How does the structure of a college chemistry examination affect pedagogy

Cómo la estructura de exámenes de una clase de química universitaria afecta la pedagogía

Rajeev R. Pandey, John Mayberry, Jace Hargis

University of the Pacific, Chaminade University , Honolulu, USA
rajeev.pandey@hotmail.com , jmayberry@pacific.edu, jace.hargis@gmail.com

Abstract

This study examines variations of assessment and connections to active learning methods, which may enhance both the accuracy of assessment, engagement and retention. Correlation data relating instruction and assessment in a multiple dimensions are presented. Multiple choice (MC) and free response (FR) exams were provided and students were also given the option to provide FR answers to the MC items. This study suggests there is little overall difference in mean or median student scores on the MC vs. FR portions of the exam, but that there is some evidence to believe that student scores on MC portions are more variable than their corresponding scores on FR portions. Some students may exhibit a difference in their abilities to answer MC vs. FR questions, but these preferences do not appear to be widespread and exhibit no biases towards one particular type of assessment.

Key words: *summative assessment, structured response test; unstructured response tests, active learning*

Resumen

Este estudio examina las variaciones de evaluación y conexiones a los métodos de aprendizaje activo, que puede mejorar tanto la precisión de evaluación, compromiso del estudiante, y la retención. Los datos de correlación con la instrucción y evaluación en varias dimensiones son presentados. Exámenes de opción múltiple y de respuesta libre fueron proporcionados y a los estudiantes también se les dio la opción de proporcionar respuestas libres a las preguntas de opción múltiple. Este estudio sugiere que hay poca diferencia global en media o mediana de las puntuaciones de los estudiantes en las porciones de opción múltiple y de respuesta libre en el examen, pero hay una cierta evidencia para creer que los resultados de los estudiantes en la parte de opción múltiple son más variables que sus correspondientes puntuaciones en las porciones de respuestas libres. Algunos estudiantes pueden mostrar una diferencia en su capacidad para responder de opción múltiple y respuesta libre, pero estas diferencias no parecen ser generalizadas y no presentan sesgos hacia un tipo particular de evaluación.

Palabras clave: *evaluación sumativa, prueba de respuesta estructurada; pruebas de respuesta no estructuradas, aprendizaje activo*

INTRODUCTION

In the Unites States (US), many K-20 do not excel in or enjoy STEM related fields of study. Since NASA's response to Sputnik, the US has tried to encourage young people to pursue math and science careers, however, the results are still less than desirable. Of particular interest are the complex sciences of chemistry and physics. Freshman chemistry is one of the most feared subjects among college students (Anderson, 2005) yet it is one of the core subject requirement for students wanting to pursue professional careers in the allied health and pharmaceutical sciences as well as certain areas of engineering and applied biological sciences. At the author's institution, for example, the majority of freshman students taking chemistry courses are either pre-med, dental or pharmacy students. There is therefore an authentic interest in exploring different ways to teach and assess introductory science classes to help students learn more effectively and empower them to create a solid foundation of science. This will, in turn, assist their success in their professional programs. In this study, we focus specifically on methods of assessment and examine the dependence of class scores and rankings on the form of assessment used for the course. We also examine connections between exam scores and various active learning methods (clickers and group work), which may enhance both the accuracy of assessment, engagement and retention of concepts.

This study was conducted collectively in two class sections of the same undergraduate general chemistry course taught by the lead author of this paper. The two class sections met one after the other throughout the spring semester on Monday, Wednesday and Friday each week from 11:00 AM to 12:15 PM and 12:30 PM and 1:45 PM respectively. The former section contained 46 students, while the second section had 17 students, producing a total of 63 students who participated in this study.

LITERATURE REVIEW

Assessment

Assessment is a vehicle for gathering information about learners' behavior. Measurement is an assignment of marks based on an explicit set of criteria. Evaluation is a process of making judgments about the level of understanding (Hargis, 2007). Each of these concepts used correctly can greatly assist a faculty member's ability to accurately assess the performance of a student. In the context of this study, we focus on summative assessment since the data examined were derived from final examination scores. Summative assessment can be narrowly defined as an instrument used to gauge a student's performance at a given time. This type of assessment typically does not provide information, which can be used to assist the students learning or remediation. It occurs at the end of a learning cycle (concept, chapter, semester, etc.) (Little, Badway & Hargis, 2008). Schmoker (2006) indicates the value of summative assessment is that it provides valuable information following a learning event to determine if foundational learning outcomes have been achieved. Race (2003) in his book "Designing Assessment to Improve Physical Sciences Learning" suggests the need for educators to get involved in getting the current methods of assessments fixed.

Dorman, Waldrip and Fisher (2008), report the use of a new instrument, the Students' Perceptions of Assessment Questionnaire (SPAQ) in middle school science classes that assesses Congruence with Planned Learning, Authenticity, Student Consultation, Transparency, and Diversity. Their method allows more focus on classroom-based perceptions of assessment rather than the traditional external accountability measures of classroom assessment. González-Espada (2008), in their work apply item difficulty and discrimination to analyze the quality of the multiple choice test items used to grade students enrolled in the "Introduction to Physical Science Laboratory" course. They suggest that by choosing multiple-choice items with optimal difficulty and discrimination, physical science instructors can develop the most effective and valid assessments possible.

Structured Assessment (Multiple Choice Tests)

Traditionally, multiple-choice examinations have been used in settings where efficiency is critical, both in the view of class time, large class sizes and publish or perish pressures on faculty members. Multiple-choice assessment can be an appropriate mechanism to accurately assess student performance, if the items are written attending to empirical research on effective tests. Attributes which have been identified for good practices include testing for important ideas instead of trivial facts in isolation; clearly worded questions and response options; uncluttered figure layouts; and consistent grammar. Practices to avoid include the use of interrelated items, irrelevant clues, direct quotations from the text, or trick questions; terms such as all, none, never, always, none or all of the; clues in the stem; use of long, complex sentences; unnecessary distractors; answers located in a previous question; and the use of "C" as a common response. Basic guidelines for writing multiple-choice items include addressing a single plausible concept for each item; providing three to five options; placing repeated words in the stem; avoiding window dressing; and using options of similar lengths placed at the end of statements.

Towns (2014) has published a guidance for chemistry faculty from the research literature on multiple-choice item development in chemistry that could allow faculty to create assessments that are reliable and valid, with

greater ability to discriminate between high- and low-achieving students. Along with good multiple choice test construction, the instructor should be responsible with analyzing the test items after the instrument is deployed. To do this, most automated grading machines provide foundational statistical analysis including difficulty level and discrimination index. If either of these indices are out of an acceptable range, the instructor should develop a consistent strategy to address these items (for example, discarding questions, providing point compensation, or retests). Finally, sufficient statistical analysis might include measures of central tendency (mean, median, mode) and dispersion (range, standard deviation, variance). Ultimately, each instructor should have an intentional, clearly developed strategy for addressing test result data and be able to defend an instrument's validity (measuring what it is intended to measure-content, criterion, construct) and reliability (consistency). Even after addressing the test construction parameters, the instructor should be aware of potential errors, both random and systematic. Random errors occur when responses differ un-systematically from one measurement to another and reduce the reliability of an instrument. Systematic errors occur when a test consistently measures something other than the intended effect. For example, a test for mathematics containing word problems given to English speakers of other languages will have limited validity.

Extensive research has been completed on the effectiveness of multiple-choice items. For example, a meta-study conducted by Vyas and Supe (2008) determined that three or four responses (two or three distractors) are optimal for discriminating between student achievements. They concluded that there was no significant advantages to the use of additional distractors and hence, it would be more efficient for faculty members to spend less time in creating distractors. Campbell (2015) reports a study in which variable number of response choices are used based on question type, and students are encouraged not to guess by giving them partial credit for answers left blank. Johnstone and Ambusaidi (2001 & 2001), through their work have discussed certain limitations of "conventional" fixed response tests and offer three fixed-response questions designed to overcome some of the limitations.

Unstructured Assessment (Free Response Tests)

Typically unstructured assessments are thought of as short or extended response items where students read a question stem and provide their response without the assistance of choices. These items are common in math and science disciplines where "work", formulas, and equations are critical. One advantage of unstructured items is that that they are relatively easy to prepare (Haynie, 1983).

Guidelines for writing unstructured, short response assessments include stating the question in a way that only specific and unique information can be correct; omitting only significant words from a statement; and specifying a degree of precision for numerical problems. In addition, several specific items are better than one broad item; the use of optional essay items should be avoided; a list of main points or rubric should be constructed; material should be reviewed by graders before assigning scores; scoring should be anonymous; and students should be informed in advance of how they will be scored (Zimmerman, Sudweeks, Shelley, & Wood, 1990).

Regardless of the category of assessment (summative/formative), or structure (multiple choice/free response), curriculum and instruction play a vital role in how students are able to process the information, and subsequently their achievement on assessments. To this end, Black and William (1998) remark that assessment is not an isolated incident, but one, which reflects an active learning environment with frequent feedback mechanisms.

Active Learning

It seems an obvious notion that the way in which we teach should parallel the way in which we assess students. Hence, if we provide information in a linear way, students will encode in a linear fashion. Subsequently, if we assess in a linear way, students will be able to decode this information. Although the manner in which we process information is more complex, general rules of coding/decoding do occur. In this study, active in-class teaching methods were regularly used throughout the semester in the form of student response systems (clickers) to train students on MC test taking and collaborative project-based learning in the form of group activity were used to enhance problem solving skills on FR type questions. Improvement in student problem solving skills has been demonstrated by having students work collaboratively in groups (Copper et. al., 2008). For the group activity, students were grouped together based on their scores from a survey on the first day of the class using a Quasi-Diagnostic Instrument that included a series of questions to be answered on a survey sheet. The questions were roughly based on two broad categories: assertiveness and emotional control Students were assigned points for their responses and these scores were used to classify each student as one of the four personality types: driver, analytical, amiable or expressive. Groups were then formed using one student of each personality type (Bender, 1997). Although some people view chemistry as a linear concept, in actuality, there are numerous permutations, which persist both in theory and application.

The literature is extensive on the effectiveness of active learning on gaining attention, processing from working memory to long term memory, and retention (Bean, 1996; Johnson, Johnson, & Smith, 1998; Vernon, & Blake, 1993; and Hake, 1998). The purpose of this study is not to defend, or validate active learning, but to determine an association between active learning and assessment type (structure vs. unstructured).

METHODS

This study was conducted during the final examination of a second semester undergraduate general chemistry course in the Chemistry department at a private university in northern California during the spring semester of 2011. Sixty-three students participated in this study. The majority of students were freshman although a significant number of sophomore students were also included. Most of these students were either pre-dental or pre-pharmacy majors who are required to take a two-semester undergraduate organic chemistry course after passing this general chemistry class.

This study was designed to compare and contrast the use of multiple choice (MC) and free response (FR) questions in student evaluations. Sixty-three students in a second semester undergraduate general chemistry course were randomly assigned to two groups: Group A, which consisted of 33 students and Group B, which consisted of 30 students. Both groups were given a common final examination consisting of 49 questions. The structure of the first 33 questions were MC for both groups. The two groups differed in regard to the format of responses for the last 16 questions. Group A was given questions 1-8 as MC questions and questions 9-16 as FR questions; whereas Group B was given the reverse scenario: questions 1-8 as FR and 9-16 as MC. (Note: Although this design is similar to a two-way ANOVA, the responses of students in the various factor combinations (question numbers and answer format) are not independent since the same students who answered Q 1-8 as MC questions also answered Q 9-16 in free response.) This design was conceived out of practicality (all students needed to complete the exam) and fairness (students should have the same number of MC/FR questions). An additional measure to guarantee fairness, additional white space on the exam papers were provided for explanations in the MC portion and this content was evaluated for partial credit in the assignment of final grades. The scores which take the addition of partial credit points into account are referred to as adjusted scores (Adj) in the following discussion. Overall, scores are computed from the Adj and FR scores.

The objectives were to examine (a) overall differences in student responses to FR and MC questions; and (b) correlations between FR and MC scores for individual students. We realize that even if investigations into the former objective yield no significant differences, the latter is still important to determine if the same "type" of students perform well on FR and MC questions or if individual differences can be found despite overall similarities.

RESULTS

The summary statistics for the various portions of the exam are shown in the table 1. Mean scores on the common portion suggest that Group B had a slight inherent advantage over Group A, but this difference in groups was not statistically significant (One-way ANOVA for difference of group means, p-value = 0.64) as should be expected from the random assignment. The similarities between the two groups are further (and perhaps more strongly) evidenced by the similarity in quartiles and medians between the two groups. The distributions of scores on Q 1-8 and Q 9-16 exhibited a strong left skew and were multi-modal so we used non-parametric measures of comparing responses on these items. The overall scores on Q 1-8 and Q 9-16 indicate that there was a significant difference in question difficulty (Wilcox Signed-rank test for a difference in median overall scores, p-value = 0.002). Similarities between groups, however, suggest that comparing and contrasting the MC and FR responses between groups ((i) Group A vs.

Group B on Q 1-8; and (ii) Group A vs. Group B on Q 9-16) are likely the most accurate measures of assessing one of our objectives, namely overall performance differences in student scores on MC vs. FR exam questions. Tests for group median differences in these two scenarios did not yield any significant results (Mann-Whitney test, p-value = 0.15 for (i), 0.20 for (ii)).

It is interesting to note that despite these similarities, there is a highly significant difference between overall achievement on the common portion (Q 1-33) and total scores on the non-common portion after adjustment for partial credit (Signed-rank test for difference in percentage scores, p-value < 0.001). However, it is difficult to determine if this difference is due to the MC vs. FR assessment structure or other factors such as question difficulty and placement of exam questions.

Table 1. Summary statistics for the various portions of the exam.

Questions	Group	Mean	SD	Min	First Quartile	Median	Third Quartile	Max
Common (99)	A (MC)	64.64	12.30	30	57	69	72	90
	B (MC)	66.20	14.40	36	58.5	69	75	93
Q1-8 (40)	A (MC)	26.21	10.46	0	20	30	35	40
	A (Adj)	31.61	8.02	11	26.5	35	37	40
	B (FR)	30.07	7.02	12	26.75	31	35	40
	Overall	30.87	7.54	11	27	33	36	40
Q9-16 (40)	A (FR)	29.79	7.31	15	23	32	35	40
	B (MC)	31.50	8.82	10	28.75	35	40	40
	B (Adj)	34.33	7.18	14	31	38	40	40
	Overall	31.95	7.5	14	28	34	38	40

One notable difference in MC and FR questions was in the variability of responses; MC scores were bi-laterally more variable (as measured by the standard deviation or inter-quartile range) than FR scores when contrasted between groups (comparisons (i) and (ii)) and questions (comparing Group A on Q1-8 vs. 9-16 and Group B on Q1-8 vs. 9-16). This difference in variability makes intuitive sense since questions are "all or nothing" in a MC setting.

We now move on to address our second objective; differences in individual student performances on MC vs. FR items. The non-normality of Q 1-8 and Q 9-16 scores may render measures of association based on Pearson's correlation coefficient unreliable so we compare individual performances by counting concordant and discordant pairs in rankings based on MC vs. FR items. For example, Student 1 is tied for the fifth lowest score on Q 1-8 in Group A and has the ninth lowest score on Q 9-16 while Student 9 is tied for ninth lowest on Q 1-8, but fifth lowest on Q 9-16. Therefore, Students 1 and 9 form a discordant pair: Student 1 ranks lower than Student 9 on Q 1-8, but higher on Q 9-16. Kendall's Tau-b is a non-parametric statistic related to the difference in the proportions of concordant and discordant pairs in the sample. Positive values of Kendall's Tau-b indicate a higher proportion of concordant pairs and the closer this statistic is to 1, the more concordant the two ranking systems. The values of Kendall's Tau-b for comparing scores on Q 1-8 vs. Q 9-16 are 0.52 for Group A and 0.38 for Group B. While these are statistically significant correlations, it is surprising that they are not higher given the previously discussed group similarities on these MC vs. FR items. One potential explanation for this discrepancy is that some individuals actually do differ in their abilities to answer MC vs. FR questions, but that these differences average out when comparisons are made at the group level.

To further examine the effect of question type on individual differences, we examined the differences in rankings before and after partial credit was assigned to MC questions (Adj ranking – MC ranking). Rankings were assigned in descending order: rank 1 was assigned to the highest score, rank 2 to the next highest and so on. Note that ties were assigned the average value of the corresponding ranking (for example, if two students tied for the fifth highest scores, both received a ranking of 5.5). The scatterplot in Figure 1 presents each students decrease in ranking as a function of their MC ranking and illustrates the fact that although individual ranking changes ranged from -7 to +9.5, there was only a small and insignificant negative correlation between MC rankings and ranking change (Kendall's Tau-b = -0.09, r = 0.13, p-value for non-zero slope of regression line = 0.49); i.e. there was at least no bias in ranking changes. An individual with a ranking in the higher range seemed just as likely to increase and decrease in ranking when partial credit was included as an individual in the lower rankings.

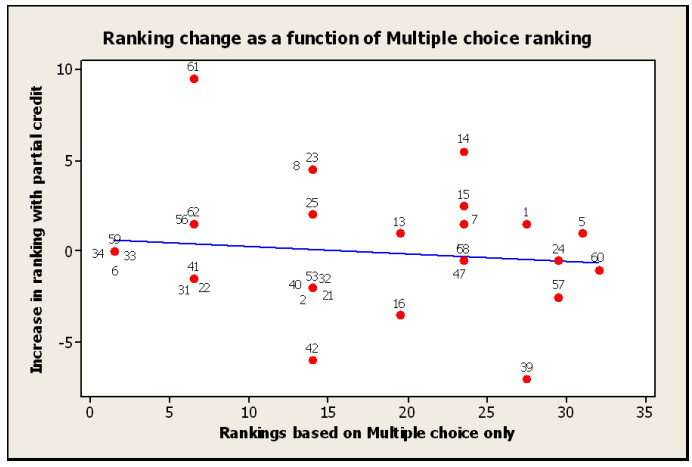

Figure 1. Student decrease in ranking as a function of their MC ranking.

A closer looks at the extreme cases suggests some possible factors which may lead to very large discrepancies in ranking changes. For example, student 61 ranked 9.5 places lower on Q 1-8 after partial credit was included. This student also ranked higher on their common MC portion (8th highest) than overall (10th) and ranked still lower on Q 9-16 FR (13th). Thus this may be an example of a student who performs much better relative to their peers on MC than FR. In contrast, student 42 who ranked 6 places higher after the inclusion of partial credit ranked extremely high on Q 9-16 FR (tied for second) and ranked 9th overall, but ranked only 13th on the common MC portion. Hence, this student may be an example of a student who performs much better relative to their peers on FR and MC questions. The other student with a drastic ranking change (student 39) illustrates a completely different scenario. The students ranked relatively high on the common MC (8th overall) and low on Q 9-16 FR (tied for 18th) even though their ranking on Q 1-8 increased by 7 spots after the inclusion of partial credit on this portion. It would be interesting to perform future investigations into individual question type preferences and effects on assessment.

We also compared student performance on group activity quizzes (which took the form of FR activities) and clicker quizzes (which took the form of MC activities) with their performance on MC portion (Common MC) and free response portion (Q 1-16 Total) of the final exam. The resulting correlations are shown in the table 2.

Table 2. Student performance on group activity and clicker quizzes with MC and FR.

	Q 1-16 Total	Clicker Quizzes	Group work
Common MC	0.77	0.54	0.42
Q 1-16 Total		0.50	0.43
Clicker Quizzes			0.70

All correlations are highly significant (P-value < 0.001)

Although both Clicker and Group work scores were similarly correlated with student performance on both portions of the exam, they are also highly correlated with one another. There is no evidence of a relationship between Group work scores and performance on the final exam after accounting for Clicker quiz scores (Extra sum of squares *F*-test, p-values = 0.59 and 0.34 when Common MC and Q 1-16 totals are respectively used as response variables). In other words, it appears that Clicker quiz scores were a better predictor of student performance on both the MC and FR portions of the exam.

DISCUSSION

This study suggests there is little overall difference in mean or median student scores on the MC vs. FR portions of the exam, but that there is some evidence to believe that student scores on MC portions are more variable than their corresponding scores on FR portions. In addition, some students may indeed exhibit a difference in their abilities to answer MC vs. FR questions, but these preferences do not appear to be widespread and exhibit no biases towards high/low achieving students.

Assessment results on the active learning techniques used in this study indicate that the group activity quizzes (FR) over the semester had three percent better average than individual clicker based quizzes (MC). This result supports the hypothesis that group work promotes better learning via social cognition. It would be interesting to have a direct comparison of individual clicker quizzes and clicker quizzes in groups. Between the two active learning methods, we found that clicker quizzes are a much better predictor of performance on both the sections of the final exam. However, since the exam is an individual student effort this result should not be too surprising. It would be interesting in the future to compare student performance on individual FR quizzes throughout the semester with their performance on the final exam.

Assessment based group size ranging from two to four students when compared to individual student scores on an in-class quiz did not show any major differences in scores. However, this comparison was limited to one single assessment, more studies of this type would be needed to further understand the role of group size in student learning outcomes.

The two forms of assessment methods explored here are very different from each other, and require students to develop different cognitive strategies. Therefore, to most accurately assess what students actually comprehend, apply, analyze, and connect, the ideal assessment strategy is to deploy both MC and FR format. This approach incorporated into course work and assessment would promote a multi-dimensional approach to problem solving and thereby enhance learning for a wide range of student abilities, aptitude and interest. The students' apprehension towards a particular testing format can be addressed by moving beyond linear teaching and learning methods and incorporating active learning methods. Active learning methods can both engage students, as well as enable professors to collect formative assessment data, which allows them to redirect, or possibly remediate instruction in-situ. A high frequency of formative, real-time low-risk assessment provides opportunities for students to share what they know at a given time, as well as providing a broader voice for the instructor to make decisions on whether to precede onto another topic.

The two sets of questions i.e. Q1-8 and Q9-16 were chosen to be of similar level of difficulty with a few matching questions on a given topic, however, they are not exactly similar and therefore can be regarded as non-normal. Comparing the student performance on these questions in terms of multiple-choice between the two groups, it is seen that the group that had Q9-16 as MC did better in terms of mean score by about five points on the MC section compared to the group that had Q1-8 as MC. While at first one may attribute it to the non-normality of the two sets of questions, it is interesting to note that when one compares the performance of the two groups for the same questions in terms of FR questions, the difference in the mean scores is nearly zero.

Analysis of student performance between the two sets of MC questions with the addition of partial credit allotted for incorrect MC questions for work shown while solving the MC problems, shows an improvement in the mean MC score for group A, i.e. the group taking Q1-8 as MC by a factor of five points. For group B, the mean MC total improved by a factor of slightly less than three points. To summarize the student performance on this test, while taking into account factors like the non-normality of MC questions, we find that group B performed better on the MC questions compared to group A. This is also supported by the fact that group B edged ahead of group A by about two points on the mean score of the Common (99) section of the test which was same for both the groups and was all MC.

The tests used in-class and for this study were created from a question bank provided by the publisher of the textbook used for the course. The majority of the questions are in multiple choice (MC) format, in addition to several open/free response (FR) questions for each chapter. For this study, 16 MC questions from the topics covered over the semester were selected and divided as described in the methods section of this article. The MC and the FR questions for the two groups were similar in difficulty level. The answer key for the MC questions is provided by the publisher and was verified for accuracy by the professor, who completed the questions and compared to the results provided by the publisher. For the FR question, a grading key rubric was designed for all the questions with partial credit allocated at different steps in the problem, depending on the amount and quality of work provided.

The results presented are combined results from both the class sections (46 and 17, for a total of 63 students), and the performances and assessment results between the two sections have not been investigated at the individual student level. The goal of the study was not to address individual students, or their performance, but a random aggregate, total results to assist instruction. No significant difference was noticed between the two class sections qualitatively. It would, however be interesting to further compare and contrast the individual performances between the two sections especially in the context of the difference in class size. The maximum number of students taught in a single class section during the course of a semester by the lead author was 60 students in the prior academic year and the minimum number of students taught was 17 during the course of this study. It would therefore be special interest to study the effect of student - teacher ratio at the college level in terms of student pedagogy. Similarly, comparison between active learning techniques employed in this study versus the traditional teaching technique of plain lecturing and assessment in form of few midterms and a final would further our understanding student learning in this subject and other high or low achieving students.

CONCLUSIONS

In this paper we report variations of assessment in form of MC questions and FR questions and its connection to active learning methods with a goal to possibly enhance both the accuracy of assessment, engagement and retention. Students were assessed with MC and FR exams along with the option to provide FR answers to the MC items. Results suggest that there is little overall difference in mean or median student scores on the MC vs. FR portions of the exam. However, there is also some evidence to believe that student scores on MC portions are more variable than their corresponding scores on FR portions. Some students may exhibit a difference in their abilities to answer MC vs. FR questions, but these preferences do not appear to be widespread and exhibit no biases towards one particular type of assessment.

BIBLIOGRAPHY

Anderson, A. Students fearful of freshman chemistry. Athens GA: The Red and Black Publishing Company, Inc. 2005.

Bean, J. Engaging ideas: The Professor's guide to integrating writing, critical thinking, and active Learning in the classroom, 2nd ed. San Francisco, CA: Jossey-Bass Publishers. 2011.

Bender, P. U. Leadership from within. Toronto: Stoddart Publishing Co. Limited. 1997.

Black, P., & William, D. Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa*, 80(2), 1998.

Campbell, M. L., Multiple-choice exams and guessing: Results from a one-year study of general chemistry tests designed to discourage guessing, *Journal of Chemical Education*, DOI: 10.1021/ed500465q, published online: April 02, 2015.

Cooper, M. M., Cox Jr., C. T., Nammouz, M., Case, E., & Stevens, R. An assessment of the effect of collaborative groups on students' problem-solving strategies and abilities. *Journal of Chemical Education*, 85(6), 866-872, 1998.

Dorman, J. P. and Waldrip, B. G. and Fisher, D. L. Using the Student Perceptions of Assessment Questionnaire (SPAQ) to develop an assessment typology for science classes. *Journal of Science Education*, 1 (9), 13-17, 2008.

González-Espada, W. J. Physical science lab quizzes: Results from test item analysis, *Journal of Science Education*, 9(2), 81-85, 2008.

Johnstone, A. H., Ambusaidi, A. Fixed – response: What are we testing? *Chemistry Education Research and Practice in Europe*, 1, 323-328, 2000.

Johnstone, A. H., Ambusaidi, A. Fixed – response questions with a difference, *Chemistry Education Research and Practice in Europe*, 2, 313-328, 2001.

Hake, R. Interactive-engagement vs. traditional methods: A six thousand student survey, *American Journal of Physics*, 66(1), 1998.

Hargis, J. Teaching project-based assessment in 12 days in a developing country. *Journal of Excellence in College Teaching*, 18(3), 129-142, 2007.

Haynie, W. J. Student evaluation: The teacher's most difficult job. *Monograph Series of the Virginia Industrial Arts Teacher Education Council*. Monograph 11, 1983.

Johnson, D., R., Johnson, J., & Smith, K. Active learning: Cooperation in the college classroom, 2nd ed. Edina, MN: Interaction Book Co. 1998.

Little, T., Badway, N., & Hargis, J. Student learning outcomes assessment in Allied Health Education. *Journal of Faculty Development*, 22(2), 89-95, 2008.

Race. P. Why do we need to "repair" our assessment processes? A discussion paper,

Journal of Science Education, 4(2), 73-76, 2003.
Schmoker, M. Results now. How we can achieve unprecedented improvements in teaching and learning. Alexandria, VA: Association for Supervision and Curriculum Development. 2006.
Towns, M. H. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, 91(9), 1426-1431, 2014.
Vernon, D., & Blake, R. Does problem-based learning work? A Meta-analysis of evaluative research. *Academic Medicine*, 68(7), 550-563, 1993.
Vyas, R., & Supe, A. Multiple choice questions: A literature review on the optimal number of options. *National Medical Journal of India*. (3):130-3, 2008.
Zimmerman, B. B., Sudweeks, R. R., Shelley, M.F., & Wood, B. How to prepare better tests: Guidelines for university faculty. Provo, UT: Brigham Young University Testing Services. 1990.

Received 22-01-2015 /Approved 30-04-2016

A rethinking of assessment practice: an experience with a stage test

El replanteamiento de la práctica de la evaluación: una experiencia con una prueba en etapas

André Luis Trevisan, Regina Luzia Corio de Buriasco

Department of Mathematics , Federal Technological University of Paraná, Londrina, PR, Departament of Mathematics, State University of Londrina, PR, Brazil, andrelt@utfpr.edu.br, reginaburiasco@gmail.com.

Abstract

This article presents a study based on an experience involving a different assessment instrument (a stage test) in a second-year high school class. The work involving this use was developed over a semester, in order to take the written test as a learning task, and was developed under the perspective of assessment as a formative instrument present in the educational process both as a teaching and students' learning processes diagnostic tool and as a way to investigate pedagogical practice. Reflections originating from the written production of students in one of the questions of the test as well as a critical analysis of the instrument itself and the attitudes as a teacher are presented. The approach adopted is qualitative/ interpretative in the light of Content Analysis.

Keywords: *school learning assessment, stage test, written production analysis.*

Resumen

Este artículo presenta un estudio sobre una experiencia con un instrumento diferenciado de evaluación (una prueba en etapas) en una clase de la escuela secundaria de segundo año. El trabajo relacionado con este uso se desarrolló a lo largo del primer semestre, con el fin de tomar la prueba como una tarea de aprendizaje, y bajo la perspectiva de la evaluación como un instrumento de formación presente en el proceso educativo, tanto para el proceso de enseñanza y aprendizaje de los alumnos, como para herramienta de diagnóstico y una manera de investigar las prácticas pedagógicas. Las reflexiones se originaron a partir de la producción escrita de los estudiantes en una de las preguntas de la prueba, así como un análisis crítico del instrumento en sí y las actitudes que como docente se presentan. El enfoque que se adopte es cualitativa / interpretativa a luz del análisis de contenido.

Palabras clave: *evaluación del aprendizaje escolar, prueba en etapas; análisis de la producción escrita.*

INTRODUCTION

In this article, we present excerpts of a study about the use of a six-stage test in Math lessons. Although it is an investigation in the teaching of Mathematics, it can be applied to education of the sciences, and it brings contributions using modern active methods into teaching and assessment practices.

In opposition to the concept of Math as "an erudite discipline whose teaching is provided to all ages", Freudenthal (1979, p. 318) understands Math as a natural and social activity, the evolution of which follows that of the individual and meets the needs of an expanding world.

For Freudenthal (1979), Math is a both natural and social human activity, just like the speaking, drawing and writing. It is included among the first known cognitive activities to be taught. However, it evolved and changed, including its Philosophy and method, under the influence of social changes.

Under the Realistic Mathematics Education, a movement that gained power in Holland in the late 1950s and had as its forefather the mathematician Hans Freudenthal, students must be seen as active participants in the educational process. Situations that demand math organization should be proposed to the students. From these situations math concepts will arise as well as opportunities to reinvent math through a process of reality *mathematization* (De Lange (1987, 1999, 2003), Freudenthal (1979), Gravemeijer (2008), Gravemeijer & Terwel (2000), Van Den-Heuvel Painhuizen (1996)).

An assessment consistent with the RME must as an educational activity be formative and treat Math as a human activity, focused on meaningful activities. It should take into account that, during their development process, students go through several levels of mathematization and "create" their own math, offering them (imaginable) realistic contexts. Several other authors (Buriasco (2000); De Lange (1987, 1999); Esteban (2001, 2009); Hadji (1994); Van Den Heuvel-Panhuizen (1996); Viola dos Santos, Buriasco & Ciani (2008)) have referred to assessment as a formative instrument in the education process, both as a means to diagnose teaching and learning processes and as an instrument of pedagogical practice investigation. Along these lines, analyses involving the written production of students developed at GEPEMA (Group of Studies and Research in Mathematics Education and Assessment (http://www.uel.br/grupo-estudo/gepema/) are carried out under the perspective of assessment as an investigation practice and learning opportunity.

Besides being aware of this, we, as teachers, have often wondered how we could "operationalize" an assessment perspective to help us interpret, include, regulate and mediate teaching and learning processes. Barlow (2006, p.165) gives us a hint: it is necessary to kill the imaginary evaluator, by questioning and rejecting myths and rites and false appearances as well as to know how to revive it, by preserving or recreating "that which carries meaning and is rich in potential efficacy".

In an attempt to "kill" my own imaginary evaluator when I started my Doctoral studies I found myself in the position of a teacher/ researcher trying to reconceptualize assessment practice. So, at that time, my idea of assessment came down to "taking tests", thus changing assessment practice would imply modifying the instrument. Accordingly, I decided to experiment with a different written test format with my classes, inspired by some studies that used two-stage tests. It included a written test accomplished in two phases: in the classroom and with limited time (first phase) and generally at home, with more time (second phase). According to De Lange (1987), the two-stage test gives students the opportunity to reflect upon their work: after being taken at school for the first time, the test is corrected and commented on by the teacher and then returned to the students for additional work.

Menino and Santos (2004) and Santos (2004) report experiences about the use of the two stage test as an assessment tool applied to different levels of education. For Menino and Santos (2004), the second stage is based on "runs" offered by the teacher at the end of the first phase.The student performs the second stage in a period agreed to beforehand, working especially with open questions. According to Santos (2004), the second stage must include test questions of an open nature, such as exploration and research tasks. In these questions that allow for any degree of development of the