1-2019

Which of the Following Is True: We Can Write Better Multiple Choice Questions

Emily Rauschert
Cleveland State University, e.rauschert@csuohio.edu

Suann Yang
State University of New York College at Geneseo

Rachel M. Pigg
Presbyterian College

Recommended Citation

Rauschert, Emily; Yang, Suann; and Pigg, Rachel M., "Which of the Following Is True: We Can Write Better Multiple Choice Questions" (2019). *Biological, Geological, and Environmental Faculty Publications*. 247.
https://engagedscholarship.csuohio.edu/scibges_facpub/247

ECO 101

Which of the Following Is True: We Can Write Better Multiple Choice Questions

Emily S. J. Rauschert[1], Suann Yang[2] and Rachel M. Pigg[3]
[1]Cleveland State University, 2121 Euclid Ave., Cleveland, Ohio 44115 USA
[2]The State University of New York College at Geneseo, 1 College Circle, Geneseo, New York 14454 USA
[3]Presbyterian College, 503 South Broad Street, Clinton, South Carolina 29325 USA

Summary and Comments from Workshop 49: "Which of the Following Is True: We Can Write Better Multiple Choice Questions" presented at the 102nd Ecological Society of America Meeting.

Introduction

Multiple choice questions (MCQs) are commonly used as a way to assess student learning, especially in (but not restricted to) very large classes. Nearly all students are exposed to multiple choice testing during K-12 education as well as in standardized tests such as the SAT and the GRE.

There are many benefits to using MCQs. First of all, they provide an objective way to grade, compared to essay questions. In addition, they are quicker to grade, particularly when class sizes are large. Within a single test, we typically can ask more multiple choice questions across more topics (i.e., better content domain sampling; Haladyna 1997) compared to other question types. Further, we can consistently score the same multiple choice questions across years to examine whether changes in curriculum are having a positive effect on student learning.

On the other hand, it can be difficult to write large numbers of novel MCQs. Also, instructors may have concerns about their limitations in assessing all the skills we would like students to acquire. Common criticisms include that MCQs mainly test rote memorization, not higher order skills. Finally, instructors worry that students will get answers correct merely because they guessed the right answer and not because they have mastered the material (Stanger-Hall 2012).

Rauschert, E. S. J., S. Yang, and R. M. Pigg. 2019. Which of the Following Is True: We Can Write Better Multiple Choice Questions. Bull Ecol Soc Am 100(1):e01468. https://doi.org/10.1002/bes2.1468

To help address these concerns, we can use our scientific training and evidence about how people learn to design better MCQs. Turning our scientific attention to the practice of how we teach is at the heart of what has been termed “scientific teaching” (Handelsman et al. 2004). Writing effective multiple choice tests has been studied by a number of researchers, and there are many evidenced-based practices in multiple choice testing that would be helpful for ecology faculty to know. Fortunately, even short training sessions, such as one-hour workshops, can improve faculty members’ abilities to write effective multiple choice questions (Dellinges and Curtis 2017). To help connect ecological educators with best practices in MCQs, we held a workshop at the ESA meeting in Portland, Oregon, in August 2017. We report here some of our main findings from the literature on MCQs, combined with insights from the workshop participants, to provide ideas about how to teach ecology using MCQs more effectively.

Designing MCQs

Alignment of multiple choice questions to their corresponding learning outcomes, or what we want students to learn, is an important step toward accurately measuring the effectiveness of our teaching (Wiggins and McTighe 1998). Designing learning outcomes at the outset is an important tenet of scientific teaching (Dirks et al. 2014). By setting specific instructional goals, we can guide our students toward these goals with assignments prior to an examination (called formative assessments). Then, we assess whether or not students have reached these goals with an examination (called summative assessments). To align our assessments and communicate to our students what they should be able to do in the examinations, learning outcomes should convey both the concept and cognitive skill needed by the student.

Several frameworks exist for classifying the cognitive skills specified in learning outcomes. Bloom’s Taxonomy of Educational Objectives is recommended for use in biology (Bloom et al. 1956, Anderson and Krathwohl 2001, Crowe et al. 2008). Revised in 2001, Bloom’s Taxonomy utilizes two dimensions: Knowledge and Cognitive Process (Anderson and Krathwohl 2001). The Knowledge dimension consists of factual (terminology and specific details), conceptual (interrelationships among elements of factual knowledge), procedural (skills and techniques), and metacognitive knowledge (self-knowledge). The categories of the Cognitive Process dimension vary in cognitive complexity, from simple to most complex: remember, understand/comprehend, apply, analyze, evaluate, and create (Anderson and Krathwohl 2001). Categories at the complex end of the scale can require students to achieve some of the simpler cognitive processes. For example, in order to identify the biome represented in a climate diagram (understand/comprehend), a student would need to be able to describe the characteristics of various biomes (remember). The Biology in Bloom Tool (BBT; Crowe et al. 2008) provides examples of how different biological skills and concepts are classified within the Cognitive Process dimension.

As a method of assessment, MCQs can impact student learning gains when designed carefully and used to gauge student learning prior to an examination (Black and Wiliam 1998). Although MCQs are frequently used to evaluate cognitive skills such as recall or comprehension, they can also be employed to measure higher order cognitive skills such as application, analysis, and evaluation (e.g., Palmer and Devitt 2007, reviewed by Haladyna 1997). To design MCQs that test a range of cognitive skills, it is useful to understand the basic structure of good MCQs.

The anatomy of a good multiple choice question

The basic structure of a multiple choice question is a stem followed by several choices (Fig. 1). The stem is usually a question or a phrase with some words missing. At least one of the choices is the correct answer, and the incorrect choices are termed distractors. A good stem includes the central idea of the question and most of the text of the question. Stems should be as concise as possible so that students can read and understand them quickly. Although this is common, negatively stated stems (e.g., “All of the following EXCEPT”) should be avoided if possible (Haladyna 2004). To reduce cognitive load, it is also better to write a stem that is a question rather than having missing words and a blank in the middle of a sentence (Haladyna 1997). Grammatical ambiguities, such as problems with a/an or is/are, should be avoided.

Good choices are generally short and all of similar length. They should have parallel grammatical and content structure as the correct choice (Haladyna 1997). They should be independent of each other; choices like “A and B are true” or “All/none of the above” are generally not effective (Haladyna 1997). Like for stems, positive phrasing in choices is clearer than negative phrasing. Lastly, good choices should be plausible answers and can contain typical errors that students make.

Many faculty struggle to come up with numerous plausible distractors. Sometimes asking a version of a question as an open-ended question can generate several answers that students would find plausible. Even though many faculty feel compelled to have 4–5 options, three choices (one correct, two distractors) is optimal (Owen and Froman 1987, Haladyna and Downing 1993, Rodriguez 2005). In practice, even if we create five options, students will usually easily rule out one or more and only select among the three most plausible. Adding more distractors can make it more difficult for students who read slowly without actually allowing better differentiation between students’ understanding of the content (Rodriguez 2005).

Testing both lower and higher cognitive skills with MCQs

Although we may be sure of the learning outcome that is to be assessed, constructing the corresponding multiple choice question can be a challenge, especially for more complex cognitive processes. We recommend preparing students for difficult summative (examination) MCQs with formative (practice) MCQs at the same cognitive level, because students who are accustomed to being tested on only lower

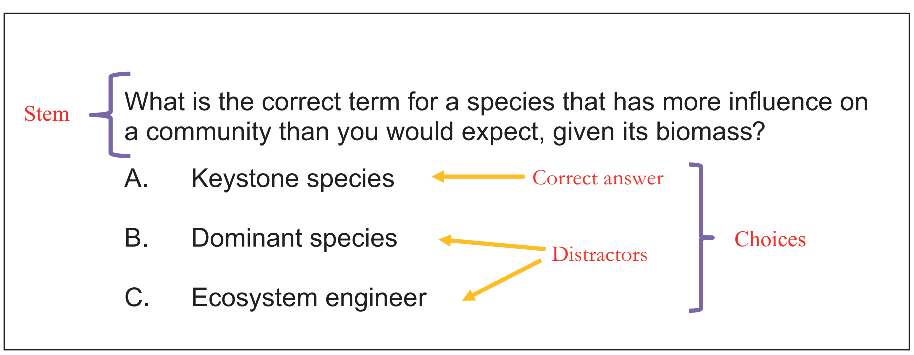

Fig. 1. The anatomy of a multiple choice question: The stem is the question or base phrase, and among the possible choices, there is at least one correct answer as well as several distractors, or incorrect choices.

Table 1. Item shells for different cognitive processes. Adapted from Haladyna (1997, 2004) to correspond with Crowe et al. (2008).

Cognitive process	Item shell	Example
Lower order cognitive skills		
Remember	• What is the best definition of __? • What is the best word to use for _?	What is the best definition of net primary productivity?
Understand/Comprehend	• What is an example of __? • What is the difference between __ and __?	What is an example of density-dependent regulation of population growth?
Higher order cognitive skills		
Apply	• Predict what happens when __. • What is a consequence of __? • Given __, what is the cause of __? • Solve a word problem or apply a formula in a new situation.	Predict what happens to production efficiency of polar bears when a lack of sea ice increases energy expenditure
Analyze	• To test the hypothesis __, a scientist would need to __. • What is the best conclusion for what is shown in this graph of __?	What would be the best way to test the effect of nitrogen on a phyto-plankton species in a local pond?
Evaluate	• What is the most effective/ appropriate for __? • What is the best [solution] for __?	An experiment finds that metapopulation dynamics of a butterfly is absent when habitat loss has occurred. What is the best explanation for this result?

cognitive skills with MCQs may under-prepare for MCQs that test higher cognitive skills (Scouller 1998, Martinez 1999, Stanger-Hall 2012). This means we need to write even more questions, adding to the challenge of composing questions for complex cognitive processes. Item shells, which show the structure of a question without the outcome-specific details, are a useful starting point (Table 1; Haladyna and Shindoll 1989, Haladyna 1997, see also Crowe et al. 2008). We suggest analyzing previous examinations and question structures in test banks to identify additional possibilities for item shells.

Going further with MCQs

There are several ways to expand the use of multiple choice questions on tests. One way is through pairing an open-ended question to an MCQ, such as with two-tiered diagnostic instruments (Treagust 1988). Originally developed for identifying patterns in student misconceptions, a two-tiered diagnostic instrument quantifies student content knowledge in a first step using an MCQ, while the second step reveals the depth of knowledge by asking students to explain their choice. For example, Diagnostic Question Clusters (DQCs) are concept inventories that utilize this two-tiered question structure, along with other types of questions, to identify student misconceptions in ecological topics (D'Avanzo et al. 2012, Hartley et al. 2012). A simple way to use a two-tiered approach for any topic would be to ask students to defend their answer after an MCQ, to help determine whether they understand the concept or are simply guessing.

Another way to expand our use of MCQs is to develop scenarios-based question sets, which are a series of multiple choice questions that are associated with a single proceeding scenario (or short story).

The scenario describes a problem, relationship, and/or process that provides the context needed for students to answer the questions that follow. This framework is similar to case studies (Herreid 1994) or DQCs (D'Avanzo et al. 2012), in that students are being asked to apply their knowledge from multiple areas (or learning outcomes) to the specific situation or problem described in the scenario.

Using scenario-based questions in examinations has many advantages. A single scenario can replace multiple, complex question prompts, which reduces the cognitive load students experience during an examination (Haladyna 2004). Questions that test higher order cognitive skills are already demanding, and grouping questions around a common scenario reduces the amount of processing students need to do compared to if they need to transition between multiple problem contexts. Instructors can also distribute the scenarios (without their associated questions) to students before the examination, which gives students more time during the examination to read questions carefully and consider their answers. In this way, students can demonstrate their ability to solve the problem during the examination rather than using test-taking time for reading comprehension. Further, the scenarios themselves provide students with a great study tool to practice applying concepts and learning objectives to new situations before their summative assessment. Scenarios also give the instructor an opportunity to introduce students to newly published research (and/or the instructor's own research) by transforming that research into a scenario. For example, recently published work by Klein et al. (2016) describes how mycorrhizal fungi can transfer carbon among individual trees. This research could be transformed into a scenario, in which students could answer questions regarding photosynthesis, translocation of sugars in plants, interspecific interaction types, and much more.

Interpreting test statistics to validate or improve MCQs

To determine whether MCQs have successfully tested learning outcomes, we can interpret test statistics. Campus grading services (or the scanning software of a test scoring machine) can usually provide reports that summarize test statistics and individual item (question) analyses. These reports are helpful to evaluate student learning, the quality of a particular test item, and the quality of the entire test. Typical metrics from such a report include mean and standard deviation per test and/or per item, item difficulty, point-biserial correlation, and Cronbach's alpha and/or KR 20 (Fig. 2; Kehoe 1995). Individual item analyses are especially useful when questions are designed with distractors that encompass specific misconceptions in ecology (e.g., D'Avanzo et al. 2012, reviewed in Munson 1994, see also Summers et al. 2018). Item analyses can also be used to identify questions with errors in phrasing. How useful test statistics and item analyses are for pinpointing improvements does depend on how close the alignment is among learning outcomes, class activities, and test questions.

Conclusion

Our workshop was attended by 27 instructors at different stages in their careers and from a range of institutions: community colleges to research universities, private to public, and small to large. A survey at the end of the workshop revealed that participants were enthusiastic about expanding their repertoire of multiple choice questions and discovering new strategies for using them. Our short workshop shows that participants can revitalize their interest in MCQs by examining ways to use them effectively and appropriately.

Mean and standard deviation: At the test level, arithmetic mean of student scores for the test. At the individual item level, the arithmetic mean of the number of points earned by each students on that item. Standard deviations about each of these can be calculated as well to show the extent of variability in performance.

Item difficulty: Proportion or percentage of students answering a question correctly; smaller values result from harder questions (or incorrect keying or ambiguous question)

Point-biserial correlation: Measures how well the question discriminates between students with higher and lower levels of knowledge by quantifying the relationship between scores on the question and total test scores; positive if higher performers answered the question correctly more frequently than lower performers did, and negative if the opposite occurred. Ranges from –1.0 to 1.0.

Cronbach's alpha, KR 20: These two metrics measure (in different ways) the internal consistency of the test, or the consistency of students' responses across the items on the test; students performing well (or poorly) should do so in a consistent fashion. Values closer to 1 show greater consistency.

Fig. 2. Test statistics and item analysis for MCQs.

Feedback from participants suggested our workshop was useful, especially reiterating the importance of aligning questions to objectives, identifying the attributes of good and bad MCQs, and how to interpret test statistics to make our MCQs better. Feedback also revealed that participants were intrigued by the possibility of using scenario-based questions to expand the ways in which MCQs can be utilized in any classroom. Several participants expressed that they would have liked more time to practice and discuss the tools that were presented in the workshop, underscoring the need for more opportunities for pedagogical training in ecology.

Acknowledgments

We appreciate the valuable insights and contributions of the workshop participants.

Literature Cited

Anderson, L., and D. Krathwohl. 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Longman, New York, New York, USA.

Black, P., and D. Wiliam. 1998. Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice 5:7–74.

Bloom, B. S., M. D. Englehard, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. Taxonomy of educational objectives: the classification of educational goals. Longmans, Green and Company, New York, New York, USA.

Crowe, A., C. Dirks, and M. P. Wenderoth. 2008. Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE—Life Sciences Education 7:368–381.

D'Avanzo, C., C. W. Anderson, L. M. Hartley, and N. Pelaez. 2012. A faculty-development model for transforming introductory biology and ecology courses. BioScience 62:416–427.
Dellinges, M., and D. Curtis. 2017. Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. Journal of Dental Education 81:948–955.
Dirks, C., M. P. Wenderoth, and M. Withers. 2014. Assessment in the college science classroom. W.H. Freeman and Company, New York, New York, USA.
Haladyna, T. M. 1997. Writing test items to evaluate higher order thinking. Allyn and Bacon, Needham Heights, Massachusetts, USA.
Haladyna, T. M. 2004. Developing and validating multiple-choice test items. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.
Haladyna, T. M., and S. M. Downing. 1993. How many options is enough for a multiple-choice test item. Educational and Psychological Measurement 53:999–1010.
Haladyna, T. M., and R. R. Shindoll. 1989. Item shells: a method for writing effective multiple-choice test items. Evaluation and the Health Professions 12:97–106.
Handelsman, J., et al. 2004. Scientific teaching. Science 304:521–522.
Hartley, L. M., J. Momsen, A. Maskiewicz, and C. D'Avanzo. 2012. Energy and matter: Differences in discourse in physical and biological sciences can be confusing for introductory biology students. BioScience 62:488–496.
Herreid, C. F. 1994. Case studies in science–a novel method of science education. Journal of College Science Teaching 23:221–229.
Kehoe, J. 1995. Basic item analysis for multiple-choice tests. ERIC/AE Digest. https://pareonline.net/getvn.asp?v=4&n=10
Klein, T., R. T. W. Siegwolf, and C. Körner. 2016. Belowground carbon trade among tall trees in a temperate forest. Science 352:342–344.
Martinez, M. E. 1999. Cognition and the question of test item format. Educational Psychologist 34:207–218.
Munson, B. H. 1994. Ecological misconceptions. Journal of Environmental Education 25:30–34.
Owen, S. V., and R. D. Froman. 1987. What's wrong with three-option multiple choice items? Educational and Psychological Measurement 47:513–522.
Palmer, E. J., and P. G. Devitt. 2007. Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. BMC Medical Education 7:49.
Rodriguez, M. C. 2005. Three options are optimal for multiple-choice items: a meta- analysis of 80 years of research. Educational Measurement: Issues and Practice 24:3–13.
Scouller, K. 1998. The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. Higher Education 35:453–472.
Stanger-Hall, K. F. 2012. Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE—Life Sciences Education 11:294–306.
Summers, M. M., B. A. Couch, J. K. Knight, S. E. Brownell, A. J. Crowe, K. Semsar, C. D. Wright, and M. K. Smith. 2018. EcoEvo-MAPS: an ecology and evolution assessment for introductory through advanced undergraduates. CBE—Life Sciences Education 17:ar18.
Treagust, D. F. 1988. Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education 10:159–169.
Wiggins, G. P., and J. McTighe. 1998. Understanding by design. Association for Supervision and Curriculum Development, Alexandria, Virginia, USA.