Challenging Multiple-Choice Questions to Engage Critical Thinking

Dennis D. Kerkman, PhD
Professor of Psychology
Park University

Andrew T. Johnson, PhD
Professor of Psychology
Park University

This article examines a technique for engaging critical thinking on multiple-choice exams. University students were encouraged to "challenge" the validity of any exam question they believed to be unfair (e.g., more than one equally correct answer, ambiguous wording, etc.). The number of valid challenges a student wrote was a better predictor of exam scores than the number of invalid challenges or GPA. The technique also allows instructors to gain insight into the sources of students' errors that may be useful in improving instruction.

So-called "multiple-guess" exams have a rather bad reputation in certain quarters of academe. Despite their numerous detractors in the USA (e.g., Gould, 1996; Sacks, 2000) and abroad (Last, 2006), multiple-choice tests are frequently employed by those who teach large classes, primarily because grading large numbers of essay exams is prohibitively time-consuming. However, many professors look askance at this assessment technique, asserting that multiple-choice questions tend to focus on rote memory rather than comprehension of the subject matter or thinking critically about it. Here, we present evidence that encouraging students to "challenge the question" is associated with higher scores on exams and can provide instructors with new information to improve instruction.

…encourages students to 'challenge the question' is associated with higher scores on exams and can provide instructors with new information to improve instruction.

It is widely agreed that critical thinking is one of the foremost goals of higher education, but there is less agreement on what critical thinking is, or what constitutes evidence of it. Jones and his colleagues (Jones, Dougherty, Fantaske, & Hoffman, 1997; Jones et al., 1995) found a consensus among 500 educators, policy makers, and employers concerning the definition of critical thinking: "…critical thinking describes reasoning in an open-ended manner, with an unlimited number of solutions. It involves constructing a situation and supporting the reasoning that went into a conclusion" (Halpern, 2001, p. 254). In defining critical thinking, Halpern (2001) notes that "When we think critically, we are evaluating the outcomes of our thought processes—how good a decision is or how well a problem is solved" (p. 254).

The inherently forced-choice nature of multiple-choice questions would seem to preclude them as indicators of or occasions for critical thinking. However, when students answer multiple-choice questions they do evaluate multiple response options in order to decide how well each solves the problem. Thus, multiple-choice questions would appear to satisfy at least that part of the definition of critical thinking. If multiple-choice questions also posed the opportunity for open-ended responding, such as arguing that the question itself is inherently flawed, or that none of the options is superior to the others, then multiple-choice questions might provide clearer evidence of critical thinking.

Previous research identifies relationships between critical thinking and multiple-choice test performance. Several sources indicate that multiple-choice test items involve critical thinking processes (Appleby, 1990; Scialfa, Legare, Wenger, & Dingley, 2001; Williams & Clark, 2004; Yoder & Hochevar, 2005). Wallace and

Williams (2003) and Williams, Oliver, Allin, Winn, and Booher (2003) found that scores on a standardized measure of critical thinking correlated positively and significantly with college students' performance on multiple-choice tests. However, we know of no prior research that has attempted to engage critical thinking on multiple-choice exams by encouraging students to challenge the validity of the exam questions themselves.

Initially, we developed the "Question Challenge" policy as a way to deal with students' complaints about the difficulty of our multiple-choice exam questions. In an effort to avoid "sour grapes" complaints after the students had received their test scores, we informed students that if they thought a question was unfair, they could write a challenge to the question before they knew whether or not they got it right. If their challenge convinced the instructor that the question was unfair, then everyone in the class would receive an extra point, and this extra point would be named in honor of the successful challenger, as a form of social reinforcement for thinking critically about questions.

The challenge policy has been well-received by students. Furthermore, students who wrote challenges, even challenges that were not accepted, seemed to be getting better scores on the exams. In addition, the unsuccessful challenges provided the instructor with useful insights into the student's thought processes. It appeared that the challenge policy induced a critical thinking mental set that encouraged students to analyze the question, rather than just reacting to it on the basis of rote association. If this is true, then students who write challenges should score higher on the exam than those who do not. Further, those whose critical thinking is more accurate, as indicated by writing challenges that were valid, should score even higher than those whose challenges were judged to be invalid.

Initially, we developed the 'Question Challenge' policy as a way to deal with students' complaints about the difficulty of our multiple-choice exam questions.

Here, we report a study designed to test three hypotheses: (a) the number of valid challenges that students wrote would be correlated with their scores on multiple-choice exams, (b) the number of invalid challenges would also be correlated with multiple-choice exam scores, and (c) the number of valid challenges would be a significantly better predictor of exam scores than the number of invalid challenges. An obvious "third variable explanation" would be that general scholastic ability, rather than a specific critical thinking mental set, could be the source of higher scores on multiple-choice exams and writing more valid challenges. To address that issue, we compared the relative predictive power of students' GPAs versus their valid and invalid challenges.

Method

Participants

Complete data were available for 10 male and 111 female undergraduate students who were enrolled in a social psychology class. All were native speakers of American English. Mean age was 21.62 (SD = 1.80). The average number of years of college completed was 3.0 (SD = .63).

Procedure

Each exam contained approximately 25 items based on the text (Taylor, Peplau, & Sears, 2005) and 25 from lectures, videos, Internet exercises, and discussions. All questions were composed by the first author of this study (First Author). Exams covered three or four chapters of the text and occurred approximately every four weeks. On the first page of each exam, the following instructions appeared:

CHALLENGING QUESTIONS: If you think that two or more of the answers are equally correct, or that none of the answers is really correct, or that the question is ambiguously worded, then write the word "CHALLENGE" and the question number at the top of THIS PAGE (PAGE 1). Then, on the back of the last page, write the reason for your challenge. You must explain your challenge well enough that I am convinced that you are correct. For example, "Challenge Question 4 – ambiguous." is not sufficient. However, "Challenge Question 4 – according to the book, option B 'modeling', and option C 'imitation' both mean the same thing, learning by watching someone else." is good enough. Remember, if you challenge a question successfully, then EVERYONE who takes the exam gets an extra point added to their score, and I will name that extra point in your honor. ALWAYS ANSWER EVERY QUESTION, EVEN IF YOU PLAN TO CHALLENGE IT. CHOOSE THE ANSWER THAT YOU THINK I THINK IS CORRECT.

Two content-area experts (the authors) examined the validity of the challenges. A challenge was judged as valid when both experts accepted the argument. For example, one question read:

Which of the following characteristics makes it easier to change a person's attitude?

A. internal locus of control.
B. high-self-esteem.
C. being highly authoritarian.
D. being relaxed.

A student wrote the following challenge: "Are you talking about the personality characteristics of the speaker or the listener? The question is ambiguous." Much to the instructor's chagrin, this question turns out to be a classic example of "deep structure ambiguity" (Chomsky, 1957), so the challenge was judged to be valid.

In contrast, consider the following question:

To voluntarily help someone without expecting anything in return is called
A. the norm of social responsibility
B. empathy
C. altruism
D. prosocial behaviors

A student challenged this question, stating "Altruism and prosocial behavior are the same thing. Both are about helping someone without expecting a reward." In this case, the student's challenge is invalid, because altruism is the specific subcategory within the more general category of prosocial behavior in which no reward is expected. Interestingly, this invalid challenge reveals the source of the student's confusion that can be useful for instructing future classes. It reflects a failure to understand how the logic of class inclusion (Inhelder & Piaget, 1964) applies to the relation between these two abstract concepts: All altruistic behaviors are prosocial, but not all prosocial behaviors are altruistic.

Results

Each student's correct answers, valid, and invalid challenges for exams 3 and 4 were summed. Means (and SD's) for total exam scores, number of valid challenges, number of invalid challenges, and GPA (as of the beginning of the semester in question) were 81.29 (10.40), .67 (1.02), .86 (1.01), and 3.15 (.60), respectively.

GPA correlated significantly with exam scores, $r(19) = .44$, $p = .022$, but not with the number of valid or invalid challenges. The number of valid challenges was not significantly correlated with the number of invalid challenges, but was significantly correlated with exam scores, $r(19) = .65$, $p = .001$. The number of invalid challenges that students wrote was also significantly correlated with their exam scores, $r(19) = .38$, $p = .045$.

Stepwise multiple regression including GPA, number of valid challenges, and number of invalid challenges as predictors showed that the most efficient model for predicting exam scores involved a single predictor: the number of valid challenges, $R2 = .43$, $F(1, 19) = 14.09$, $p = .001$. Even when GPA was forced to enter the equation first, the number of valid challenges that a student wrote still accounted for a significant increase in exam score variance, part-$r = .54$, $t(18) = 3.20$, $p = .005$.

Discussion

The results support all three hypotheses. Writing challenges to multiple-choice questions predicted performance on multiple-choice exams, even if the challenge was invalid. However, the number of valid challenges was a better predictor of exam scores than the number of invalid challenges. It accounted for nearly half of the variance in exam scores and predicting exam scores over and above general academic skill, as measured by the student's GPA at the beginning of the semester.

For valid challenges to account for nearly half of the variance in exam scores suggests that critical thinking can play a role on multiple-choice exams when students are explicitly asked to challenge and critique the questions. Encouraging students to challenge exam questions engages their critical thinking processes and opens a window for the instructor to view what is going on in students' minds when they take multiple-choice exams. For example, we now use information from valid challenges to clarify potential confusions during lectures, such as the aforementioned class inclusion relationship between prosocial behavior and altruism. As a result of numerous invalid challenges from students, one of us (Kerkman) now provides examples at the outset of each exam to clarify what he means by "choose the best answer," (e.g., "Rome is in (a) the universe, (b) Europe, (c) Italy, (d) the Coliseum." The best answer is (b), Italy, because it is the most specific answer without being too specific. A small area of Rome lies within the Coliseum, but this answer is too specific, because there is a great deal of Rome that is not in the Coliseum). The Challenge technique also provides the instructor with a way to refine the item-pool from one semester to the next by eliminating flawed questions that have been successfully challenged by students.

Encouraging students to challenge exam questions engages their critical thinking processes and opens a window for the instructor to view what is going on in students' minds when they take multiple-choice exams.

Further, the challenge technique makes it clear to the students that the classroom setting is more democratic and less authoritarian than some may initially perceive it to be. This is particularly true of students from other cultures (e.g., most Amerindian and many Asian cultures), where criticizing the instructor is considered to be extremely disrespectful. When the instructor explicitly encourages and rewards students for challenging the instructor's questions, the students know that they have an opportunity for input and a right to "to petition the government for a redress of grievances" (U.S. Const. amend. I). Thus, the challenge technique described here serves to promote the free and frank exchange of ideas in the classroom setting that is essential to critical thinking in all its forms and has formed the very foundation of the academic enterprise since the time of Socrates.

In conclusion, while there are several strategies that shift test takers to a critical thinking mindset, we propose that there is an additional one involving challenging the actual questions. We argue that there is a deeper processing of questions and responses. The first process is question comprehension and selection of the correct response. The second process involves evaluating the questions for quality. This secondary process is one of the highest levels of Bloom's Taxonomy and demands critical thinking processing. Furthermore, there is a third process of

creating the challenge – creation is the highest level of the Revised Bloom's Taxonomy (Krathwohl, 2002).

While we have found evidence for our claim, it is not yet determined that there may be alternative strategies that may directly activate a "creation" mindset and have the same performance effects. Future research could examine creation strategies and their effects.

It is noteworthy that the challenge technique is quite general and can be readily applied in virtually any content area. In principal, we see no reason why is should be restricted to the multiple-choice format. Matching, short answer, or even essay questions can, and we believe should, be open to criticism by those whose performance evaluations are based on them.

References

Appleby, D. (1990). A cognitive taxonomy of multiple-choice questions. In Makosky, Sileo, Whittemore, Linda, Landry, Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 79-82). Washington, DC: American Psychological Association.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Gould, S. J. (1996). *The mismeasure of man*. New York, NY: W. W. Norton & Company.

Halpern, D. F. (2001). Why wisdom? *Educational Psychologist*, *36*(4), 253-256.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York, NY: Harper & Row.

Jones, E. A., Dougherty, B. C., Fantaske, P., & Hoffman, S. (1997). *Identifying college graduates' essential skills in reading and problem-solving: Perspectives of faculty, employers and policymakers* (Contract No. R117G10037/CDFA84.G, 117). Washington, DC: U.S. Department of Education/OERI.

Jones, E. A., Hoffman, S., Moore, L. M., Ratcliff, G., Tibbetts, S., & Click, B. A. (1995). *National assessment of college student learning: Identifying college graduates' essential skills in writing, speech and listening, and critical thinking* (NCES 95-001). Washington, DC: U.S. Government Printing Office.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, *41*, 212-218.

Last, J. (2006, June). The dumbed-down GCSE exam with no need for writing. *Times Educational Supplement*; June 12, 2006, Issue 4689. Retrieved from http://www.tes.co.uk/search/story/?stor y_id=2245901

Sacks, P. (2000). *Standardized minds: The high price of America's testing culture and what we can do about it*. Cambridge, MA: Perseus Books.

Scialfa, C., Legare, C., Wenger, L., & Dingley, L. (2001). Difficulty and discriminability of introductory psychology test items. *Teaching of Psychology*, *28*(1), 11-15.

Taylor, S. E., Peplau, L. A., & Sears, D. O. (2005). *Social Psychology* (12th ed.). Upper Saddle River, NJ: Pearson Education Inc.

U.S Const. amend. I

Wallace, M. A., & Williams, R. L. (2003). Multiple-choice exams: Explanations for student choices. *Teaching of Psychology*, *30*(2), 136-138.

Williams, R., & Clark, L. (2004). College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice exams. *Educational Research*, *46*(3), 231-239.

Williams, R. L., Oliver, R., Allin, J. L., Winn, B., & Booher, C. S. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology*, *30*(3), 220-223.

Yoder, J., & Hochevar, C. (2005). Encouraging active learning can improve students' performance on examinations. *Teaching of Psychology*, *32*(2), 91-95.

Footnotes

[1] In keeping with standard statistical practice, one female student was deleted from the analyses because her number of challenges was more than three standard deviations above the group mean, and therefore was deemed to be a statistical outlier.

Dennis D. Kerkman received his Bachelor's degree in Psychology from the University of Kansas, his Master's degree in Psychology from the University of Georgia, and PhD in Developmental & Child Psychology from the University of Kansas. His research interests include cognitive development of problem solving skills in mathematics and science, and cultural differences. Before joining Park University in 2003, he served as a postdoctoral research associate at Carnegie Mellon University, a professor at Texas State University, and as an educational research consultant to the Mexican government. He enjoys fishing and traveling.

Andrew Johnson is Professor of Psychology at Park University. He received his PhD from Kansas State University in Experimental Psychology with an emphasis in Cognition. His research interests are in: metacognition, metapedagogical processes, figurative language processing, cognitive maps, scholarship of teaching, assessment and rubric creation, and science and pseudoscience. He is an active in the Society for the Teaching of Psychology and serves as a consultant-evaluator for Educational Testing Services for assessment of the Advanced Placement Psychology exams and development of CLEP Psychology exams and materials.