Investigating a collaborative group exam as an instructional tool to address student reasoning difficulties that remain even after instruction

Alistair McInerny
Department of Physics, North Dakota State University, 1340 Administration Ave, Fargo, ND 58105

Mila Kryjevskaia
Department of Physics, North Dakota State University, 1340 Administration Ave, Fargo, ND 58105

Many students tend to provide intuitively appealing (but incorrect) responses to some physics questions despite demonstrating (on isomorphic questions) the formal knowledge necessary to reason correctly. These inconsistencies in reasoning are persistent and remain even after evidence-based instruction. This project probed whether a collaborative group exam could serve not only as an innovative assessment tool but also as an instructional intervention that helps address persistent reasoning difficulties. Specifically, students were given opportunities to revisit their answers to questions known to elicit intuitively appealing responses in a collaborative group exam component immediately following a traditional individual exam. The efficacy of this approach was compared to that of a more traditional instructor-led exam review session. Both approaches yielded moderate improvements in performance on the final exam. However, additional multi-faceted data analysis provided further insights into student reasoning difficulties that suggested further implication for instruction and research.

I. INTRODUCTION

Many research-based instructional materials and techniques produce positive impacts on various aspects of student learning, including conceptual understanding and reasoning [1-6]. At the same time, a growing body of research suggests that, even after targeted instruction designed to address persistent student difficulties, many students continue to reason inconsistently [3-6]. In particular, some students are able to demonstrate necessary conceptual understanding on some physics tasks but fail to do so on isomorphic tasks that require the application of the same knowledge and reasoning but also tend to elicit strong intuitively appealing ideas [7-10].

An overarching goal of our project is to identify factors and instructional circumstances that appear to enhance productive student reasoning in physics. Existing classroom interventions developed by physics education researchers appear to improve the level of consistency in student reasoning with various degrees of success. In this project we probe the efficacy of another (perhaps less conventional) form of instructional intervention, the collaborative group exam, as a strategy to help students identify and resolve inconsistencies in their reasoning that remain even after instruction. In this paper, we discuss the motivation for this work, specific aspects of implementation, and implications for instruction.

II. MOTIVATION AND THEORETICAL FRAMEWORK

The dual process theory of reasoning developed in cognitive psychology has been used to account for many observed inconsistencies in student reasoning in physics [11-13]. The theory suggests that two processes are involved in most reasoning tasks: process 1 is quick, subconscious, and intuitive while process 2 is slow, logic-based, and deliberate. The most critical aspect of the interactions between the two processes is that process 1 cannot be turned off; we perceive the world around us through the lens of the quick and automatic process 1. Process 2 may only intervene after process 1 has formed an intuition-based mental model of a given situation. Productive intervention by process 2 requires correct and relevant background knowledge (e.g., understanding of relevant physics concepts). However, because process 2 is often impaired by reasoning biases of its own, the presence of relevant background knowledge alone may not be sufficient. For example, individuals tend to look for evidence that supports what they already believe to be true (i.e., confirmation bias). To catch a mistake, process 2 must be placed on alert by detecting reasoning "red flags." We argue that becoming aware of one's own reasoning and developing the ability to recognize (and act upon) reasoning red flags represents a critical step for developing an expertise in physics. Moreover, the benefits of developing such cognitive reflection skills extend to other areas of human functioning beyond a physics classroom and therefore should be fostered in college instruction.

It has been shown that collaborative group work is effective in engaging students in socially mediated metacognition in which group members share their individual thinking, evaluate ideas, receive feedback, and monitor each other's reasoning [14-18]. As such, it is likely that a collaborative work environment may be effective in helping students both identify reasoning red flags and mediate intuitive ideas via analytical reasoning. In recent years, a different form of collaborative group work, the collaborative exam, has gained momentum in the science education community [19-36]. An emerging body of research suggests that collaborative exams have marked benefits that include improved performance [25-32], increased motivation [33], and decreased test anxiety [26]. While collaborative exams appear to be a promising and innovative educational tool, many aspects of its efficacy are still under investigation [33, 34]. For instance, instructors often raise concerns that collaborative exams simply promote the propagation of correct answers, do not facilitate individual growth, and are challenging to implement. At the same time, emerging evidence suggests that collaborative exams do often promote individual learning if students are given adequate time to meaningfully examine their responses (even if none of the students were able to arrive at a correct answer on their own) [24]. In addition, the impact of group exams is enhanced even further in classrooms where a collaborative group work is a norm, thus capitalizing on the alignment between assessment and instruction [34].

As stated above, collaborative group work fosters socially mediated metacognition which may help students learn to recognize reasoning red flags and develop strategies for resolving inconsistencies. The high-stakes environment of an exam setting may boost this effect further by enhancing student motivation to arrive at a correct response with correct reasoning. Because students were graded based on the quality of their reasoning, they may have been particularly motivated to examine their thinking as opposed to just accepting an answer as correct. The latter may happen more frequently during regular classroom instruction.

III. METHODS

This study was conducted in two semesters of an introductory calculus-based mechanics course serving primarily non-physics majors at a mid-size, research-focused land grant university. In both semesters, different instructors implemented active learning techniques such as peer instruction, tutorials, and collaborative group work.

In semester 1, we employed the two-stage exam design, featuring an individual component and a group component, during a 2-hour class period. First, students completed the test individually and submited their written responses (~60 min); then, following a short break, students were given an opportunity to work in collaborative groups (~40 min). In our study, the collaborative portion included a subset of

questions from the individual component, including those questions known to reveal persistent incorrect intuitive responses. Questions from the individual component that are irrelevant to this study were also included. During the collaborative group component, students were allowed to choose their own group partners, but they overwhelmingly stayed within the groups formed during regular classroom instruction. Although students were encouraged to discuss responses to the collaborative component in their groups, they were required to submit their own answers with detailed explanations of their reasoning. In semester 1, 68 students completed individual and group components on 3 midterm exams. Student performance on the group components contributed 20% to their midterm grades. Students did not receive any formal feedback from an instructor during or after the exams. Semester 1 is considered to provide "treatment" conditions.

In semester 2, 48 students completed identical exams individually (no group component was included); however, the instructor conducted a follow-up classroom session dedicated to reviewing solutions to the exam and answering student questions. In both semesters, exam solutions were not posted. Students did not have access to the exam problems to review or "study from" before the final exam. We consider semester 2 to be the "control" condition. We believe that semester 2 represents traditional exam design and therefore serves as a baseline to compare the collaborative exam treatment against. In this study, we intended to probe the efficacy of the collaborative exam as a learning tool rather than an assessment tool, with a specific focus on probing the impact of this intervention on student performance on questions that tend to elicit intuitively appealing (but incorrect) ideas that persist even after classroom instruction.

To probe the level of consistency in student reasoning, we used the screening-target methodology, which employs a pair of isomorphic questions. A screening question probes whether a student possesses the knowledge and skills necessary to analyze a given situation correctly. A target question requires the application of the same formal knowledge and skills but includes surface features that tend to elicit incorrect intuitively appealing ideas.

An example of a screening-target question pair is shown in Fig. 1. On the screening question, most students correctly recognize that, because box A remains at rest, the net force on the box must be zero; therefore, the force of static friction must be equal in magnitude to the applied 30 N force. The target question requires the application of the same reasoning because both boxes remain at rest while identical 30N horizontal forces act on each box. However, ~25% of the students who answer the screening question correctly do not use the correct reasoning approach on the target question. Instead, they tend to argue that the magnitude of the friction force on box A is less than that on box B because the coefficient of static friction between the surface and box A is smaller. Through the lens of dual process theory, we argue

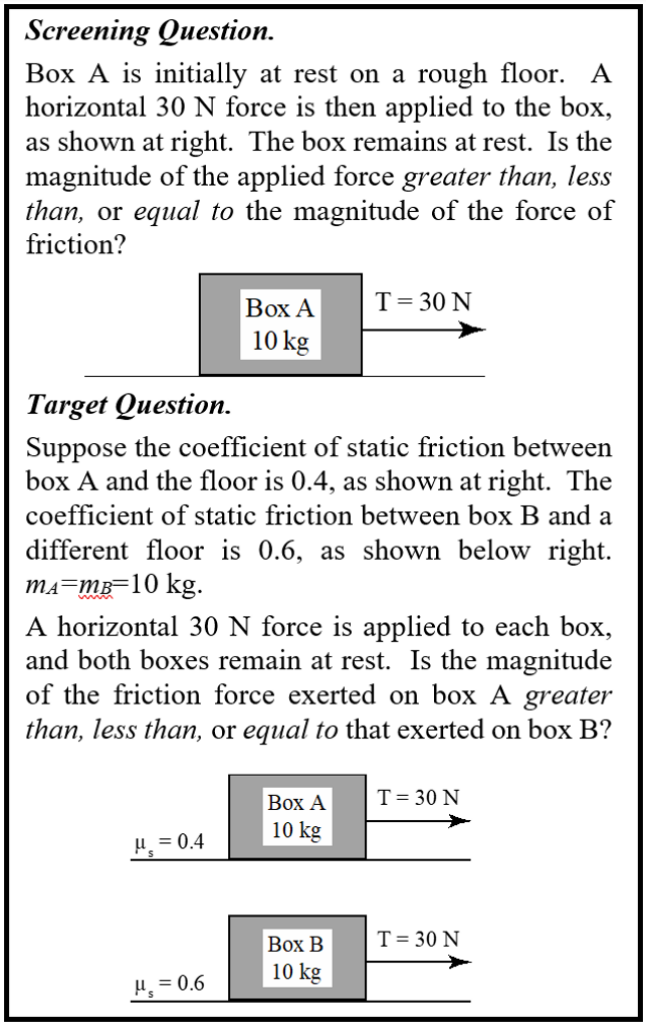

FIG. 1. An example of a screening-target question pair

that the inclusion of the extraneous information on the target question cues an automatic but incorrect response that "higher μ implies higher friction." It appears that students who make this type of error immediately and subconsciously embrace this response as correct, while the mathematical relationship (irrelevant in this case) between kinetic friction f_k and μ_k, $f_k = \mu_k N$, provides further confirmation of the intuitive μ-based response.

The screening-target pair in Fig. 1 served as one of the 5 pairs of questions included across the individual components of the 3 midterm exams. The other 4 pairs were designed in the context of kinematics graphs [6], Newton's 3rd law [35], dynamics of circular motion, and work and energy [36]. In both semesters, 1 question pair was included on exam 1 and 2 pairs were included on exams 2 and 3. In semester 1, each target question was also included in the collaborative group exam component. As stated above, students were required to provide in-depth reasoning since their work was graded based on the quality of explanations rather than the correctness of answers.

To assess the effect of both conditions on student performance on questions that tend to elicit incorrect intuitive (rather than correct formal) reasoning, all five target questions were also included on the final exam in both semesters. We recognized that the most significant limitation of this approach was the possibility of students giving memorized responses. However, the decision to include the same set of five target questions on the final exam was made after careful considerations. First, since student intuitively appealing ideas are often cued by specific features of a task, designing new versions of target questions may introduce new variables without necessarily addressing the possibility of a memorized response. As such, we opted for a more parsimonious design. Second, given the persistent nature of student difficulties, we were interested in probing the impacts of the two interventions under the most favorable conditions. Moreover, results of data analysis discussed in Section IV suggest that the memorization of correct responses is not a major factor affecting student performance on the final exam. The time between testing on a midterm exam and re-test on the final exam varied from several months (for midterm 1) to several weeks (for midterm 3).

IV. RESULTS & DISCUSSION

Student individual responses were coded in a binary format with a score of 1 or 0 given to correct or incorrect responses, respectively. Then, each pair of student screening-target responses received one of four possible codes [i,j], with i and j representing performance on the screening and target questions, respectively.

Three approaches to data analysis and interpretation were employed: (1) a course-level analysis involving comparison of all aggregated [i,j] codes pre- and post-treatment in the two conditions, (2) a student-level analysis using a matched pre- and post-treatment data for each student, and (3) a question-level analysis conducted to probe shifts in student performance on each question.

Analysis of performance pre-treatment. A course-level analysis of student performance on the individual components of the three midterms revealed nearly identical results in the two semesters (see Table 1) suggesting no difference in the student populations before treatments. Close to half of the total student responses to the five pairs of screening-target questions were correct and consistent (codes [1,1], 42% and 47%): students answered both screening and target questions correctly. Nearly a fifth of all responses were coded as [1,0] revealing inconsistencies in reasoning that suggest that a fraction of the students who are able to apply correct conceptual understanding on screening questions tended not to deploy those correct reasoning approaches on target questions which elicit intuitively appealing responses. More significantly, a third of all correct responses on the screening questions were followed by an inconsistent response on the corresponding target question ([1,0]/([1,1]+[1,0])) suggesting that even in the presence of correct conceptual understanding many students had not yet developed strategies to recognize intuitively appealing, but incorrect, ideas and to override them with correct reasoning acquired during formal physics instruction. Understanding and addressing this type of reasoning errors is an overarching goal of this project.

TABLE 1. Student individual performance on midterms.

Codes	Semester 1 (collaborative group exam condition)	Semester 2 (instructor-led exam review condition)
[1,1]	42%	47%
[1,0]	20%	17%
[0,0]	27%	25%
[0,1]	11%	11%

Approximately a quarter of all responses revealed that students had not developed a basic conceptual understanding as evident by the incorrect responses to both screening and target questions (codes [0,0]).

A small fraction of codes [0,1] demonstrated inconsistencies in student reasoning; however, further research is needed to pinpoint the sources of this error (student carelessness on screening questions, guessing on target questions, etc).

The sankey diagrams in Fig. 2 provide some help with data visualization: the vertical bars on the left side of each diagram represent prevalence of specific codes on the midterms (also included in Table 1). The bars on the right side represent student performance on the final exam.

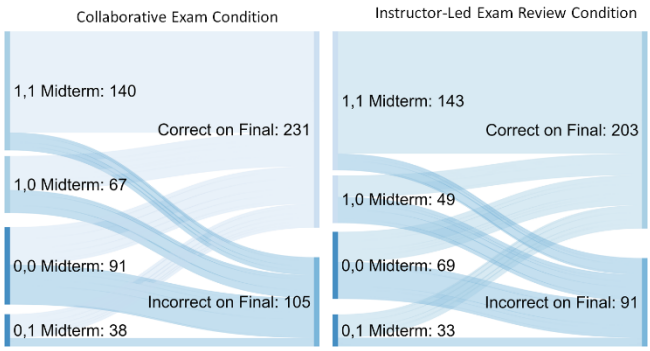

FIG. 2. Sankey diagram showing shifts from midterm responses to final responses (in aggregate) for both treatments.

A student-level analysis confirmed the result above, that no significant difference in student populations was detectable before the treatments. Specifically, for individual midterm responses, the average numbers of correct responses to target questions (per student) were $< N^{Midterm}_{Target,Collab} >= 2.7$ and $< N^{Midterm}_{Target,Review} >= 2.9$ for the collaborative exam and instructor-led exam review conditions respectively, with nearly equal variance. This difference is not statistically significant (t-test, $p>0.05$).

Analysis of student performance post-treatment. In both conditions, student performance on the five target questions

included on the final exam improved significantly. The average number of correct responses per student increased from the numbers reported above to $< N^{Final}_{Target,Collab} > = < N^{Final}_{Target,Review} >= 3.5$. This increase is statistically significant (t-test, $p<0.05$). While these averages are equal, the effect size (Cohen's d) of the collaborative exam condition, $d_{Collab}=0.64$, is larger than that in the exam review condition, $d_{Review}=0.44$, primarily due to a smaller variance in the student performance on the final exam in the collaborative exam condition ($\sigma^2_{Collab} = 0.84$, $\sigma^2_{Review} = 1.47$). Still, we find the moderate effect of the collaborative exam ($d_{Collab}=0.64$) to be reassuring given that the students in this condition did not receive any formal feedback from a source of "authority" (e.g., the instructor or exam solutions). In addition, the smaller variance in performance on the final exam in that condition seems to suggest that the collaborative exam treatment may be more equitable.

The two vertical bars on the right-hand side of each sankey diagram in Fig. 2 illustrate course-level student performance on the final exam. In both conditions, the fractions of correct responses were the same (69%), which is consistent with the student-level analysis of the shifts in the average number of correct responses per student. However, the sankey diagrams reveal an interesting (and remarkably consistent) pattern in the "flow" of student responses from midterms to the final which allows for additional insights into nuanced aspects of the shifts in student performance. Nearly all [1,1] codes assigned to the midterm performance are also linked to correct answers to the target questions on the final exam suggesting that in the presence of conceptual understanding (indicated by the correct performance on the screening questions) student correct responses to the target questions appear to be stable over time.

Approximately half of the incorrect responses to the target questions on midterms ([1,0]+[0,0]) switched to correct answers on the final exam independent of performance on the screening questions or the treatment condition. Multiple interpretations are possible. One may expect that students who demonstrated correct conceptual understanding on a screening question would be more likely to improve their reasoning on the target question and therefore would be more likely to reason correctly on the final exam. The absence of this dependence may suggest that the improved performance on the final exam is a result of memorization. We argue, however, that the tendency to memorize is not a major factor in the observed improvement due to the following. First, as stated above, the exam solutions were not posted for review at any point during the semester. Second, the most significant improvement in performance was observed on the question included on midterm 1 (a few months before the final) with the smallest improvement observed on one of the questions included on midterm 3 (two weeks before the final). Third, and most importantly, a question-level analysis revealed that students appear to improve more on those questions that yielded higher performance on a midterm (at least 60% correct). This finding suggests that some intuitive ideas that remain even after formal instruction could be further addressed during (or after) summative assessment by implementing either a collaborative group exam component as a part of the assessment process, or by following up with an instructor-led exam review. At the same time, other intuitive patterns of reasoning appear to be less responsive to the "quick" interventions examined here and may require more targeted classroom instruction that takes into account both student conceptual difficulties and tendencies to reason intuitively. Our explorative data analysis helped identify contexts in which incorrect reasoning patterns do not appear to be responsive to the treatments described here (i.e., dynamics of circular motion, and work and energy); however, further research is needed to generalize to other instructional conditions.

V. CONCLUSIONS

In this project, we probed the impacts of a collaborative group exam and an instructor-led exam review in addressing student reasoning difficulties that remain even after classroom instruction. Results suggest that both approaches led to comparable improvements in performance on questions that tend to elicit incorrect intuitively appealing responses even in the presence of correct conceptual understanding. Nevertheless, we advocate for the implementation of the collaborative group exam approach in courses in which student group work is established to be a norm. A synergy between the classroom instruction and this assessment technique may provide additional benefits to student learning not examined in this study (e.g., developing social networks of support) as well as the potential for more equitable improvements in student reasoning.

Further, we argue that the two approaches examined in this study may serve not only as instructional techniques but also as research tools for identifying those patterns of incorrect student reasoning that require priority during instruction. Indeed, our results revealed that improvements in student performance appear to be higher on questions that already yielded fairly satisfactory performance on midterms. This suggests that perhaps the corresponding classroom instruction was already effective in addressing conceptual and reasoning difficulties so that some students simply needed a gentle nudge provided by the examined interventions. At the same time, on those questions that yielded less satisfactory performance on midterms, the effects of the interventions were minimized, thus suggesting the need for more rigorous targeted instruction.

ACKNOWLEDGMENTS

This material is based upon work supported by the national science foundation under the grants Nos. DUE-1821390, DUE-1821123, DUE-1821400, DUE-1821511, DUE-1821561, DUE-1431940, DUE-1431541, DUE-1431857, DUE-1432052, and DUE-1432765.

[1] M.K Smith, et al. Why peer discussion improves student performance on in-class concept questions, Science **323**, 5910 (2009).

[2] M. Menekse, et al. Differentiated overt learning activities for effective instruction in engineering classrooms, Journal of Engineering Education, **102**, 3 (2013).

[3] L. C. McDermott and E. F. Redish, Resource Letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).

[4] L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Resource Letter RPS-1: Research in problem solving, Am. J. Phys. **72**, 1147 (2004).

[5] D. E. Meltzer and R. K. Thornton, Resource Letter ALIP–1: Active-Learning Instruction in Physics, Am. J. Phys. **80**, 478 (2012).

[6] A.F. Heckler, The Ubiquitous Patterns of Incorrect Answers to Science Questions: The Role of Automatic, Bottom-up Processes, *Psychology of Learning and Motivation*, **55** (2011)

[7] C.R. Gette, et al. Probing student reasoning approaches through the lens of dual-process theories: a case study in buoyancy *Physical Review Physics Education Research*, **14** 2018.

[8] C.R. Gette and M. Kryjevskaia, Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses, *Physical Review Physics Education Research*, **15** 2019.

[9] M. Kryjevskaia, et al. Answer first: Applying the heuristic-analytic theory of reasoning to examine student intuitive thinking in the context of physics, *Physical Review Special Topics - Physics Education Research*, **10**, 2 2014.

[10] Kryjevskaia, Mila, et al. Failure to engage: Examining the Impact of Metacognitive Interventions on Persistent Intuitive Reasoning Approaches, in *Physics Education Research Conference Proceedings*, 2015.

[11] J.S.B.T Evans, The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, **13**, 3 (2006).

[12] D. Kahneman, *Thinking, Fast and Slow* (Macmillan, New York, 2011).

[13] D. Sands, in *ICPE-EPEC 2013 Proceedings*, edited by L. Dvorák and V. Koudelková ˇ (MATFYZPRESS, Prague, Czech Republic, 2013)

[14] M. Goos, et al. Socially mediated metacognition: Creating collaborative zones of proximal development in small group problem solving, Educational studies in Mathematics, **49**, 2 (2002)

[15] L. S. Vygotsky and M Cole, *Mind in Society: The Development of Higher Psychological Processes* (Harvard U.P., Cambridge, MA, 1978).

[16] H. Shirouzu, et al. Cognitively active externalization for situated reflection, Cognitive Science, **26**, 4 (2002)

[17] M. A. Siegel, Filling in the distance between us: Group metacognition during problem solving in a secondary education course, Journal of Science Education and Technology, **21**, 3, (2011).

[18] M.M. Chiu and S.W. Kuo, From metacognition to social metacognition: Similarities, differences, and learning, Journal of Education Research, **3**, 4 (2010).

[19] R.N. Cortright, et al. Student retention of course content is improved by collaborative-group testing, Advances in Physiology Education **27**, 3 (2003).

[20] J.E. Cooke, et al. Online and Cliker Quizzing on Jargon Terms Enhances Definition-Focused but not Conceptually Focused Biology Exam Performance, CBE—Life Sciences Education, **18**, 2, (2019).

[21] P. Heller, Patricia, et al. Teaching problem solving through cooperative grouping, Part 1: Group versus individual problem solving, American Journal of Physics, **60**, 7, (1992).

[22] P. Heller, and M. Hollabaugh. Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, American Journal of Physics, **60**, 7, 1992.

[23] H.Jang, N. Lasry, K. Miller, and E. Mazur, Collaborative exams: Cheating? Or Learning? American Journal of Physics **85**, 3, (2017).

[24] L.K. Durrant, G. Pierson, E.M. Allen, Group testing and its effectiveness at selected nursing concepts Journal of the Royal Society of Health, **105**, 3 (1985).

[25] M. Lusk, and L. Conklin, Collaborative testing to promote learning, *Journal of Nursing Education*, B, 3, (2003).

[26] C.E Wieman, G.W. Rieger, and C.E. Heiner, Collaborative assessment that supports learning, *The Physics Teacher*, **52**, 1, (2014).

[27] J.G. Lambiotte, et al. Cooperative learning and test taking: Transfer of skills, *Contemporary Educational Psychology*, **12**, 1, (1987).

[28] S.A. Stearns, *College Teaching*, Collaborative exams as learning tools **44**, 3, (1996).

[29] R.F. Yuretich, et al. Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course, Journal of Geoscience Education, **49**. 2, (2001).

[30] B.H. Gilley, and B. Clarkston, Collaborative Testing: Evidence of Learning in a Controlled In-Class study of Undergraduate Students Journal of College Science Teaching, **43**, 3, 2014.

[31] K. Knierim, H. Turner, R.K. Davis, Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades, Journal of Geoscience Educati*o*n, **63**, 2, (2015).

[32] P.G. Zimbardo, L.D. Butler, and V.A Wolfe, Cooperative college examinations: More gain, less pain when students share information and grades, The Journal of Experimental Education, **71**, 2, (2003).

[33] G.W Rieger, and C.E. Heiner, Examinations that support collaborative learning: The students' perspective, Journal of College Science Teaching, **43**, 4, (2014).
[34] S.I. Efu, Exams as learning tools: A comparison of traditional and collaborative assessment in higher education, College teaching, **67**, 1, 2019.
[35] A. Elby, Helping Physics students learn how to learn, American Journal of Physics **69**, S54 (2001).
[36] B.A Beth, P.R.L Heron, and P.S Shaffer, Student understanding of energy: Difficulties related to systems, *American Journal of Physics* **80**,2 (2012).