ASSESSMENT

Application of Bloom's Taxonomy Debunks the "MCAT Myth"

Analyses of questions that evaluate critical thinking, from college placement and medical school admission examinations, suggest improvements to college teaching methods.

Alex Y. Zheng,[1] Janessa K. Lawhorn,[1] Thomas Lumley,[2] Scott Freeman[1]*

Biology education and medical education are under scrutiny. The essence of the critique is that introductory biology courses for undergraduates and basic science courses in medical schools overemphasize "factual minutiae" [(1), p. 1343] over the things that working biologists and physicians claim students actually need to master—specifically, critical thinking and professional skills (2–4). The Medical College Admission Test (MCAT) has been accused of hindering efforts to introduce more critical thinking into introductory biology courses (3), and the Advanced Placement (AP) Biology course has come under fire for stressing rote memorization (5, 6). Are these criticisms valid? If so, why is there a disjunction between what future biologists and physicians need to know and what they are taught?

To test the hypothesis that some biology-related exams emphasize factual recall, we used Bloom's taxonomy (7–9) to quantify the level of learning that students are asked to demonstrate on a sample of course exams and standardized tests. Our approach was inspired by literature that uses Bloom's taxonomy to evaluate test questions and student learning (9). The taxonomy identifies six levels of understanding. The first four levels (knowledge, comprehension, application, and analysis) are hierarchical, meaning that a student must know and understand a topic in order to apply it to a new situation or analyze it. Only levels 3 through 6 (application, analysis, synthesis, and evaluation) are considered higher-order thinking (10, 11).

The exams we analyzed came from five sources: AP Biology, introductory biology courses for undergraduate majors from three universities in our region, the biology section of the MCAT, the biology Graduate Record Examination (GRE), and five first-year medical school courses from an institution with a traditional curriculum.

Three experts in biology education assessed 586 questions drawn at random from these five sources and used Bloom's taxonomy to assign a Bloom's rating of 1 to 6 to each question (see chart, p. 415). The ratings were done blind, and inter-rater agreement was high (11). Before comparing ratings, we weighted them to reflect their point value relative to the other questions sampled from the same exam (11). Using the weighted ratings, we compared the proportion of higher-order questions among exam sources and the mean ratings among exam sources using F tests based on linear regression (see table, above). Significance levels for pairwise tests were based on Hommel's correction for multiple comparisons.

	n	Weighted proportion of higher-order questions	Unweighted ratings	Weighted ratings
AP Biology	157	0.36 ± 0.06	1.97 ± 0.07	2.35 ± 0.12
Undergraduate	111	0.51 ± 0.07	2.24 ± 0.09	2.43 ± 0.14
GRE	108	0.35 ± 0.05	2.32 ± 0.09	2.32 ± 0.10
MCAT	109	0.45 ± 0.05	2.55 ± 0.09	2.57 ± 0.09
Medical School	101	0.21 ± 0.04	1.97 ± 0.09	1.93 ± 0.08

Representations of higher-order questions. Bloom's ratings from five types of exams (means ± SEM). GRE and MCAT all had only multiple choice (MC) questions; Medical School had only two non-MC questions.

Debunking the "MCAT Myth"

The strongest pattern in our data is the strength of the MCAT exam, which is a good predictor of success in medical school (12, 13). The MCAT has a greater proportion of questions that demand higher-order thinking than our sample of first-year medical school courses ($P = 0.002$); on average, its questions also have a higher average-weighted rating ($P < 0.0001$). These results provide quantitative support for opinions that at least some traditional first-year medical courses are too low-level (1, 4) and, instead, support instructional models that emphasize clinically oriented problem-based learning or case-based learning in the first and second year of medical school [e.g. (14)].

When the MCAT is compared with AP Biology or the undergraduate exams in our sample, there is no significant difference in average-weighted Bloom's rating or the proportion of higher-order questions. This is noteworthy because the MCAT is strictly multiple choice, whereas the AP exam and several of the undergraduate exams include essay and short-answer questions. Questions that demand written answers have a much higher average-weighted Bloom's rating [AP exam, $P = 0.003$, and the undergraduate exams in our sample, $P = 0.0002$]. When only multiple-choice questions are included in the analysis, the MCAT and GRE exams have weighted Bloom's ratings that are significantly higher, on average, than the other three sources of exams (table S1 in supporting online material) (11).

The MCAT also has the lowest percentage of knowledge-level questions (see chart, p. 415) and the highest average raw question rating of any of our five exam sources (see table, above). These data belie the perception that the MCAT is based heavily on content knowledge and that biology courses must therefore focus more on facts than on skills such as teamwork and the analysis of quantitative evidence (15).

Taken together, these observations suggest that the biology portion of the MCAT fulfills its stated goal of assessing problem-solving ability and critical thinking, in addition to mastery of basic biology concepts (16).

Should the AP Biology Course and Undergraduate Biology Courses Be Reformed?

We found no significant differences in the proportion of higher-level questions or weighted mean Bloom's ratings in AP biology exams versus the undergraduate exams in our sample. This is surprising, given the extent of criticism that the AP Biology course has received and the College Board's current effort to reform the course with the goal of emphasizing higher-order thinking.

Two observations are relevant here. First, an F test confirms significant hetero-

[1]Department of Biology, Box 355320, University of Washington, Seattle, WA 98195, USA. [2]Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA.

*Author for correspondence. E-mail: srf991@u.washington.edu

geneity among the three undergraduate institutions in our sample in terms of proportion of higher-level questions ($P = 0.03$) (*11*). Thus, even our small sample indicates that some undergraduates are being challenged with a greater proportion of higher-order questions in their introductory courses than others. This result suggests that some undergraduate courses need to be reformed to emphasize higher-order thinking skills. Second, the AP Biology exam we analyzed contained 120 lower-level multiple-choice questions worth 60% of the total score and four higher-level essay questions worth 40% of the total score. These exams tested a low-level understanding of a wide array of topics but a higher-order understanding of just four. Thus, criticism (*5*, *6*) of the AP Biology as low-level may be valid, even though the four essay questions are weighted heavily enough to preclude statistically significant differences with the undergraduate exams in our sample.

In addition, the College Board currently validates its exams by having college students take AP exams (*17*). Because of the AP Biology exam's unusual structure, AP coursework should instead be validated by having students from AP courses take college exams.

Disjunction Between Goals and Practice

Is there is a conflict between what undergraduates are being taught and what professional biologists and practicing physicians think students need to master? According to our data, this disjunction may be present in biology courses that assess students using only multiple-choice questions and in some traditional medical school curricula. Part of the solution may lie in dispelling the MCAT myth—"freeing" introductory biology instructors to assess at a higher level by convincing them that the MCAT tests more than factual recall. Faculty could also be reassured that because of the hierarchical nature of levels 1 through 4 of Bloom's taxonomy, content isn't "lost" when exams focus on applying and analyzing knowledge and concepts.

Another part of the answer is more practical. It is much more difficult to write multiple-choice questions at the application and analysis levels of Bloom's taxonomy than at the knowledge or comprehension levels (*18*). Experienced practitioners estimate that writing one "good" multiple-choice question takes an hour (*19*). In addition, students perform better on questions that test factual recall than on questions that test their ability to explain or apply concepts (*20*). Writing

and taking exam questions that test memorization is much faster and easier—for both student and teacher—than writing and taking exam questions that test higher-order thinking. Our data suggest that there is an urgent need for instructors to create and share multiple-choice questions that test at higher levels of Bloom's taxonomy. The higher orders of thinking are not restricted to essay questions—the MCAT and GRE show that multiple-choice questions can be constructed to test these higher orders of Bloom's taxonomy.

Using Bloom's Taxonomy as a Tool

The general goal of this study was to explore how researchers might use Bloom's taxonomy to quantify and compare levels of assessment from different exams. Because our sample of undergraduate biology courses or first-year medical school courses is small, further analysis should be done for a larger sample of schools,
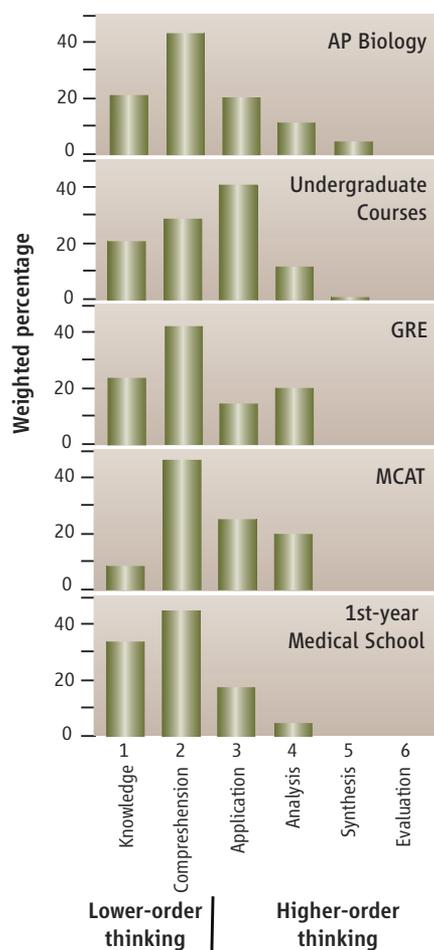
including medical schools with first-year courses that stress clinical application in a problem-based or case-based learning framework.

Because the AP, GRE, and MCAT exams are nationally standardized, however, our assessment of the biology questions on these exams should be robust. It would be interesting to compare the results reported here with analyses of exams from other AP courses in math and science, the Biochemistry GRE, Dental Admission Test, Nursing Entrance Exam, and Pharmacy College Admission Test. Bloom's taxonomy offers a tool for assessing the level of instruction in individual classrooms, monitoring curriculum changes, and making interinstitutional comparisons.

References and Notes

1. M. Cooke, D. M. Irby, W. Sullivan, K. M. Ludmerer, *N. Engl. J. Med.* **355**, 1339 (2006).
2. B. Alberts, *Cell* **123**, 739 (2005).
3. National Research Council (NRC), *BIO 2010: Transforming Undergraduate Education for Future Research Biologists* (National Academies Press, Washington, DC, 2003).
4. M. E. Whitcomb, *Acad. Med.* **81**, 413 (2006).
5. NRC, *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools* (National Academy Press, Washington, DC, 2002).
6. W. B. Wood, *Cell Biol. Educ.* **1**, 123 (2002).
7. B. S. Bloom (Ed.), *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain* (David McKay, New York 1956).
8. D. R. Krathwohl, *Theory Pract.* **41**, 212 (2002).
9. L. W. Anderson, L. A. Sosniak, Eds. *Bloom's Taxonomy: A Forty-Year Retrospective* (Univ. Chicago Press, Chicago, 1994).
10. A. N. Bissell, P. P. Lemons, *BioScience* **56**, 66 (2006).
11. Methods and supplemental data are available on *Science* Online.
12. T. Donnon, E. O. Paolucci, C. Violato, *Acad. Med.* **82**, 100 (2007).
13. N. R. Kuncel, S. A. Hezlett, *Science* **315**, 1080 (2007).
14. K. Hoffman, M. Hosokawa, R. Blake Jr., L. Headrick, G. Johnson, *Acad. Med.* **81**, 617 (2006).
15. J. B. Labov, *Cell Biol. Educ.* **4**, 7 (2005).
16. Association of American Medical Colleges (AAMC), *2007 MCAT Essentials* (AAMC, Washington, DC, 2007).
17. D. L. Morgan, M. P. Michaelides, *Setting Cut Scores for College Placement* (Research report no. 2005-9, The College Board, New York, 2005).
18. J. C. Masters *et al.*, *J. Nurs. Educ.* **40**, 25 (2001).
19. J. K. Farley, *Nurse Educ.* **14**, 10 (1989).
20. K. T. Knecht, *Am. J. Pharm. Educ.* **65**, 4 (2001).
21. We thank A. Crowe, C. Dirks, and M. P. Wenderoth for providing the exam question ratings; J. Nelson and C. Marzban for help with calculating measures of interrater reliability; S. P. Millard for assistance with statistical analyses; an anonymous reviewer for insight into the implications of this study for how AP exams are validated; and an anonymous reviewer for advice on statistical analyses. We are grateful to the College Board; the AAMC; the Educational Testing Service; and instructors at Brigham Young University, Montana State University, and the University of Washington for permission to use exam questions.

Frequency of exam questions at six levels on Bloom's taxonomy of learning. Proportion of exam questions, weighted to reflect their relative point value, characterized by the six levels of Bloom's taxonomy (*11*).