski (2008)			Periodic	≥5	0.414
Mohamed (2008)			Periodic	≤4	0.464
O'Dwyer and Childs (2015)	Non-closed	Cumulative	Consistent		0.115
Oliver-Hoyo *et al.* (2004)	Non-closed		Consistent	≤4	0.133
Shachar and Fischer (2004)		Single	Consistent	≤4	0.504
Shatila (2007)		Cumulative	Periodic	≤4	−0.165
Tarhan and Sesen (2012)	Non-closed	Single	Consistent	≥5	1.55
Williamson and Rowe (2002)	*Non-closed*	*Split*[c]	Consistent	≤4	0.137

Italics indicates this information was provided *via* direct correspondence with the author. [a] Midterm 1–4 were classified as above; the final exam was classified as missing. [b] Midterm 1–3 were classified as above; the final exam was classified as missing. [c] Split articles contained both categories within the construct.

employing CL consistently, eight employed it for an entire semester with the remaining ranging from three to eight weeks.

Group size. Most articles reported a range of student group sizes used in the CL implementation. Articles were classified based on the maximum group size reported. The decision to focus on maximum group size is owing to the common instructional decision to place a cap on group size, while groups smaller then the cap can arise out of logistics (*e.g.* forming groups of four with an odd number of students). Of the 23 articles that reported group size, the most common maximum group sizes reported were four students (13 studies) or five students (six studies). Articles were thus categorized as having maximum group sizes of four or less *versus* five or more students per group.

Thus each assessment was categorized based on the above assessment constructs and each sample was categorized based on the CL usage construct and group size construct. In cases where insufficient information was available to categorize on one of the above constructs, an effort was made to contact the corresponding author of the article for additional information. Ultimately, if the information could not be obtained, the sample in the study was counted as missing toward that construct and not included in either of the two categories associated with the construct. The categorization of each article is presented in Table 1.

Statistical treatment

Each assessment was scored on Cohen's *d* as a measure of effect size and converted to Hedges' *g*. When studies reported pre-test differences between pedagogies and the pre-test measure was identical to the outcome measure used, the effect size of the outcome metric was calculated as the difference between the effect size for the outcome minus the effect size for the pre-test. Thus if an outcome metric reported an effect size of $g = 0.8$ (positive indicates CL was more effective) and the pre-test effect size was $g = 0.3$, the effect size was recorded as 0.5. This adjustment was done to control for observed incoming differences when reported and the following analyses used these adjusted values.

The observed effect sizes for each assessment were averaged to create one effect size for each independent sample (Lipsey and Wilson, 2001). If an article had four closed-ended exams they were averaged and presented as one data point in the closed assessment category. On occasion, a study employed different types of assessment within the study, herein referred to as split articles. For example, Hemraj-Benny and Beckford (2014) used a short-answer (non-closed) assessment for Exam 1 and a multiple-choice (closed) assessment for Exam 2. Split articles were not included in either category that they split. In the Hemraj-Benny example, the split article was not considered in the Assessment Type construct. This decision is revisited in an additional analysis presented in the appendix, where split articles are added to one relevant category and the analysis conducted, then removed and added to the alternative relevant category. Split articles only impacted Assessment Type and Assessment Coverage constructs.

The meta-analysis used a random effects model, similar to Warfa's original analysis, owing to the heterogeneity present among the articles. This decision assumes that there is variability between studies that is randomly distributed. In the random effects

model, articles are weighted based on sample size and effect size observed in each article (fixed effect component) and a term that considers the variability between the studies in the article (random effect component) (Lipsey and Wilson, 2001, pp. 116–120). The assignment of article weights was performed once on the entire corpus of articles (to model variance of the population) and these weights were used throughout the remaining analyses.

For analysis of the data, first construct overlap was considered by conducting a chi-square between each possible construct. Evidence of construct overlap may mean that differences observed for one category are the result of another category, in essence a covariate relationship. Second, for each category, the weighted average effect size was determined and a 95% confidence interval was created. The confidence interval allows a determination of whether CL use in a specific category has a statistically significant impact that is greater than zero. Finally, the effect sizes for the dichotomous categories within each construct were compared using the Q_b statistic as described by Lipsey and Wilson (Lipsey and Wilson, 2001, p. 121). The Q_b statistic is analogous to conducting an independent sample *t*-test between two categories to determine whether one category has an effect size that is significantly greater than the effect size in the other category. The decision regarding split articles described earlier ensures that the samples described in each category are exclusive with no over-lap between them.

Results and discussion

A chi-square test was conducted for each of the six possible pairs of constructs. For example, Assessment Type paired with Assessment Coverage followed by Assessment Type paired with CL Usage. The effect size for the chi-square (Cohen's w) associated with each pairing was determined (Cohen, 1988). The test showed no pairings of construct had overlaps that were medium effect size ($w = 0.3$). The pairing with the most overlap ($w = 0.22$) was between Assessment Coverage and Group Size where six of the seven studies with group sizes of five students or more also used single topic assessments. The relationship was not reciprocal as only six of the twelve single-topic studies used groups of five or greater. The relationship was also not present with groups of four or less *versus* Assessment Coverage. In short, each category can be considered independent except when considering group sizes of five or greater it is worth noting that only one study with this group size used a cumulative assessment.

The number of studies for each category and the weighted mean effect size with a 95% confidence interval are presented in Table 2. The data presented in Table 2 indicates that CL is a robust intervention that has produced a positive, statistically significant outcome in each categorization except for cumulative assessments. The testing of each category relative to an effect size of zero is significant at $p < 0.05$ when the lower confidence limit is positive. The negligible effect size associated with cumulative assessments is particularly noteworthy. This effect size may represent a limitation of CL in terms of content retention as measured by a cumulative exam, a finding that echoes a call for further research on longitudinal impact of pedagogical reforms (National Research Council, 2012; Lewis, 2014). The Q_b statistic for single *versus* cumulative indicates that the higher student performance on single topic exams *versus* cumulative exams is statistically significant as well. The description of only six studies that used cumulative exams is partially attributed to the decision to omit split articles. In the follow-up analysis (see Appendix), three split articles additionally contribute to the cumulative category with consistent results. Of the nine articles that used cumulative assessments, the effect sizes ranged from −0.497 to 0.330 with four of the nine articles having negative effect sizes and two others having effect sizes less than 0.050.

CL was effective in the remaining categorizations ranging from an average effect size of 0.433 to 0.834. CL reported a higher average effect size when implemented consistently, with maximum group size of five or more and with closed ended assessments. Additionally, the Q_b statistic did not identify any of these categorizations as significantly more effective then their counterpart within the construct. Instead, the results then speak to the robustness of CL as a pedagogical tool in a variety of scenarios.

Table 2 Weighted mean effect sizes of each category

Construct	Category (*N*)	Weighted mean effect size	95% confidence limit	Q_b statistic
Assessment type	Closed (9)	0.783	[0.387, 1.178]	0.96
	Non-closed (11)	0.515	[0.153, 0.876]	
Assessment coverage	Cumulative (6)	−0.088	[−0.479, 0.392]	16.3[a]
	Single (13)	1.12	[0.78, 1.45]	
CL usage	Periodic (9)	0.433	[0.037, 0.830]	1.05
	Consistent (16)	0.678	[0.378, 0.978]	
Group size	Four or less (14)	0.443	[0.122, 0.764]	1.85
	Five or more (8)	0.813	[0.388, 1.237]	
	Every study (25)	0.586	[0.339, 0.834]	

[a] Q_b statistic significant at $p < 0.01$.

Literature guidelines have suggested group sizes of four students or fewer for CL, indicating that when group sizes exceed four students, individuals tend to communicate less frequently (Cooper, 1995; Johnson and Johnson, 2009). However, when studies reported a maximum group size of five or more the impact was larger, though statistically comparable, to those with a maximum group size of four or less. The impact of group size is likely influenced by the demands of the setting such as the instructor to student ratio and the physical placement of students to promote interactions. Lecture-halls, where seats are fixed in a row pointing in the same direction, may struggle with group sizes of five or more because students cannot easily interact. Alternatively, a round table with five seats may represent an effective group work set-up. Variables related to the setting would need to be examined further in the research literature to aid making a definitive claim related to the impact of group size, but the articles analyzed here suggest that groups sizes of five or larger can be effective.

The second research question sought to identify the range of effect sizes observed for CL in chemistry to set a benchmark for other educational interventions. Toward that end, the meta-analysis conducted here found a 95% confidence interval of the entire corpus of studies to range from 0.34 to 0.83 with a midpoint of 0.59. This range can be thought of as defining small (0.34), medium (0.59) and large (0.83) effects for attempts to improve chemistry learning through pedagogical intervention. These benchmarks can be thought of as fluid and are expected to evolve as future reviews of research in chemistry education are conducted. These results are in line with Warfa's previous meta-analysis that found an average effect size in chemistry of 0.68 or a recent meta-analysis on college-level STEM performance that found an average effect size of 0.47 (Freeman *et al.*, 2014; Warfa, 2016). Put in context, future work that employs a pedagogical intervention to impact chemistry achievement with an effect size less than 0.6 may be viewed as less effective than CL. A special exemption to this would be if the effect size were observed on cumulative assessments where such effect sizes were not observed with CL.

Limitations

The relatively small number of studies incorporated in this meta-analysis limits this study. In particular, the statistical tests of differences within each construct would benefit in terms of statistical power with additional studies included. Additionally, the covariate relationship between constructs could be explored in more detail with a larger sample. In particular, caution is warranted regarding efforts to optimize CL implementation *via* combining results across multiple constructs, as interactions across constructs could not be explored with the current sample size. While the number of articles may limit the ability to compare categories across constructs, within each construct seven of the eight categories had sufficient statistical power to identify weighted average effect sizes significantly different than zero. The remaining category, cumulative exams, shows no descriptive indication of a positive effect.

One possibility for expanding the number of studies would be to add additional keywords to the search term including, for example "group work" or the use of widely disseminated reform efforts which incorporate CL such as "Process-Oriented Guided Inquiry Learning" or "Peer-Led Team Learning" (Gosser and Roth, 1998; Moog and Spencer, 2008). Future work is planned to expand the meta-analysis to incorporate these variants in CL methodology. Such work can then consider the effectiveness of these methodologies by using them as moderators within a meta-analysis.

Conclusions

The descriptive statistics and follow-up statistical tests indicate that CL pedagogy results in learning gains on outcome measures of chemistry achievement in a variety of assessment techniques, consistency of use and group size. The lack of impact of CL on cumulative exams does give pause and calls for future study, including exploratory qualitative research, into the factors relevant for the retention of chemistry skills and content. The meta-analysis does lend an evidence-base for the instructor decisions to incorporate CL in the classroom periodically or use group sizes larger than the recommended maximum of four students. The results presented also provide a general benchmark for small, medium and large effects for pedagogical reform aiming to improve chemistry achievement, which can guide future evaluation efforts. Finally, the focus on instructor decisions as moderators in meta-analyses can serve as a future direction to provide evidence-based recommendations for instructional practice.

Appendix: determining the impact of split articles on sensitivity of the results

In categorizing articles in the meta-analysis, a small number of the articles were split across categories. In particular, one article included both closed and non-closed assessments and three other articles included both single topic and cumulative assessments. In the original analysis, these articles were not considered in comparing these groupings to avoid violating independence of observations. What follows is an additional analysis that investigates the impact of this decision on the findings presented by analyzing their contribution to each category.

Closed *versus* non-closed assessments

One article used both closed and non-closed assessments (Hemraj-Benny and Beckford, 2014). The closed assessment in this article was added to the closed category and compared to the original non-closed category in Table 3.

The inclusion of this article in the closed category resulted in a minimal change from a weighted mean effect size of 0.783 (Table 2) to 0.772 (Table 3). The resulting confidence interval and Q_b statistic lead to similar interpretations that CL had a positive, significant effect with closed assessments that were not statistically different than non-closed assessments.

Next, the split article's non-closed assessment was added to the non-closed category and compared to the original closed assessment in Table 4.

Table 3 Including split article with closed assessments

Variable (N)	Weighted mean effect size	Standard error	95% confidence limit	Q_b statistic
Closed with splits (10)	0.772	0.191	[0.397, 1.147]	0.94
Non-closed (11)	*0.515*	*0.184*	*[0.153, 0.876]*	

Table 4 Including split article with open assessments

Variable (N)	Weighted mean effect size	Standard error	95% confidence limit	Q_b statistic
Closed (9)	*0.783*	*0.202*	*[0.387, 1.178]*	0.92
Non-closed with splits (12)	0.526	0.176	[0.180, 0.871]	

Table 5 Including split articles with cumulative assessments

Variable (N)	Weighted mean effect size	Standard error	95% confidence limit	Q_b statistic
Cumulative with Splits (9)	−0.020	0.201	[−0.414, 0.372]	18.6[a]
Single (13)	*1.12*	*0.17*	*[0.78, 1.45]*	

[a] The Q_b statistic is significant at $p < 0.01$.

Table 6 Including split article with single assessments

Variable (N)	Weighted mean effect Size	Standard error	95% confidence limit	Q_b statistic
Cumulative (6)	*−0.088*	*0.245*	*[−0.479, 0.392]*	11.7[a]
Single with splits (16)	0.900	0.154	[1.202, 0.598]	

[a] The Q_b statistic is significant at $p < 0.01$.

The impact of the non-closed assessment was also minimal changing the weighted mean effect size from 0.515 (Table 2) to 0.526 (Table 4). The category non-closed assessments remained significantly greater than zero and not statistically different than closed assessments.

Cumulative *versus* single assessments

Two articles, providing three independent samples, included both single and cumulative assessments (Williamson and Rowe, 2002; Chase *et al.*, 2013). Similar to above, these three articles were added to one category at a time. First, the cumulative assessments were added to the cumulative category in Table 5.

The weighted mean effect size in the cumulative assessments had a minor change from −0.088 (Table 2) to −0.020 (Table 5). This value can still be interpreted as a negligible effect on student achievement with a confidence interval that still ranges across zero. Additionally, the effect of CL on cumulative assessments was still found to be significantly below the effect of CL on single-topic assessments.

Finally, the single topic assessments with the three split samples were added to the single topic assessment category in Table 6.

The inclusion of three articles into the single topic assessments had the largest change observed, moving from 1.12 (Table 2) to 0.90 (Table 6). The resulting value remains significantly greater than zero and significantly greater than cumulative exams.

Overall, the above analysis can be viewed as a sensitivity test of the original analysis to the decision regarding split articles. The results indicate that the most noteworthy finding, the lack of impact for CL on cumulative exams, is not impacted by the original decision to omit split articles. In addition, none of the other categories would change the interpretation of the effectiveness of CL presented in the original analysis.

Acknowledgements

Partial support for this work was provided by the National Science Foundation's Robert Noyce Teacher Scholarship Program under DUE-1439776. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The authors would like to acknowledge the prior work of Abdi-Rizak M. Warfa in identifying the articles for this meta-analysis. Additionally the authors acknowledge Abdi-Rizak M. Warfa and Jeffrey Kromrey for their helpful consultations regarding meta-analysis. Finally, the authors acknowledge the researchers of the reviewed studies for their prior published efforts and responses to correspondence requesting additional information where needed.

References

Acar B. and Tarhan L., (2008), Effects of Cooperative Learning on Students' Understanding of Metallic Bonding, *Res. Sci. Educ.*, **38**, 401–420.

Adesoji F. A. and Ibraheem T. L., (2009), Effects of Student Teams-Achievement Divisions Strategy and Mathematics

Knowledge on Learning Outcomes in Chemical Kinetics, *J. Int. Soc. Res.*, **2**, 15–25.

Anderson W. L., Mitchell S. M. and Osgood M. P., (2005), Comparison of Student Performance in Cooperative Learning and Traditional Lecture-based Biochemistry Classes, *Biochem. Mol. Biol. Educ.*, **33**, 387–393.

Barthlow M. J., (2011), *The Effectiveness of Process Oriented Guided Inquiry Learning to Reduce Alternate Conceptions in Secondary Chemistry*, PhD thesis, Liberty University.

Bilgin I., (2006), Promoting Pre-Service Elementary Students' Understanding of Chemical Equilibrium through Discussions in Small Groups, *Int. J. Sci. Math. Educ.*, **4**, 467–484.

Bilgin I. and Geban O., (2006), The Effect of Cooperative Learning Approach Based on Conceptual Change Conditions on Students' Understanding of Chemical Equilibrium Concepts, *J. Sci. Educ. Technol.*, **15**, 31–46.

Bowen C. W., (2000), A Quantitative Literature Review of Cooperative Learning Effects on High School and College Chemistry Achievement, *J. Chem. Educ.*, **77**, 116–119.

Bradley A. Z., Ulrich S. M., Jones Jr. M. and Jones S. M., (2002), Teaching the Sophomore Organic Course without a Lecture. Are You Crazy? *J. Chem. Educ.*, **79**, 514–519.

Cetin P. S., Kaya E. and Geban O., (2009), Facilitating Conceptual Change in Gases Concepts, *J. Sci. Educ. Technol.*, **18**, 130–137.

Chase A., Pakhira D. and Stains M., (2013), Implementing Process-Oriented, Guided-Inquiry Learning for the First Time: Adaptations and Short-Term Impacts on Students' Attitude and Performance, *J. Chem. Educ.*, **90**, 409–416.

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale: Lawrence Erlbaum Associates.

Cooper M. M., (1995), Cooperative Learning: An Approach for Large Enrollment Courses, *J. Chem. Educ.*, **72**, 162–164.

Demircioglu G., Ayas A. and Demircioglu H., (2005), Conceptual change achieved through a new teaching program on acids and bases, *Chem. Educ. Res. Pract.*, **6**, 36–51.

Doymus K., (2007), Effects of a Cooperative Learning Strategy on Teaching and Learning Phases of Matter and One-Component Phase Diagrams, *J. Chem. Educ.*, **84**, 1857–1860.

Doymus K., (2008), Teaching chemical bonding through jigsaw cooperative learning, *Res. Sci. Technol. Educ.*, **26**, 47–57.

Eaton L., (2006), *The Effects of Process Oriented Guided Inquiry Learning on Student Achievement in a One Semester General, Organic, and Biochemistry Course*, Masters thesis, St. John Fisher College.

Freeman S., Eddy S. L., McDonough M., Smith M. K., Okoroafor N., Jordt H. and Wenderoth M. P., (2014), Active learning increases student performance in science, engineering and mathematics, *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 8410–8415.

Goeden T. J., Kurtz M. J., Quitadamo I. J. and Thomas C., (2015), Community-Based Inquiry in Allied Health Biochemistry Promotes Equity by Improving Critical Thinking for Women and Showing Promise for Increasing Content Gains for Ethnic Minority Students, *J. Chem. Educ.*, **92**, 788–796.

Gosser D. and Roth V., (1998), The Workshop Chemistry Project: Peer-Led Team Learning, *J. Chem. Educ.*, **75**, 185–187.

Hein S. M., (2012), Positive Impacts Using POGIL in Organic Chemistry, *J. Chem. Educ.*, **89**, 860–864.

Hemraj-Benny T. and Beckford I., (2014), Cooperative and Inquiry-Based Learning Utilizing Art-Related Topics: Teaching Chemistry to Community College Nonscience Majors, *J. Chem. Educ.*, **91**, 1618–1622.

Johnson D. W. and Johnson R. T., (2009), An Educational Psychology Success Story: Social Interdependence Theory and Cooperative Learning, *Educ. Res.*, **38**, 365–379.

Johnson D. W., Johnson R. T. and Smith K. A., (1998), Cooperative Learning Returns to College: What Evidence Is There That It Works? *Change: The Magazine of Higher Learning*, **30**, 27–35.

Kirik O. T. and Boz Y., (2012), Cooperative learning instruction for conceptual change in the concepts of chemical kinetics, *Chem. Educ. Res. Pract.*, **13**, 221–236.

Lewis S. E., (2014), Investigating the Longitudinal Impact of a Successful Reform in General Chemistry on Student Enrollment and Academic Performance, *J. Chem. Educ.*, **91**, 2037–2044.

Lipsey M. W. and Wilson D. B., (2001), *Practical Meta-analysis*, Thousand Oaks: Sage Publications.

Lipsey M. W., Puzio K., Yun C., Hebert M. A., Steinka-Frey K., Cole M. W., Roberts M., Anthony K. S. and Busick M. D., (2012), *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*, Report NCSER 2013-3000, National Center for Special Education Research, Institute of Education Sciences.

Lyon D. C. and Lagowski J. J., (2008), Effectiveness of Facilitating Small-Group Learning in Large Lecture Classes, *J. Chem. Educ.*, **85**, 1571–1576.

Mohamed A.-R., (2008), Effects of Active Learning Variants on Student Performance and Learning Perceptions, *J. Scholarship Teach. Learn.*, **2**, 1–15.

Moog R. S. and Spencer J. N., (2008), *Process-Oriented Guided Inquiry Learning*, Washington, DC: American Chemical Society.

National Research Council, (2012), *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, Washington, DC: Board on Science Education, Division of Behavioral and Social Science and Education, Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research.

O'Dwyer A. and Childs P., (2015), Organic Chemistry in Action! What is the Reaction? *J. Chem. Educ.*, **92**, 1159–1170.

Oliver-Hoyo M. T., Allen D., Hunt W. F., Hutson J. and Pitts A., (2004), Effects of an Active Learning Environment: Teaching Innovations at a Research I Institutions, *J. Chem. Educ.*, **81**, 441–448.

Shachar H. and Fischer S., (2004), Cooperative learning and the achievement of motivation and perceptions of students in 11th grade chemistry classes, *Learn. Inst.*, **14**, 69–87.

Shatila A., (2007), *Assessing the Impact of Integrating POGIL in Elementary Organic Chemistry*, Doctoral thesis, University of Southern Mississippi.

Tarhan L. and Sesen B. A., (2012), Jigsaw cooperative learning: acid–base theories, *Chem. Educ. Res. Pract.*, **13**, 307–313.

Warfa A.-R. M., (2016), Using Cooperative Learning to Teach Chemistry: A Meta-analytic Review, *J. Chem. Educ.*, **93**, 248–255.

Williamson V. M. and Rowe M. W., (2002), Group Problem-Solving *versus* Lecture in College-Level Quantitative Analysis: The Good, the Bad, and the Ugly, *J. Chem. Educ.*, **79**, 1131–1134.

Ye L., Lewis S. E., Raker J. R. and Oueini R., (2015), Examining the Impact of Chemistry Education Research Articles from 2007 through 2013 by Citation Counts, *J. Chem. Educ.*, **92**, 1299–1305.

