Graded Response Method: Does Question Type Influence the Assessment of Critical Thinking?

Sherry Fukuzawa[1,*]& Michael deBraga[2,†]

[1]Department of Anthropology, University of Toronto Mississauga, Mississauga, Canada

[2]Robert Gillespie Academic Skills Centre, University of Toronto Mississauga, Mississauga, Canada

*Correspondence: Department of Anthropology, University of Toronto Mississauga, Mississauga, ON., L5L 1C6, Canada. Tel: 1-905-569-4380. E-mail: s.fukuzawa@utoronto.ca

†Correspondence: Robert Gillespie Academic Skills Centre, University of Toronto Mississauga, Mississauga, ON., L5L 1C6, Canada. Tel: 1-905-569-4717. E-mail: michael.debraga@utoronto.ca

Received: December 4, 2018 Accepted: December 29, 2018 Online Published: February 14, 2019

doi:10.5430/jct.v8n1p1 URL: https://doi.org/10.5430/jct.v8n1p1

Abstract

Graded Response Method (GRM) is an alternative to multiple-choice testing where students rank options according to their relevance to the question. GRM requires discrimination and inference between statements and is a cost-effective critical thinking assessment in large courses where open-ended answers are not feasible. This study examined critical thinking assessment in GRM versus open-ended and multiple-choice questions composed from Bloom's taxonomy in an introductory undergraduate course in anthropology and archaeology (N=53students). Critical thinking was operationalized as the ability to assess a question with evidence to support or evaluate arguments (Ennis, 1993). We predicted that students who performed well on multiple-choice from Bloom's taxonomy levels 4-6 and open-ended questions would perform well on GRM involving similar concepts. High performing students on GRM were predicted to have higher course grades. ***The null hypothesis was question type would not have an effect on critical thinking assessment.*** In two quizzes, there was weak correlation between GRM and open-ended questions (R^2=0.15), however there was strong correlation in the exam (R^2=0.56). Correlations were consistently higher between GRM and multiple-choice from Bloom's taxonomy levels 4-6 (R^2=0.23,0.31,0.21) versus levels 1-3 (R^2=0.13,0.29,0.18). GRM is a viable alternative to multiple-choice in critical thinking assessment without added resources and grading efforts.

Keywords: critical thinking, question type, graded response, multiple choice, Bloom's taxonomy, assessment

1. Introduction

1.1 The Problem

Post-secondary institutions regularly include critical thinking as a learning outcome for their undergraduate programs (Behar-Horenstein & Niu, 2011; Ennis, 1991; Fliegel & Holland, 2013; Mahapoonyanont, Krahomwong, Kochakornjarupong, & Rachasong, 2010; Poondej & Lerdpornkulrat, 2015; Stanger-Hall, 2012; Stupnisky, Renaud, Daniels, Haynes, & Perry, 2008). This is largely in response to employers who have encouraged them to develop critical thinking in their students as transferable skills for success in the workplace (Barnett & Francis, 2012; Halpern, 2014; Hyytinen, Nissinen, Ursin, Toom, & Lindblom-Yianne, 2015). Today's workplace is such an information rich environment that it is important for individuals to be able to evaluate alternative forms of evidence (Ku, 2009). The development of critical thinking in education therefore, involves teaching techniques that emphasize "how to think" rather than "what to think" (Daud & Husin, 2004, p. 478).

1.2 Importance of the Problem

Although critical thinking is a common educational goal, teaching critical thinking is not a straightforward task. The definition of critical thinking, and consequently how to accurately assess it, is diverse in the literature (see Ennis, 1993; Facione, 2000; Fliegel & Holland, 2013; Halpern, 2001; Halonen, 1995; Kerkman & Johnson, 2014; Norris, 1989). Most definitions acknowledge that critical thinking is a process by which individuals form a solution and evaluate their thought process by demonstrating the reasoning they used to support their conclusions (Ennis, 1993; Facione, 2000; Halpern, 2001; Kerkman & Johnson, 2014). Wilson-Mulnix (2012, p. 465) states that the fundamental skills underlying all critical thinking definitions is the ability of the thinker to "acquire, develop, and exercise the facility to grasp inferential connections holding between statements". The definition, therefore, incorporates thinking critically about the course material through an assessment of the question with the application of evidence to support or evaluate the argument (Bassett, 2016; Ennis, 1993; Gellin, 2003; Loes, Salisbury, & Pascarella, 2015; Watson & Glaser, 2008). This involves a number of skills that are necessary for critical thinking such as interpretation, analysis, inference, decision-making (i.e. evaluating your reasoning), and problem solving (Ennis, 1985; Facione, 2000; Barnett & Francis, 2012). These skills correspond with the higher levels of Bloom's taxonomy, which emphasize higher order thinking versus rote skills, such as factual recall of course material at Bloom's lower levels (Barnett & Francis, 2012; Ennis, 1985; 1993; Krathwohl, 2002; Seaman, 2011).

1.3 Relevant Scholarship

It is widely held that critical thinking is best assessed through open-ended problems or argumentative essay questions, where students can demonstrate their reasoning and the ability to defend their arguments (Ennis, 1989; 1993; Halpern, 2003; Ku, 2009; Taube, 1997; Parmenter, 2009; Sommer & Sommer, 2009; Walstad & Becker, 1994). However, in large introductory courses, open-ended questions are often not feasible for practical reasons. Multiple-choice questions are used instead due to resource constraints and the efficiency of grading (Ennis, 1993; Kim, Patel, Uchizono, & Beck, 2012; Parmenter, 2009). There is much debate about whether multiple-choice questions are capable of testing critical thinking (Bassett, 2016; Kerkman & Johnson, 2014; Ku, 2009; Stanger-Hall, 2012). The very nature of multiple-choice means that students are required to select the correct option. This precludes them from demonstrating their reasoning and evaluating their own criteria to form a solution. They are unable to demonstrate synthesis, analysis, and problem-solving (Bassett, 2016; Ennis, 2003; Halpern, 2003; Norris, 1989; Norris, 2003). As a consequence, multiple-choice testing has been associated with information reproduction rather than information synthesis (Fink, 2003). Ironically, Parmenter (2009) found that students prefer multiple-choice testing over open-ended exams because multiple-choice often requires less study effort, involves less difficult questions, and includes the process of elimination or even guessing to find the correct answer. Alternatively, students thought that multiple-choice did not reflect their comprehension and knowledge of the material as well as open-ended questions, and that multiple-choice was less fair because it did not allow for partial credit (Parmenter, 2009).

Kerkman & Johnson (2014) point out that multiple-choice questions do involve a component of the definition of critical thinking because students must evaluate the relative validity of each option to decide which one best answers the question. Several studies have suggested that carefully constructed multiple-choice questions using the higher cognitive levels of Bloom's taxonomy (Bloom, 1956; Krathwohl, 2002; Seaman, 2011) may assess higher order thinking skills in students (Kibble & Johnson, 2011; Kim et al., 2012; Morrison & Free, 2001; Naastreom, 2009; Nevid & McClelland, 2013; Tiemeier, Stacy, & Burke, 2011; Su, Osisek, & Starnes, 2005; Zheng, Lawhorn, Lumley, & Freeman, 2008). Problem-based questions, practical scenarios, case studies, as well as data or image analysis have also been suggested as methods to develop critical thinking skills in multiple-choice testing (Azer, 2003; Bassett, 2016; Holmes, Wieman & Bonn, 2015; Ku, 2009; Morrison & Free, 2001; Su, Osisek, Montgomery & Pellar, 2009). Bassett (2016) and Ennis (1993) added a metacognition component to multiple-choice questions that allowed students to provide a rationale for their option choice. Similarly, Kerkman & Johnson (2014) gave students a multiple choice "challenge" in which students could argue against the correct solution after the exam answer key was distributed. The drawback to these creative attempts to develop critical thinking in multiple-choice tests is that they require additional resources and they have not been tested against open-ended exams. Hickson, Reed, & Sander (2012) found no difference in grade outcomes between multiple-choice and open-ended questions (called constructed response questions in their study) in a large sample (N=7754) from two first year undergraduate economics courses. They used the levels of Bloom's taxonomy to construct the questions and suggested that similar grade outcomes indicated a comparable level of higher order thinking between the question types. Similarly, Barnett & Francis (2012) found that critical thinking was significantly greater in students who were given higher order questions based on Bloom's taxonomy on quizzes regardless of whether they were multiple-choice or essay format. In fact, in their study students who were given higher order multiple-choice questions had a significantly greater increase in critical thinking as measured by the Watson Glaser Critical Thinking Appraisal (short form) over students who were given factual essay questions (Barnett & Francis, 2012).

1.4 Hypothesis and Correspondence to Research Design

Recently, deBraga, Boyd & Abdulnour (2015) suggested that ranking options rather than just choosing one correct answer in a multiple-choice question may be a way to incorporate critical thinking into the convenience of a multiple-choice format. This Graded Response Method (GRM) allows students to rank multiple-choice options in order of their relative accuracy to answer the question. They demonstrated a positive correlation (p = 0.388, N=31) between the GRM and an objective critical thinking test (i.e. Watson-Glaser Critical Thinking Appraisal) (deBraga et al., 2015; Watson & Glaser, 2006). The GRM is based on the graded response model that was developed by Samejima (1969) as a statistical archetype founded on the allocation of partial credit by the ordering of item responses (see Matteucci & Stracqualursi, 2006). Su et al., (2009) tested a similar concept to increase higher order thinking in multiple-choice exams in their nursing program. They created multiple-choice questions where the students were asked to choose the most appropriate option in a clinical scenario when all of the options were plausible. They found improvements in p-values (item difficulty) and point biserial index (ability to discriminate knowledge of course content) with the revised questions over standard multiple-choice. The authors suggested that students elicited higher order thinking because they were required to have a high discriminative capacity with the ability to evaluate multiple coinciding factors in order to answer the questions (Su et al., 2009).

The purpose of this pilot project was to examine the assessment of critical thinking in GRM questions by comparing them with open ended short answer questions and multiple-choice questions composed from the cognitive levels of Bloom’s taxonomy in an introductory undergraduate course in anthropology and archaeology (N=53 students). Critical thinking was operationalized in terms of the ability to assess a question with the application of evidence to support or evaluate the argument (Barrett, 2016; Ennis, 1993; Gellin, 2003; Loes, Salisbury, & Pascarella, 2015; Watson & Glaser, 2008). We predicted that students who performed well on multiple-choice questions from Bloom’s taxonomy levels 4 through 6 and open-ended short answer questions would also perform well on the GRM questions involving the same concepts. In addition, it was predicted that these high performing students would have higher overall grades in the course. ***Our null hypothesis was that question type would not have an effect on the assessment of critical thinking***.

2. Method

The introduction to biological anthropology and archaeology is a first-year undergraduate course at a large, research focused, Ontario University. This study was conducted during the summer session 2016. There is no pre-requisite for the course and the majority of students take the course in the summer session for a science breadth requirement. Two quizzes worth 15% each of the final grade (20 multiple-choice, 7 GRM, 5 open-ended short answer) and one final exam worth 40% of the final grade (60 multiple-choice, 8 GRM, 8 open-ended short answer) were comprised of all three question types (i.e. multiple-choice (MC), graded response method (GRM) and short answer (open ended questions). The multiple-choice (MC) questions were composed according to the levels of Bloom’s taxonomy. The three question types involved the same course material so that the investigators could assess whether question type was related to the ability of the participants to correctly answer the question. GRM questions were composed by the instructor (a primary investigator in the study) and reviewed by three teaching assistants in the course.

2.1 How Assessment Was Undertaken

Students were given an overview of the GRM in an introductory lecture by the investigators. In addition, two practice GRM questions were included at the end of their weekly lab assignments so that question familiarity was not a limiting factor in their ability to answer. Students could work in their lab groups to answer these practice GRM questions. They were then taken up and discussed with the teaching assistant. Collaborative small group, written lab assignments were worth 25% of the final grade, and participation in an online problem-based learning exercise was worth 5% of the final grade. The Student Opinion Survey at the end of the course invited students to comment on their perception of any differences between question types. The study examined the relationship between question type and the level of higher order thinking. The analysis of the data examined the correlation between student answers (N=53) for different question types that have been ranked on similar levels of the Bloom’s taxonomy and involved the same information (e.g. the same theory). Please refer to the Appendix for an example of the question types from this study.

2.2 Participant (Subject) Characteristics

Participants were registered in the summer iteration of an introductory anthropology course. The student population represented a diverse demographic and was made up of 53 participants. Students were provided with a formal consent letter identifying their role in the investigation and were assured that at no time would their participation negatively impact on their course grade. To ensure this procedure, the data was used as an aggregate and therefore no individual student could be identifiable to the course instructor.

2.3 Sampling Procedures & Data Gathering

Student performance on each test was compared across all three tests and only students completing all three tests (i.e., quizzes 1 & 2 and the final exam N=53) were included in the final analysis. Grade distribution was tested for normality using the Shapiro-Wilk test. The null hypothesis assumed a-priori that grades would not be normally distributed. Normal distribution of the data is essential to properly frame the analysis as strongly skewed data would be suggestive of additional variance that could not be readily explained. The null hypothesis was rejected and data was found to be normally distributed. Data was then subjected to a maximum likelihood analysis using the VGAM statistical package with the “vglm–tobit” command, which tests for the strength of the correlation between data that is representative of vector generalized linear models (Yee, 2015; Yee & Hastie, 2003). The analysis was completed using the open source software “R-studio” version 3.3.1 (2016-06-21) -- "Bug in Your Hair" Copyright © 2016 The R Foundation for Statistical Computing Platform: x86_64-apple-darwin13.4.0 (64-bit).

Multiple-choice (MC) questions were segregated and pooled by combining the scores for all questions on each separate test and then taking the average for all MC questions that were segregated prior to administration, using Bloom’s taxonomic categories. Questions identified as belonging to Levels 1, 2, & 3 were thus averaged for each test independently. The same procedure was repeated for MC questions assigned, under Bloom’s methodology, as belonging to Levels 4, 5, & 6. The average scores for each of the MC tests was then compared to the results for each of the average GRM and short answer scores for all three tests. The null hypothesis being tested, was that student performance on the GRM would not show any preferred correlation across MC question types. In other words, there should be no difference in the strength of the correlation between GRM scores and either MC level 1-3 or MC level 4-6. In addition, the null hypothesis also predicted that the GRM performance would not serve as a statistically valid predictor of student performance on open-ended short answer questions. Finally, we examined the predictive value of each question type with respect to students’ final grades as a combined average across all three tests. This project was approved by the Social Sciences, Humanities and Education Research Ethics Board according to the Tri-Council Policy Statement on Ethics (protocol reference number 33158).

3. Results

The data below is examined sequentially starting with a test for correlation between GRM and MC levels 1-3, followed by an evaluation of the same relationship between GRM and MC levels 4-6, and finally between GRM and short answer questions. When examining the strength of the correlation between GRM and MC levels 1-3, two tests (Quiz 1 and Final Exam) showed only a weak correlation, but Quiz 2 showed a moderate correlation between GRM and MC levels 1-3 question types (Table 1). Table 3 displays the strongest correlation between open-ended short answer questions and GRM questions, with an R^2 of 0.56.

3.1 Recruitment

Students involved in this investigation were enrolled in an introductory anthropology course and represented both science and non-science students. Recruitment was automatic with enrolment in the course.

3.2 Statistics and Data Analysis

Table 1. Test for Correlation between GRM and MC Testing among Lower Level (Level 1-3) Bloom’s Taxonomy Questions

Test Type	Null hypothesis	R^2 (strength of correlation)
Quiz # 1	rejected	0.13
Quiz # 2	rejected	0.29
Final Exam	rejected	0.18

Description: Test for strength of correlation between GRM and MC level 1-3 question types across all three tests (N=53).

Table 2. Test for Correlation between GRM and MC Testing among Higher Level (Level, 4-6) Bloom's Taxonomy Questions

Test Type	Null hypothesis	R^2 (strength of correlation)
Quiz # 1	rejected	0.23
Quiz # 2	rejected	0.31*
Final Exam	rejected	0.20

Description: Analysis for all three separate tests showed a moderate to moderately strong (*) correlation between student performance on GRM questions to MC level 4-6 question types.

Table 3. Test for Correlation between GRM and Short Answer (Open-Ended) Questions.

Test Type	Null hypothesis	R^2 (strength of correlation)
Quiz # 1	rejected	0.15
Quiz # 2	rejected	0.15
Final Exam	rejected	0.56†

Description: There was a very strong (†) correlation ($R^2 = 0.56$) between open-ended short answer and GRM questions.

4. Discussion

The results of this study illustrate that the graded response method (GRM) does correlate with higher order thinking questions of multiple-choice (according to Bloom's taxonomy), and generally for the open-ended short answer questions. However, the correlation with open-ended short answer questions was not consistent throughout the course.

There were consistently higher correlations between the GRM and MC questions from Bloom's taxonomy levels 4-6 (R^2 = 0.23, 0.31, 0.21 respectively) versus levels 1-3 (R^2 = 0.13, 0.29, 0.18 respectively). This suggests that the GRM questions are comparable to higher order MC questions and supports the notion that the GRM is a viable alternative to multiple-choice in the assessment of critical thinking without added resources and grading efforts. It has been argued that because students rank the relevancy of each option to the question in GRM, they use a higher degree of discrimination and inference between statements and therefore elicit greater critical thinking skills over traditional MC (deBraga et al., 2015; Su et al., 2009). In this way, GRM may reduce the tendency in forced choice testing towards student study strategies that emphasize ways to recognize the correct option rather than think critically about the inferences between options (Ennis, 1993).

It is important to keep in mind however, that other limitations of multiple-choice testing in the assessment of critical thinking may still apply to the GRM. There are inherent limitations in forced choice testing. For example, Ennis (1993, p. 181) points out that multiple-choice options do not allow students to come up with their own solutions and use creativity in problem solving based on "differences in the background beliefs and assumptions between the test maker and the test taker". Similarly, students are not able to provide a rationale for their rankings in GRM questions.

In our study, the comparison of GRM to open-ended short answer questions was less consistent throughout the course than the GRM to MC relationship. In both quizzes there was a weak correlation between the GRM and short answer questions ($R^2 = 0.15$), however there was a strong correlation in the final exam ($R^2 = 0.56$). This could reflect an increasing familiarity with the GRM questions as the course progressed. By the final exam, students had developed a test strategy to revise the GRM after they had written their short answers for the same course material. This aligns with Hickson et al., (2012) in their study of both multiple-choice and open-ended questions. They suggest that by including more than one question type in a single test it changes the study strategy of students in comparison to a multiple-choice only setting. In our study, the short answer questions in the final exam acted as a means for students to provide a rationale for their GRM rankings. It appears that it was not until the end of the course that the students realized that the GRM and short answers covered the same practical problems (see Appendix for an example of the question types). Another factor in the discrepancy between the correlations in the final exam versus the quizzes may be that the GRM and short answer questions on the final exam made up a larger portion of their final grade (40% of the final grade) than the quizzes (15% each). The students had more time and appeared to make more of an effort to work through the GRM and short answers in the exam. A testing effect may also have played a part in the higher correlation between exam GRM and short answer questions (Toppino, Thomas, & Cohen, 2009; Veltre, Cho, & Neely, 2015). The course material for the final exam included the information from both quizzes, but the quizzes themselves did not have overlapping information. Before the exam there was an optional review session with sample quizzes where the students practiced GRM questions. Fukuzawa, Goodman, & Jankowski (2015) demonstrated in an earlier iteration of the same course that self-selected attendance at the exam review resulted in an overwhelming participation by students with term grades of 80% or more. Review session attendees performed significantly better on the final exam in comparison to other students in their grade cohort for all grade categories.

Inconsistencies between the correlations of the GRM with other question types emphasizes the importance of the relationship between course design, instruction, and forms of assessment (Barnett & Francis, 2012; deBraga et al., 2015). GRM critical thinking skills must be developed, practiced, and reinforced in the instruction of the course. This may require an entire course redesign to develop critical thinking skills through student centered teaching techniques (see deBraga et al., 2015). Su et al., (2009) changed lecture-based classes to student centered case study discussions to facilitate their revisions in the multiple-choice questions to include plausible options. Similarly, Barnett & Francis (2012) suggest that instructional techniques to encourage and develop writing skills may play a part in the student's ability to express their ideas in open-ended questions.

In our study, in an end of course survey students did express an appreciation for the practice of GRM questions in the lab. They evaluated that portion of the lab highly for the discussion it generated around specific concepts. However, they evaluated the GRM questions on the quizzes and the exam negatively. This probably reflected the collaborative nature of the GRM questions in the lab to generate discussion and engagement in the material, which was not possible during test taking. The lack of collaboration for GRM questions on the tests meant that students had less confidence in their answers. Students also found the GRM questions to be time consuming in the stressful test setting, and preferred multiple-choice or short answer. The students did acknowledge an appreciation for the ability to achieve partial grades for a GRM question over the standard multiple-choice question. However, they felt that the first and last GRM options were easy to decipher but the second and third rankings were confusing and convoluted. This may reflect the quality of the GRM questions.

GRM questions require significantly more time to effectively compose in comparison to multiple-choice questions. This is especially true in courses where the course material is largely qualitative and open to interpretation. GRM seems to work best in courses with a range of finite quantitative solutions like statistics (see Matteucci & Stracqualursi, 2006) or biology (deBraga et al., 2015). In this anthropology course, there was much discourse and discussion around the order of the options amongst the teaching assistants. Differences in their level of knowledge for course material led to different interpretations on the ranking of GRM options. It was difficult to compose GRM questions from the perspective of the varying knowledge of the students, while taking into consideration their interpretation of course material.

It was expected that the highest achieving students would also have the highest scores in all question types. Student performance on the multiple-choice and short answer questions was a strong predictor (MC: $R^2 = .44$; short answer: $R^2 = 0.55$) of a student's final grade. There was a moderate correlation between GRM performance (GRM: $R^2 = 0.25$) and final grade. This most likely reflected question type familiarity as students overall found the GRM questions difficult to navigate at the beginning of the course. It was the first time that any of the students had written GRM assessments.

4.1 Limitations of Study

The small sample size (N = 53 out of 100 students registered in the course) may have resulted in a sample bias. The use of three question types in a single assessment may have altered the testing strategy of students in comparison to one question type per assessment (Hickson et al., 2012). There is no question that the assessment of critical thinking is diverse across the literature (Norris, 1989; Wilson-Mulnix, 2012). Identifying objective standards to test critical thinking is variable depending on whether critical thinking is considered to be a composite of generalized skills or contextual within a specific discipline (Behar-Horenstein & Niu, 2011; Tsui, 2002). It seems that there are universal skills in almost all definitions of critical thinking. However, some argue that the definition of critical thinking is "too vague to guide us in developing and judging critical thinking assessment" (Ennis, 1993, p. 179). A quantitative critical thinking assessment of the GRM (e.g. Watson-Glaser Critical Thinking test) in comparison to higher order MC and open-ended questions would provide insight into GRM as an effective critical thinking tool.

5. Conclusion

The GRM is a cost-effective alternative to multiple-choice testing where options are ranked according to their relative accuracy to answer a question. As we predicted, students who performed well on MC questions from higher Bloom's taxonomy levels (i.e. levels 4-6) and open-ended short answer questions generally performed well on GRM questions involving the same concepts. Also, high performing students on all three question types had better overall course grades. The null hypothesis that question type would not have an effect on critical thinking assessment was rejected. This suggests that GRM questions are comparable to higher order MC questions and supports GRM as a viable alternative to MC in critical thinking assessment. However, there are limitations to the effectiveness of GRM questions, particularly in subjects with qualitative course material. In addition, GRM questions require significantly more time to compose than standard multiple-choice testing.

While acknowledging that there are limitations, notably in the time required to conceive GRM style questions, the apparent challenge is mitigated by the GRM's strength in providing an additional opportunity for the test designer to verify or quantify the degree of difficulty of the test question. We suggest that the use of the GRM is an excellent method for ensuring greater test question complexity, which in turn will help to contribute to better test question design even when using traditional MC tests. We recommend that future research be undertaken to standardize the components of GRM questions in courses with quantitative course material.

Acknowledgements

We would like to acknowledge the University of Toronto Mississauga for supporting the pedagogical research that was the basis for this study.

References

Azer, S. (2003). Assessment in a problem-based learning course: Twelve tips for constructing multiple choice questions that test students' cognitive skills. *Biochemistry and Molecular Biology Education*, *31*(6), 428-434. http://doi.org/10.1002/bmb.2003.494031060288

Barnett, J., & Francis, A. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology, 32*(2), 201-211. http://doi.org/10.1080/01443410.2011.638619

Bassett, M. (2016). Teaching critical thinking without (much) writing: Multiple-choice and metacognition. *Teaching Theology & Religion, 19*(1), 20-40. http://dx.doi.org/10.1111/teth.12318

Behar-Horenstein L., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching and Learning, 8*(2), 25-41. https://doi.org/10.19030/tlc.v8i2.3554

Bloom, B. (ed.), Englehurst, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. White Plains, New York: Longman.

Daud, N., & Husin, Z. (2004). Developing critical thinking skills in computer-aided extended reading classes. *British Journal of Educational Technology, 35*(4), 477-487. http://dx.doi.org/10.1111/j.0007-1013.2004.00405.x

deBraga, M., Boyd, C., & Abdulnour, S. (2015). Using the principles of SoTL to redesign an advanced evolutionary biology course. *Teaching & Learning Inquiry, 3*(1), 15-29. http://dx.doi.org/10.2979/teachlearninqu.3.1.15

Ennis, R. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44-48.

Ennis, R. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, *18*, 4-10. https://doi.org/10.3102/0013189X018003004

Ennis, R. (1991). Critical thinking: A streamlined Conception. *Teaching Philosophy*, *14*(1), 5-24. http://dx.doi.org/10.5840/teachphil19911412

Ennis, R. (1993). Critical thinking assessment. *Theory into Practice*, 32(3), 179-186. http://dx.doi.org/10.1080/00405849309543594

Facione, P. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20*(1), 61-84. http://dx.doi.org/10.22329/il.v20i1.2254

Fink, L. (2003). *Creating significant learning experiences: An integrated approach to designing college courses.* San Francisco, California, Jossey-Bass.

Fliegel, R., & Holland, J. (2013). Quantifying learning in critical thinking. *The Journal of General Education*, *62*(2-3), 160-223. http://dx.doi.org/10.1353/jge.2013.0015

Fukuzawa, S., Goodman, J., & Jankowski, C. (2015). Building a sense of belonging: The impact of a university-wide exam prep initiative. *Research on Teaching and Learning conference*, McMaster University, Hamilton, Ontario, December 9, 2015.

Gellin, A. (2003). The effect of undergraduate student involvement on critical thinking. *Journal of College Student Development*, *44*(6), 746-762. http://dx.doi.org/10.1353/csd.2003.0066

Halonen, J. (1995). Demystifying critical thinking. *Teaching of Psychology*, *22*(1), 75-81. http://dx.doi.org/10.1207/s15328023top2201_23

Halpern, D. (2001). Why wisdom? *Educational Psychologist*, *36*(4), 253-256. http://dx.doi.org/10.1207/S15326985EP3604_4

Halpern, D. (2003). The how and why of critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning: Current research, theory and practice.* Cresskill, NJ: Hampton Press.

Hickson, S., Reed, W., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment*, *17*, 200-213. http://dx.doi.org/10.1080/10627197.2012.735915

Holmes, N., Wieman, C., & Bonn, D. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences*, *112*(36), 11199-11204. http://dx.doi.org/10.1073/pnas.1505329112

Hyyttinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Yianne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, *44*, 108. http://doi.org/10.1016/j.stuedus.2014.11.001

Kerkman, D., & Johnson, A. (2014). Challenging multiple-choice questions to engage critical thinking. *InSight: A Journal of Scholarly Teaching*, *9*, 92-97.

Kibble, J., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education*, *35*, 396-401. http://doi.org/10.1152/advan.00062.2011

Kim, M., Patel, R., Uchizono, J., & Beck, L. (2012). Incorporation of bloom's taxonomy into multiple-choice questions for a pharmacotherapeutics course. *American Journal of Pharmaceutical Education*, *76*(6), 114. http://dx.doi.org/10.5688/ajpe766114

Krathwohl, D. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218. http://doi.org/10.1207/s15430421tip4104_2

Ku, K. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*, 70-76. http://doi.org/10.1016/j.tsc.2009.02.001

Loes, C., Salisbury, M., & Pascarella, E. (2015). Student perception of effective instruction and the development of critical thinking: A replication and extension. *Higher Education*, *69*, 823-838. http://dx.doi.org/10.1007/s10734-014-9807-0

Mahapoonyanont, N., Krahomwong, R., Kochakornjarupong, D., & Rachasong, W. (2010). Critical thinking abilities assessment tools: reliability generalization. *Procedia Social and Behavioral Sciences*, *2*, 434-438. http://doi.org/10.1016/j.sbspro.2010.03.038

Matteucci, M., & Stracqualursi, L. (2006). Student assessment via graded response model. *Statistica*, *LXVI*(4), 435-447. http://doi.org/10.6092/issn.1973-2201/1216

Morrison, S., & Free, K. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, *40*(1), 17-24.

Naastrom, G. (2009). Interpretation of standards with Bloom's revised taxonomy: A comparison of teachers and assessment experts. *International Journal of Research & Method in Education*, *32*(1), 39-51. http://doi.org/10.1080/17-437270902749262

Nevid, J., & McClelland, N. (2013). Using action verbs as learning outcomes: Applying Bloom's taxonomy in measuring instructional objectives in introductory psychology. *Journal of Education and Training Studies*, *1*(2), 19-24. http://dx.doi.org/10.11114/jets.v1i2.94

Norris, S. (1989). Can we test validly for critical thinking? *Educational Researcher, 18*(9), 21-26. http://doi.org/10.2307/1176715

Norris, S. (2003). The meaning of critical thinking test performance: The effects of abilities and dispositions on scores. In: D. Fasko (Ed.), *Critical thinking and reasoning: Current research, theory, and practice*. Cresskill, NJ: Hampton Press.

Parmenter, D. (2009). Essay versus multiple-choice: Student preferences and the underlying rationale with implications for test construction. *Academy of Entrepreneurship Journal*, *15*(2), 57-71.

Poondej, C., & Lerdpornkulrat, T. (2015). The reliability and construct validity of the critical thinking disposition scale. *Journal of Psychological and Educational Research*, *XXIII*(1), 23-36.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17. https://doi.org/10.1007/BF03372160

Seaman, M. (2011). Bloom’s taxonomy: Its evolution, revision, and use in the field of education. *Curriculum & Teaching Dialogue*, *13*, 29-43.

Sommer, R., & Sommer, B. (2009). The dreaded essay exam. *Teaching of Psychology*, *36*(3), 197-198. http://dx.doi.org/10.1080/00986280902959820

Stanger-Hall, K. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *Life Sciences Education*, *11*(3), 294-306. http://dx.doi.org/10.1187/cbe.11-11-0100

Stupnisky, R., Renaud, R., Daniels, L., Haynes, T., & Perry, R. (2008). The interrelation of first-year college students’ critical thinking disposition, perceived academic control, and academic achievement. *Research in Higher Education*, *49*(6), 513-530. http://dx.doi.org/10.1007/s11162-008-9093-8

Su, M., Osisek, P., & Starnes, B (2005). Using the revised Bloom’s taxonomy in the clinical laboratory: Thinking skills involved in diagnostic reasoning. *Nurse Educator, 30*, 117-122. http://dx.doi.org/10.1097%2F00006223-200505000-00014

Su, M., Osisek, P., Montgomery, C., & Pellar, S. (2009). Designing multiple-choice test items at higher cognitive levels. *Nurse Educator*, *34*(5), 223-227. http://dx.doi.org/10.1097/NNE.0b013e3181b2b546

Taube, K. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *Journal of General Education*, *46*, 129-164.

Tiemeier, A., Stacy, A., & Burke, J. (2011). Using multiple choice questions written at various Bloom’s taxonomy levels to evaluate student performance across a therapeutics sequence. *Innovations in Pharmacy*, *2*, 1-11. https://doi.org/10.24926/iip.v2i2.224

Toppino, T., & Cohen, M. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology*, *56*(4), 252-257. https://doi.org/10.1027/1618-3169.56.4.252

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *Journal of Higher Education*, *73*(6), 740-763. https://doi.org/10.1353/jhe.2002.0056

Veltre, M., Cho, K., & Neely, J. (2015). Transfer-appropriate processing in the testing effect. *Memory, 8*(2), 1229-1237. https://doi.org/10.1080/09658211.2014.970196

Walstad, W., & Becker, W. (1994). Achievement differences on multiple-choice and essay tests in economics. *The American Economic Review, 8*4(2), 193-196.

Watson, G., & Glaser, E. (2008). *Watson-Glaser critical thinking appraisal short form manual*. Upper Saddle River, NJ: Pearson Education.

Wilson-Mulnix, J. (2012). Thinking critically about critical thinking. *Educational Philosophy and Theory*, *44*(5), 464-479. http://dx.doi.org/j.1469-5812.2010.00673.x

Yee, T. W. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer. https://doi.org/10.1007/978-1-4939-2818-7

Yee, T. W., & Hastie, T. J. (2003) Reduced-rank vector generalized linear models. *Statistical Modelling*, *3*, 15-41. https://doi.org/10.1016/j.csda.2013.01.012

Zheng, A., Lawhorn, J., Lumley, T., & Freeman, S. (2008). Application of Bloom’s taxonomy debunks the “MCAT myth”. *Science*, *319*, 414-415. http://dx.doi.org/10.1126%2Fscience.1147852

Notes

Note. This is an example of a multiple-choice, GRM and open-ended question covering the same course material from the final exam. The answer is in bold for the multiple-choice and the GRM questions.

You are a paleoanthropologist working in North America. You have discovered a skull that dates to the Paleocene. It has been identified as *Carpolestes*. What is the first step for an assessment of the evolutionary significance? (Level 6 Bloom's taxonomy)

a. *Carpolestes* provides evidence that Omomyidae originated in North America. I would look for evidence of immobile eye sockets similar to tarsiers.

b. *Carpolestes* provides evidence that Adapidae in North America was the common ancestors of all prosimians. I would examine the specimen for a 2.1.2.3 dental formula.

c. *Carpolestes* provides evidence that Propliopithecidae originated in Asia. I would look for a prehensile tail because it is only found in primates of North America and Asia.

d. *Carpolestes* provides evidence that Plesiadapis in North America may be the oldest arboreal fossil. I would look for evidence of a developed clavicle demonstrating brachiation.

e. ***Carpolestes* provides evidence of Plesiadapis in North America. I would look for traits that are shared by all extant primates.**

Graded Response Method: Rank the following options from the most correct to the least correct. Remember that each option is a separate question on the scantron.

An unknown fossil fragment that includes a face with a complete postorbital bar has recently been found in North America. It has been dated at 55 million years old. You are a palaeoanthropologist called to the site where the specimen was discovered. What is the possible evolutionary significance of this specimen? What else would you investigate about the specimen?

a. This specimen is significant because the complete postorbital bar suggests that the specimen belongs to the Order Primate.

b. This specimen is significant because the complete postorbital bar suggests that the specimen may be a common ancestor for prosimians. I would want to know if the eye socket is completely enclosed by bone.

c. This is significant because the date and primate traits of the specimen suggest that this specimen may be one of the earliest definitive primates. I would want to know if it is a Plesiadapiforme.

d. This specimen is significant because the location of the primate is North America and no living primates are found in North America

Answer: C, A, B, D

Open-ended Question:

Explain the debate concerning whether or not Plesiadapidae is a primate. What is the evolutionary significance of this debate? (4 marks)