

Providing Insight into the Relationship Between Constructed Response Questions and Multiple Choice Questions in Introduction to Computer Programming Courses

Joseph B. Herzog R.B. Annis School of Engineering University of Indianapolis Indianapolis, IN, USA herzogjb@uindy.edu	Patricia Snell Herzog Lilly Family School of Philanthropy Indiana University Purdue University Indianapolis Indianapolis, IN, USA psherzog@iupui.edu	Paul Talaga R.B. Annis School of Engineering University of Indianapolis Indianapolis, IN, USA talagap@uindy.edu	Christopher M. Stanley R.B. Annis School of Engineering University of Indianapolis Indianapolis, IN, USA stanleyc@uindy.edu	George Ricco R.B. Annis School of Engineering University of Indianapolis Indianapolis, IN, USA riccog@uindy.edu
--	--	--	--	--

Abstract— This Research-to-Practice Work in Progress (WIP) investigates the format of student assessment questions. In particular, the focus is on the relationship between student performance on open-ended, constructed-response questions (CRQs) versus close-ended, multiple-choice-response questions (MCQs) in first-year introductory programming courses. We introduce a study to evaluate whether these different response formats return distinct or comparable results. In order to assess this, we compare and correlate student scores on each question type. Our focus is on assessments (exams and tests) in first-year classes. The paper investigates two first-year programming courses with a total of seven sections and approximately 180 combined students. The subject of the sequential set of courses is the procedural C programming language. Based on extant studies comparing student performance on MCQs to their performance on open-ended questions, we investigate whether MCQ scores predict CRQ scores. Preliminary results on the comparison between student performance on these two question formats are presented to assess whether MCQs produce similar results as CRQs, or whether MCQs yield unique contributions. Possible avenues for future work are also discussed.

Keywords—Introduction to Programming, Constructed Response Questions, Multiple Choice Questions

I. INTRODUCTION

This Research-to-Practice Work in Progress (WIP) presents the design and first results of an ongoing study investigating the question of whether multiple choice questions (MCQs) lead to similar results compared to open-ended constructed response questions (CRQs). The study focuses on introductory computer science classes (introduction to programming and object oriented programming).

For the purposes of this study, MCQs are student assessments based on singular, close-response option questions. The format of these MCQs includes a range in the number of response options from two to four or more responses and distractors. In contrast, CRQs, as defined by Keuchler and Simkin, are “questions [that] require respondents to create their

own answers” [1]. CRQs are also known as written questions or open-ended questions. In particular, for this work, the CRQs are primarily questions that require students to create their own computer program to solve a problem, to de-bug a program, fix syntax, or complete a partial program. The MCQs are relatively comparable to these open-ended questions in content but are notably distinct in format. For example, a MCQ asks students to select one among possible options for a correct line of syntax, whereas a comparable CRQ asks students to find and fix a similar syntax error.

This study compares student performance on these distinctly formatted questions in order to examine whether the open-ended response format provides unique assessment relative to the close-ended response format. The MCQ response format is more efficient for students to complete, and for instructors to grade. Thus, the typical justification for the more labor-intensive, CRQ response format is that these questions provide the students a different form of assessment that yields unique results.

However, this widely held assumption has received limited direct empirical investigation, and the validity of this assumption has implications for teaching best practices. If students indeed perform distinctly on CRQs, relative to MCQs, then there is some support for the need to provide this more labor-intensive assessment format. If instead the students perform similarly on CRQs and MCQs, then questions are raised as to the utility of the more labor-intensive CRQ format and support is garnered for instructor investment in creating high-quality MCQs, which is more labor intensive during the preparation stage.

II. PRIOR WORKS

There is a long history of interest in educational studies as to whether assessment methods affect student outcomes. For example, more than 20 years prior, Scouller examined the relationship of a MCQ examination to an assignment essay [2]. This early study evidenced that, in comparison to the deep learning evidenced in essays, students evidenced more surface learning in MCQ assessment formats. However, more

This project was generously supported by the University of Indianapolis Office of the Provost and the Faculty Academy of Excellence and Innovation.

contemporary research instead finds that MCQ examinations are capable of producing desirable student outcomes, and additionally that – when well-constructed, MCQs are preferable due to their: (a) relative efficiency in student completion and instructor grading, (b) objectivity and reliability in evaluating all students the same regardless of instructor fatigue and other grading idiosyncrasies, and (c) accuracy in discriminating high-performing students from low performing-students [3-12].

Indeed, in response to the misperception that MCQs necessarily assess low-order cognitive skills, Sim & Rasiah state: “Concerns have been voiced that most MCQs tend to measure factual recall and recognition of isolated facts. But if carefully constructed, MCQs (especially one-best-answer-type) may also assess higher-order thinking skills.” [5]. Additionally, Nicol found the MCQs can enhance learning by providing students with greater learning autonomy generated from quicker feedback [7]. Moreover, more than simply relative substitutes, MCQs have been found to be better than open-ended questions in assessing higher-order cognitive skills [8], such as comprehension and application in Bloom’s taxonomy [13], deeming open-ended questions to be inferior due to their greater resource burdens and lack of reliability [8].

However, most these studies were in disciplines other than engineering, and thus the majority of the open-ended response formats focus on essays or case study analyses, which are often not relevant or readily applicable to engineering content. Yet, in a more closely related discipline of chemical engineering, Case & Fraser found that MCQs were an effective assessment of an intervention designed to improve students’ concrete thinking ability [14]. Additionally, Sorensen found that results from online MCQ assessments was predictive of summative end-of-semester assessment, and that students preferred the e-learning format of the MCQs [15]. With even greater applicability, several papers from IEEE and FIE also evidence the utility of MCQs and their relative benefit to other formats [16-18].

For example, in a study comparing MCQ formats to descriptive examinations, Brown made the following five observations: (1) “Examinations using multi-choice questions can be automatically graded in a fraction of the time and cost of manually grading descriptive question; (2) Multiple-choice questions are very much more difficult to write than descriptive questions; (3) Written properly, multiple-choice examinations correlate strongly with assessments by descriptive assessments; (4) For students numbering more than 30 or for ongoing automated assessment multi-choice assessments are highly advantageous if not essential; (5) There is a strong correlation between MCQ examinations and the more traditional descriptive examinations.” [16].

Yet, Ventouras et al. found that the utility of MCQs is highly dependent on the formulation of quality distractors, as low-quality response options increased the probability of students answering correctly by chance [19]. While MCQ assessments inherently have a degree of guesswork that can uncertainly improve student performance relative to their actual ability [20], eliminating low-quality distractors has been found to increase the accuracy of MCQs to assess student comprehension [21].

Moreover, Simkin & Keuchler investigated MCQs versus CRQs on assessments that were specifically designed for

computer programming classes and hypothesized that MCQs are fair in evaluating student understanding of course concepts [22]. In a similar study, Mbonigaba & Oumar investigate how MCQs and CRQs match relative to student cognitive ability in an Introduction to Management Sciences course [23]. Vasan et al. looks at the relationship of these assessment methods in an Anatomy course [24].

The implication of this theory of the relative utility of MCQs is that the time-intensive component of CRQs may be inefficient in assessing student understanding. If MCQs are comparably accurate but considerably more efficient in measuring student understanding, then MCQs would be preferable in most cases, due to the consistency of their objectivity and time efficiency for student completion and for instructor grading. However, this theory remains understudied, especially within engineering, and Simkin & Keuchler call for others to investigate this hypothesis in college-level classes, such as programming [22].

Informed by these prior works, this paper aims to add more data to the field and respond to the call for others to continue to reveal insights into this area, specifically in Introduction to Programming courses [22]. We here respond to this call with an empirical study constructed by an interdisciplinary team. Collectively, our disciplines represent engineering, computer science, engineering education, and sociology. We thus integrate social science theories on learning and data collection techniques on human subjects with techniques of engineering education, along with the content expertise of computer science.

Our work contributes to the field by further empirically investigating the comparison between student performance on MCQs and CRQs in programming courses. Analysis of open-ended problems has a robust and significant place in engineering education history and recent work. From design studies, for instance, analysis of open-ended problems is the norm [25]. Frameworks of open-ended analysis in education and engineering education span across teamwork and collaborative research [26, 27], to work on capstone design courses [28]. A key thread through these lines of work is the establishment of new ways of understanding student learning and development; and yet attention to MCQs online as an innovative assessment of student learning is underdeveloped.

III. METHODS

This work investigates student MCQ score and CRQ scores in first-year computer programming courses. The courses include Introduction to Programming and a first-year Object-Oriented Programming with a total of seven sections and approximately 180 students. Sections are defined as unique meeting times of the course, each with a unique roster, and sometimes with different instructors. Each section had between two and four total assessments (exams or tests); where tests are just like exams, but with fewer total questions. Each assessment had between 4 and 30 MCQs and 1-5 CRQs. The MCQs were a combination of two-response options (e.g. true-false questions) and single-selection from four-or-more response options,

including terminology selection and code analysis. The CRQs had multiple subparts to the questions that consist of programming problems. These CRQs evaluate proper programming style, identifying and fixing both syntax and logical errors in the code, completing a partially-written code, and creating a code from scratch to solve a simple problem. This resulted in nearly 400 total data points comparing MCQ scores to CRQ scores, where each data point is a single assessment attempt, containing both MCQs and CRQs. For each Assessment, the MCQs and CRQs covered the same topics.

Assessments were designed mostly independently by three different instructors across these courses. For all the assessments, the MCQ part was closed-book, while the CRQ part was open-book. Some of the MCQs were hand-written, while others were completed online with no back-tracking and with random answer order. The CRQs were delivered online and students used an integrated development environment (IDE) for creating a computer program. Pearson r correlation coefficients are computed to assess the strength of the relationship between MCQ and CRQ results.

IV. RESULTS

Fig. 1a plots all of the assessment results for three different data sets. Each data point plots a single student's score for a single assessment located at the MCQ score and the CRQ score. For example, if a single class had three assessments, then each individual student has three data points, one for each assessment. Fig 1b is a two-dimensional histogram, plotting the frequency of the data points for a small range of percent correct for CRQ and MCQs. Fig 1b is another way to represent the same data in Fig. 1a, and all of the data from Fig 1 comes from three different datasets that are plotted individually in Fig. 2, and are described below. Collectively, all of the data has a correlation coefficient of 0.55 between the MCQs and the CRQs, indicating a relationship. These results are somewhat similar to results in Simkin & Keuchler's Fig. 3 [1].

A. Dataset A

Dataset A (Fig. 2a) is collected from three sections taught by Instructor I. It includes two different courses, Introduction to Programming and Object Oriented Programming (OOP). The Introduction to Programming data come from two different semesters of data, where one semester has two assessments (a midterm and a final exam). For the midterm exam there were four MCQs, and the final had 18. The data from the other semester comes from three assessments with 8, 10, and 10 MCQs. The data from the OOP class comes from one assessment which has 24 MCQs. The assessments in this dataset had between 2-5 CRQs. The correlation coefficient of the entire Dataset A is 0.45, a moderately high relationship.

B. Dataset B

Dataset B (Fig. 2b) is collected from three sections taught by Instructor J. It also includes two different courses, Introduction to Programming and OOP. The Introduction to Programming dataset comes from two sections in the same semesters with each section having three assessments during the semester, for each section there was not a significant difference between each

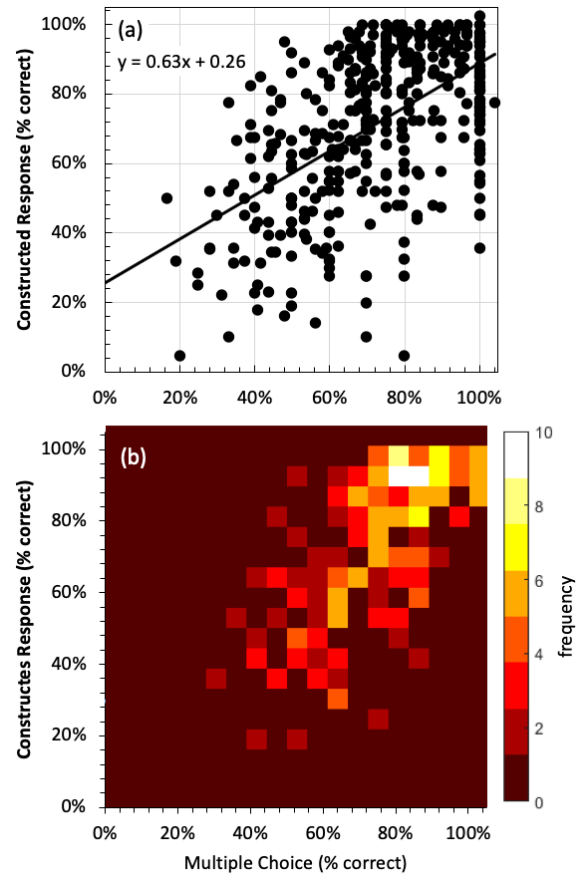


Fig. 1. Relationship between MCQs and CR question of all data with each data point plotted in (a), and (b) is a 2D histogram of the results.

assessment. The three assessments for the Intro course had 26, 24, and 33 MCQs, and the assessments in this dataset had between 1-2 CRQs. This dataset has a moderately high MRQ-CRQ correlation coefficient of 0.50.

C. Dataset C

Dataset C (Fig. 2c) is collected from two sections in the same semester of an Intro course taught by Instructor K, having a total of four assessments (three midterm and one final exam). The assessments had 24, 32, 32, and 32 total MCQs. The MCQs were the same question for each section with random order and random answer order as well. The CRQs for the assessments were 90% similar between each section. The assessments had 2-3 CRQs. Instructor J and K worked closely when preparing the MCQs and CRQs for the Introduction to Programming course and used course notes from Instructor I. Dataset C has a correlation coefficient of 0.75, which is the strongest relationship of all 3.

D. Effect of number of MCQs

Fig. 3 plots the correlation coefficient between MCQs and CRQs ($\rho_{M,C}$) for each individual assessment as a function of the number of MCQs on that assessment. Fig. 3 shows that there is a positive relationship between the number of MCQs and the correlation for both the A and C data together. However, the

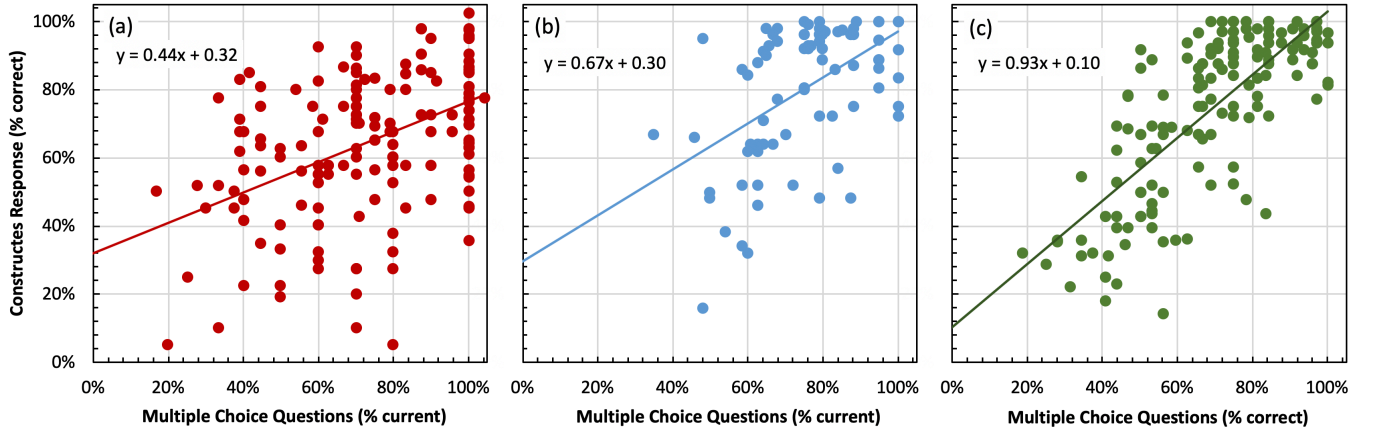


Fig. 2. Relationship between MCQs and CRQs from three different data sets (a), (b), and (c).

correlation coefficient between $\rho_{M,C}$ and the number of MCQs ($\rho_{\rho,N}$) for all three datasets combined is 0.425, not very strong. However, in considering each dataset individually, the correlation coefficient between $\rho_{M,C}$ and $\rho_{\rho,N}$ is 0.638 for dataset A, -0.734 for dataset B, and 0.759 for dataset C. This suggests that for both datasets A and C, a greater number of MCQs yields a stronger correlation between the results of MCQs and CRQs. However, for dataset B, the opposite was the case. More investigation is needed to determine the reasons for this disparity across instructors.

V. DISCUSSION

The results of this study show that: (a) there is a correlation between MCQs and CRQs, and (b) that there is variation across instructors in the strength of this correlation. More work is necessary to investigate why the variation exists. One reason is the number of MCQs (as shown in Fig. 3), as well as content. Other factors can include teaching style of each instructor (for example, instructors may differ in the extent to which they review examples for each type of question), differences between closed and open-book formats, or variations in instructor MCQ distractor quality. Additionally, difficulty of course content can be a factor, since datasets span two different classes. However, the difference in correlation $\rho_{M,C}$ is not that significant with

Introduction to programming, $\rho_{M,C} = 0.54$, and Object Oriented Programming, $\rho_{M,C} = 0.56$. Of particular importance is performing an item analysis of each question.

VI. CONCLUSION AND FUTURE WORK

In conclusion, this study compares CRQ and MCQ results in order to examine their relationship in assessing student performance. As this WIP continues, so will the range of analyses employed in order to validate the reliability of these results. Future work is needed to further investigate the differences between each dataset, which includes investigating potential effects of teaching styles, content and approach of both MCQs and CRQs, and ranges of students ability within each dataset. D. Clark's study of types of MCQs will be used to help classify each question type and study their effect on final score outcomes [21]. We are also developing a biserial (item analysis) model for the express purpose of gleaning determinant and discriminatory factors from the ranges of questions asked (both on open ended and discrete ends) [29]. With these additional analyses, this WIP will contribute to better understanding student assessment tools, including the relevance of CRQs relative to MCQs. Considering the labor-intensity of CRQs, both for students and instructors, better understanding their relative merits will advance engineering education and its scholarship.

AUTHOR CONTRIBUTIONS

JBH formulated the research idea and the method design, executed the analysis, managed the project, acquired funding, created the visualizations, and wrote the initial draft. PSH and JBH wrote the response document and made the revisions requested of reviewers. JBH, PT, and CMS developed the assessments and collected and provided the data. PSH, PT, and GR edited the draft. PSH wrote most of the Prior Works section.

ACKNOWLEDGMENTS

JBH would like to acknowledge Jonathan Mishler, currently a graduate student at Univ. of Washington in the Dept. of Bioengineering, for sharing some of his knowledge which has benefited this work. We are grateful to the FIE program co-chairs and the four reviewers for providing extensive attention and input into our paper.

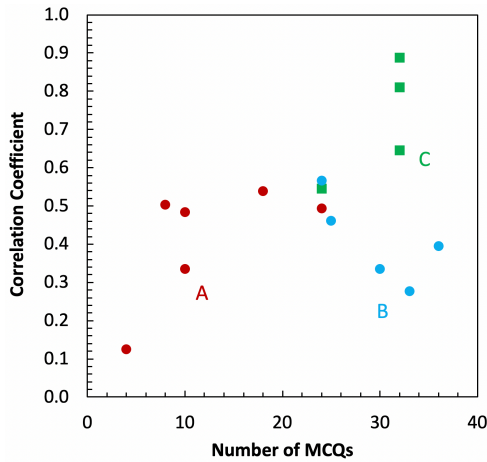


Fig. 3. Relationship between correlation coefficient, between MCQs and CRQs, and the number of MCQs on exam.

REFERENCES

- [1] W. L. Kuechler and M. G. Simkin, "Why is performance on multiple-choice tests and constructed-response tests not more closely related? theory and an empirical test," *Decision Sciences Journal of Innovative Education*, vol. 8(1), pp. 55, Jan 2010.
- [2] K. Scouller, "The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay," *Higher Education*, vol. 35, no. 4, pp. 453–472, Jun. 1998.
- [3] A.-M. Brady, "Assessment of learning with multiple-choice questions," *Nurse Education in Practice*, vol. 5, no. 4, pp. 238–242, Jul. 2005.
- [4] M. C. Rodriguez, "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research," *Educational Measurement: Issues and Practice*, vol. 24, no. 2, pp. 3–13, 2005.
- [5] S.-M. Sim and R. I. Rasiah, "Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper," *Ann. Acad. Med. Singap.*, vol. 35, no. 2, pp. 67–71, Feb. 2006.
- [6] M. Tarrant, A. Knierim, S. K. Hayes, and J. Ware, "The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments," *Nurse Education Today*, vol. 26, no. 8, pp. 662–671, Dec. 2006.
- [7] D. Nicol, "E-assessment by design: using multiple-choice tests to good effect," *Journal of Further and Higher Education*, vol. 31, no. 1, pp. 53–64, Feb. 2007.
- [8] E. J. Palmer and P. G. Devitt, "Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper," *BMC Medical Education*, vol. 7, no. 1, p. 49, Nov. 2007.
- [9] M. Tarrant, J. Ware, and A. M. Mohammed, "An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis," *BMC Medical Education*, vol. 9, no. 1, p. 40, Jul. 2009.
- [10] G. Mehta and V. Mokhasi, "Item Analysis of Multiple Choice Questions- An Assessment of the Assessment Tool," *International Journal of Health Sciences*, no. 7, p. 6, 2014.
- [11] J. Mbonigaba and S. B. Oumar, "Multiple-Choice Questions and Written Questions matched according to levels of cognitive ability in an applied course: Evidence and practical implications," *Africa Education Review*, vol. 14, no. 1, pp. 139–154, Jan. 2017.
- [12] C. A. Melovitz Vasan, D. O. DeFouw, B. K. Holland, and N. S. Vasan, "Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course," *Anat Sci Educ*, vol. 11, no. 3, pp. 254–261, May 2018.
- [13] B. S. Bloom, *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*, 2nd edition Edition edition. New York: Addison-Wesley Longman Ltd, 1956.
- [14] J. M. Case and D. M. Fraser, "An investigation into chemical engineering students' understanding of the mole and the use of concrete activities to promote conceptual change," *International Journal of Science Education*, vol. 21, no. 12, pp. 1237–1249, Dec. 1999.
- [15] E. Sorensen, "Implementation and student perceptions of e-assessment in a Chemical Engineering module," *European Journal of Engineering Education*, vol. 38, no. 2, pp. 172–185, May 2013.
- [16] R. W. Brown, "Multi-choice versus descriptive examinations," in 31st Annual Frontiers in Education Conference. Impact on Engineering and Science Education. Conference Proceedings (Cat. No.01CH37193), 2001, vol. 1, pp. T3A-13.
- [17] R. W. Brown, "Undergraduate summative assessment experiences," in 34th Annual Frontiers in Education, 2004. FIE 2004., 2004, pp. F3G-5.
- [18] Y. Duroc and T. Vuong, "Multiple-Choice Question Enhanced with Interactive Software for Autonomous Learning," in 2008 Eighth IEEE International Conference on Advanced Learning Technologies, 2008, pp. 662–663.
- [19] E. Ventouras, D. Triantis, P. Tsiakas, and C. Stergiopoulos, "Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers," *Computers & Education*, vol. 54, no. 2, pp. 455–461, Feb. 2010.
- [20] Y. Zhao, "How to Design and Interpret a Multiple-Choice-Question Test: A Probabilistic Approach," *International Journal of Engineering Education*, vol. 22, Jan. 2006.
- [21] D. J. Clark, "Testing Programming Skills with Multiple Choice Questions," *Informatics in Education*, vol. 3, pp. 161–178, 2004.
- [22] M. G. Simkin and W. L. Kuechler, "Multiple-choice tests and student understanding: what is the connection?" *Decision Sciences Journal of Innovative Education*, vol. 3(1), pp. 73, Jan 2005.
- [23] J. Mbonigaba & S. B. Oumar, "Multiple-choice questions and written questions matched according to levels of cognitive ability in an applied course: evidence and practical implications," *Africa Education Review*, 14:1, 139-154, 2017.
- [24] C. A. Melovitz Vasan, D. O. DeFouw, B. K. Holland, N. S. Vasan, "Analysis of testing with multiple choice versus open-ended questions: outcome-based observations in an anatomy course," *Anat Sci Educ* 11:254–261 (2018)
- [25] R. Buchanan, "Wicked Problems in Design Thinking," *Design Issues*, vol. 8, no. 2, pp. 5-21, 1992.
- [26] E. Durall, T. Leinonen, and E. Durall-Gazulla, "Design Thinking and Collaborative Learning," *Comunicar /*, vol. 21, no. 42, pp. 107–116, 2014.
- [27] E. C. Johnson and M. C. Loui, "Work in progress - how do students benefit as peer leaders of learning teams?," in 2009 39th IEEE Frontiers in Education Conference, 2009, pp. 1-2.
- [28] P. H. Gregson and T. A. Little, "Using contests to teach design to EE juniors," *IEEE Transactions on Education*, vol. 42, no. 3, pp. 229-232, 1999.
- [29] S. I. Inc. (2014). *The Item Analysis Report*. Available: http://www.jmp.com/support/help/The_Item_Analysis_Report.shtml