Adv Physiol Educ 41: 170–177, 2017;
doi:10.1152/advan.00101.2016.

ILLUMINATIONS

Bloom's dichotomous key: a new tool for evaluating the cognitive difficulty of assessments

Katharine Semsar and Janet Casagrand
Department of Integrative Physiology, University of Colorado, Boulder, Boulder, Colorado

Submitted 1 July 2016; accepted in final form 17 January 2017

ONE OF THE MORE WIDELY USED TOOLS to both inform course design and measure expert-like skills is Bloom's taxonomy of educational objectives for the cognitive domain (2, 13, 22). This tool divides assessment of cognitive skills into six different levels: knowledge/remember, comprehension/understand, application/apply, analysis/analyze, synthesis/create, and evaluation/evaluate (2, 6). The first two levels are generally considered to represent lower levels of mastery (lower-order cognitive skills) and the last three represent higher-order levels of mastery involving critical thinking (higher-order cognitive skills) with apply-level questions often bridging the gap between the two (e.g., Refs. 5, 8, 10, 11, 23, and 24). While Bloom's taxonomy is widely used by science educators, learning and mastering the concepts of the cognitive domain to categorize educational materials into the six levels identified in Bloom's taxonomy are not trivial tasks.

As with any complex task, experts and novices differ in the key abilities needed to cue into and evaluate information (4, 7, 9). Across disciplines, novices are less adept at noticing salient features and meaningful patterns, recognizing the context of applicability of concepts, and using organized conceptual knowledge rather than superficial cues to guide their decisions. Newer users of Bloom's taxonomy demonstrate similar difficulties as they work to gain expertise, leading to inconsistencies in Bloom's ratings (1, 8, 15) (see *BDK Development* for examples).

To help novices gain expertise in a discipline, a common educational strategy is the use of scaffolding (7, 17, 21). Scaffolding aims to control the elements of a task, allowing a novice learner to complete the easier levels of the task and build up to the more complete and complex elements of the task (17). In the context of "Blooming," a scaffolding structure would help the rater cue into the salient and most important elements of a question relating to the skill level of the problem and aid in using those elements to categorize the specific skill being tested in the problem. A scaffolding tool therefore provides a structure with which the novice could model their identification of key elements and decision making.

One such example of a scaffolding tool to use for Bloom's taxonomy is the Biology Blooming Tool (BBT) (8). The BBT is a conventional rubric for developing and identifying biology-specific skills and questions based on Bloom's taxonomy. Organized as a table, each column of the rubric table outlines the key skills assessed at a given Bloom's level (starting with the lowest level, "remember"), provides examples of exam questions, and delineates the type of exam questions that can be asked at that level. Unfortunately, in our own attempt to Bloom exam questions and course materials using a modified BBT, we had difficulty getting three independent raters to consistently rate materials. Therefore, we set out to design a new Bloom's training tool that would provide additional, specific scaffolding that directly addressed the inconsistencies among our raters and thus might lead to greater consistency among raters. Here, we present a description of the development and evaluation of that tool: Bloom's dichotomous key (BDK).

BDK Development

The development and analysis of the BDK was conducted under Institutional Review Board protocol 0108.9 (exempt status).

Rationale and initial independent rater training. The development of the BDK grew from an attempt to evaluate the Bloom's level of course content before and after course reform efforts in a neurophysiology course (J. Casagrand and K. Semsar,7a). One way we sought to assess the effectiveness of the course reform was to use Bloom's taxonomy to categorize the cognitive level of course exams and other course materials before and after reform as an indirect, retrospective measure of changes in student understanding. Thus, using Bloom's taxonomy as an indirect measure of student understanding could provide a way to gauge how the neurophysiology course had changed over time and whether students were able to demonstrate deeper levels of understanding of course content.

To reduce potential bias while "Blooming" course materials, we began by recruiting three independent raters. One rater was a current graduate student in the department who had previously been a teaching assistant for the course, and two raters were former graduate students who remained in the department as postgraduates, one of whom had been a teaching assistant for the course and the other who had taken the course as an undergraduate. In selecting these raters, we were careful to choose raters who were familiar with the course content and knowledgeable of neurophysiology because it was important that they had a sufficient understanding of the course content to recognize what knowledge and problem-solving skills a student would need to answer each question, such as whether students were being asked to apply concepts in new contexts or remembering material exactly as presented in class. However, the three raters had varying expertise and experience using Bloom's taxonomy to assess the cognitive skill level of course

J. Casagrand, Dept. of Integrative Physiology, Univ. of Colorado, 354 UCB, Boulder, CO 80309-0354 (e-mail: Janet.Casagrand@colorado.edu).

material. One rater had been extensively trained in Bloom's taxonomy and used Bloom's taxonomy over several years of working as a science education specialist. Another rater was also working as a science education specialist but had received minimal training before working on this project. Our third rater had no prior exposure to Bloom's taxonomy before this project.

To familiarize our raters with the process of "Blooming" course materials, we initially provided the raters with an overview of Bloom's taxonomy, associated terms, and sample questions in a conventional rubric modeled after the BBT. We had raters practice categorizing 26 sample neurophysiology questions. Unfortunately, we were dissatisfied with the degree of categorization similarity among the three raters. On average, the raters only matched the authors' question categorizations 46% of the time (Table 1). In addition, raters were deviating almost a full Bloom's category (0.65) from the average rating (Fig. 1), and the percentage of questions for which all three raters agreed was only 19% (Table 1). This prompted discussions between the authors and raters that suggested significant discrepancies and variability in the raters' reasoning processes in assigning ratings to the sample questions.

Think-aloud interviews. These discussions led us to perform individual think-aloud interviews with each rater to better discern how raters were using the rubric to make decisions. During each think-aloud interview, the rater verbalized his/her thought processes and reasoning as he/she used the rubric to categorize each sample exam question. (Raters were familiar with the course and questions were labeled as to which exam they came from so that raters would know what was taught.) If the rater did not provide reasoning, he/she was prompted to explain their choice, but that was the only prompt given by the interviewer. When all three interviews with raters were complete, we examined the raters' reasoning during their decision-making processes, specifically looking at reasons given for categorizations for which raters disagreed with each other.

During think-aloud interviews, we observed several inconsistencies in rater decision-making, most of which were similar with published accounts of difficulties in "Blooming" (1, 3, 5, 8, 15, 23). First, raters did not always take into consideration what information had been previously provided to students, an important aspect in determining the appropriate Bloom's level (also described in Refs. 1, 3, and 8). For example, if students are given the answer to a specific higher-level question in lecture and then asked the same question on an exam, answering the question only requires recall, not a higher level of understanding (see *example 1* in Fig. 2). The remaining inconsistencies all centered around raters focusing on different information within the question. For example, as also described in detail by Lemons and Lemons (15), raters would sometimes categorize questions based on the perceived difficulty of a question rather than what students would need to do to answer the question (i.e., the cognitive skills required). If a rater thought it was a more difficult problem, the rater might skew the rating to a higher level without reference to what students were actually being asked to do with the information, and vice versa (see *examples 2* and *3* in Fig. 2). Another common reason for inconsistency among our raters and similar to inconsistency described by others (5, 8, 23) stems from raters cueing in to different skills or information needed to answer a single question. This often involved questions in which more than one concept or piece of information was being tested, and the different concepts/information required different cognitive skills. Most often, raters would stop at the lower-level categories and not take into account that higher-order questions about a concept also include mastery of lower level cognitive skills related to that concept (see *example 4* in Fig. 2). In addition, category inconsistencies were commonly related to the raters' use of buzzwords or action verbs for categorization rather than the specific information and context in the question. For example, questions asking for the best answer were sometimes categorized as evaluate due to the appearance of making a judgment, even if, based on the context of what was taught, the question was at a remember, comprehend, or apply level (see *example 5* in Fig. 2). From other experiences, we know the term "predict" also commonly leads to similar inconsistencies (see *example 6* in Fig. 2). Finally, we found that questions involving data were especially prone to large categorization variation, as did Crowe et al. (8). We found that questions with data sets sometimes led raters to jump directly to the Bloom's category of apply or analyze without giving close attention to the question. However, in our examination of how students can be asked to interpret data in different questions, nearly all Bloom's skill levels could be represented, from deciding whether data are consistent with a hypothesis (evaluate) to drawing conclusions about what the data mean (analyze) to simply redescribing the data (comprehend).

Building the dichotomous key. Through the think-aloud process, we noticed the issues listed above paralleled the processes known to reflect cognitive processes of novices in general. For example, novices generally either fail to notice,

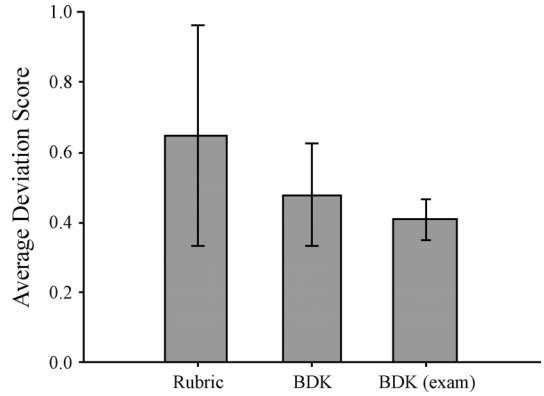
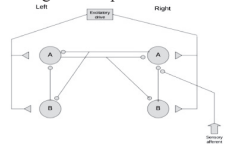


Fig. 1. Comparison of the mean average deviation scores (±SD) for the 3 raters on the initial 26 sample questions with the conventional rubric (a modified Biology Blooming Tool) and with the Bloom's dichotomous key (BDK) and on the 155 exam questions [BDK(exam)]. Average deviation scores (i.e., how much each rater deviated from the average rating of the three raters) were calculated for each rater based on the method described in Zheng et al. (23), as follows: Average deviation $= \Sigma_{i=1}^{n}|$(average rating of 3 raters) − (individual rater's rating)$|/n$

Table 1. *Percent rater agreement*

	Number of Exam Questions	Agreement With Authors, %	At Least Two Raters Agree, %*	All Three Raters Agree, %
Rubric	26	46	88	19
BDK	26	67	88	38
BDK (exams)	155	Not applicable	81	41

BDK, Bloom's dichotomous key. *Not always the same two raters.

Bloom's Taxonomy Challenge	Example Question	Bloom's Level/Quotes/Misstep Examples
Example 1: Bloom's level depends on what was taught	Tetraethylammonium (TEA) is a drug that blocks voltage-gated K^+ channels. Make a drawing to illustrate an action potential before the application of TEA and an action potential after the application of TEA.	*Case 1*: If the instructor talked about the effects of TEA in detail and showed a before/after graph: *Remember* *Case 2*: If the instructor talked about the effects of TEA in detail: *Comprehend* *Case 3*: If the instructor just taught action potential basics: *Synthesize/create*
Example 2: rating based on perceived item difficulty rather than cognitive skill	The drawing below illustrates a pattern generator circuit. Triangles denote excitatory synapses; circles denote inhibitory synapses. The excitatory drive provides a tonic descending drive. In addition to the connections shown, *cell A* also synapses onto a motor neuron on the same side. *Cell A* discharges in the pattern shown below.  What cells are most responsible for the alternation of activity?	Authors' Bloom's level: *Analyze*. Based on what is shown in class, students have practiced analyzing pattern generator circuits but have not seen this circuit. To solve this problem, they have to analyze different parts of the circuit to determine which cells are responsible for the pattern depicted. Rater misstep (with the rubric): *They [students] see something similar in class, so this shouldn't be too difficult. I'd say comprehend.* Rater with the BDK: After saying "yes" to *question 7* ("Are there data to interpret?"), the rater says "yes" to *question 9*. *Yes, they (students) have to do both, decide what the data means and need to know which parts of the diagram are relevant.*
Example 3: rating based on perceived item difficulty rather than cognitive skill	In a cell in which membrane resistance (R_m) is high relative to axial resistance (R_a), the length constant would be _____ (large or small) and you would expect the action potential to be _____ (faster or slower) than in a cell in which R_m is low relative to R_a? *A*. Small, faster *B*. Small, slower *C*. Large, faster *D*. Large, slower	Authors' Bloom's level: *Comprehend*. Although these concepts have been taught in class, they have not been phrased in such this way. Rater misstep (with the rubric): *They (students) struggle with this…Apply.* Rater with the BDK: After saying "maybe" to *question 7* ("Are there data to interpret?"), the rater says "no" to *question 12* and then "yes" to *question 13*. *Yes, showing the relationship.*
Example 4: need to evaluate separate parts of a question	What are the directions of the chemical, electrical, and net driving forces acting on K^+ when the membrane potential is −55 mV? *A*. Inward, outward, outward *B*. Outward, inward, outward *C*. Outward, inward, inward *D*. Outward, outward, inward *E*. None of the above.	Authors' Bloom's level: *Apply*. Students not only have to understand the concepts but also have to answer the direction of the net driving force; they must calculate the direction. Rater misstep (with the rubric): *Comprehend. Students have to put it together and use (their information) and show they understand these concepts.* Rater with the BDK: Getting to *question 12*: *Yes. They haven't had K^+ before so they'll have to calculate this.*
Example 5: use of higher-order cognitive skill language for lower-order cognitive skill questions	Which of the following factors BEST accounts for posttetanic potentiation? (Only one answer is correct.) *A*. Increased synthesis of transmitter *B*. Slower breakdown of transmitter in the synaptic cleft *C*. A build up of Ca^{2+} presynaptically *D*. A build up of Ca^{2+} postsynaptically *E*. increased sensitivity of the postsynaptic membrane	Authors' Bloom's level: *Remember*. Students should remember the answer from lecture. Rater misstep (with Rubric): *If it was just name or list one factor, then it would be recall. But BEST is in bold so there must be more than one possible answer. Then you're making a judgement call. Evaluate.* Rater with the BDK: (They are told that this definition is given in class.) *Question 1*: *If they are given this definition, so yes (remember).*
Example 6: use of high-order cognitive skill language for lower-order cognitive skill questions	Predict how blood pressure changes with increasing heart rate. (But instructor taught this.)	Authors' Bloom's level: *Remember or comprehend*, depending on what exactly was taught. Rater (from the workshop) misstep: *"Predict" goes with apply.*
Example 7: multiple answers but only one solution	The firing rate of group Ia afferents is affected by: *A*. Changes in steady-state length of extrafusal fibers *B*. Velocity of length changes of extrafusal fibers *C*. Vibration of a muscle *D*. *A* and *B* but not *C* *E*. A and *C* but not *B* *F*. *B* and *C* but not A *G*. *A*, *B*, and *C*	Authors' Bloom's level: *Remember*. Rater (from the workshop) misstep: Raters would reach *question 4*, which used to read ("Is there more than one answer?") and say *yes*. While there is more than one answer (*A*, *B*, and *C* are all correct), there is only one valid solution (that all three are correct).
Example 8: multiple answers but only one solution	What muscle fiber types are active during maximal force generating contractions? *A*. Type I (slow oxidative) *B*. Type IIA (fast oxidative glycolytic) *C*. Type IIB (fast glycolytic) *D*. All of the above	Authors' Bloom's level: *Remember* Rater (from the workshop) misstep: Same logic as *example 7*.
Example 9: multiple valid solutions for the "comprehend" Bloom's level	Provide an example of homeostasis that we haven't talked about in class and isn't in your textbook.	Authors' Bloom's level: *Comprehend* Rater (from the workshop) misstep: Raters would reach the current *question 4*, which reads ("Is there more than one valid solution?") and say *yes*. However, from there, the only choices had been *evaluate*, *synthesize/create*, or *analyze*. Yet, this was clearly a *comprehend* question. Thus, we added a path to *question 16*.

Quotes from raters (in italics) are closely paraphrased.

Fig. 2. Challenges faced in using Bloom's taxonomy and example missteps made by raters. Quotes from raters (in italics) are closely paraphrased.

or do not discriminate well, the salient features within complex patterns. Novices also organize their knowledge based on surface features rather than underlying structure. They also jump quickly to conclusions and do not always recognize the entire context of a problem. Likewise, in initial categorizations, our raters chose different features of problems to use in their categorization decision, relied on buzzwords to categorize items, and misclassified questions

because they had not considered what had previously been taught (Fig. 2).

To address these specific issues and provide raters with the additional scaffolding for the Bloom's categorization process, we developed a new training tool: the BDK (Table 2). For categorization processes such as these, a dichotomous key is a natural scaffolding tool because it allows users to identify and categorize items in a systematic and reproducible fashion (12). Different from a conventional rubric or flowchart, it is a series of steps, each with two choices, that focuses on key character-

Table 2. *The BDK*

- Categorize the question based on what students are being asked to do, not on how challenging the question may be. (For example, a "comprehend" question for a difficult concept could be a more challenging problem than an "analyze" question on an easier concept.)
- Evaluate questions with reference to what material we know students were exposed.

Question 1. Could students memorize the answer to this specific question?
Yes: go to *question 2*.
No: go to *question 4*.
Question 2. To answer the question, are students repeating nearly exactly what they have heard or seen in class materials (including lecture, textbook, laboratory, homework, clicker, etc.)?
Yes → See *Remember*
No: go to *question 3*.
Question 3. Are students demonstrating a conceptual understanding by putting the answer in their own words, matching examples to concepts, representing a concept in a new form (words to graph, etc.), etc.?
Yes → See *Comprehend*
No: Go back to *question 1*. If you are sure the answer to *question 1* is yes, the question should fit into "remember" or "comprehend."
Question 4. Is there potentially more than one valid solution* (even if a "better" one exists or if there is a limit to what solutions can be chosen)?
Yes: go to *question 5*.
No: go to *question 8*.
Question 5. Are students making a judgment and/or justifying their answer?
Yes → See *Evaluate*
No: go to *question 6*.
Question 6. Are students synthesizing information into a bigger picture (coherent whole) or creating something they haven't seen before (a novel hypothesis, novel model, etc.)?
Yes → See *Synthesize/create*
No: go to *question 7*.
Question 7. Are students being asked to compare/contrast information?
Yes → See *Analyze*
No: go to *question 16*.†
Question 8. To answer the question, do students have to interpret data (graph, table, figure, story problem, etc.)?
Yes: go to *question 9*.
No: go to *question 14*.
Question 9. Are students determining whether the data are consistent with a given scenario or whether conclusions are consistent with the data? Are students critiquing validity, quality, or experimental data/methods?
Yes → see *Evaluate*
No: go to *question 10*.
Question 10. Are students building up a model or novel hypothesis from the data?
Yes → See *Synthesize/create*
No: go to *question 11*.
Question 11. Are students coming to a conclusion about what the data mean (they may or may not be required to explain the conclusion) and/or having to decide what data are important to solve the problem (i.e., picking out relevant from irrelevant information)?
Yes → See *Analyze*
No: go to *question 12*.
Question 12. Are students using the data to calculate the value of a variable?
Yes → See *Apply*
No: go to *question 13*.
Question 13. Are students redescribing the data to demonstrate they understand what the data represent?
Yes → See *Comprehend*
No: go back to *questions 4* and 8.
Question 14. Are students putting information from several areas together to create a new pattern/structure/model/etc.?
Yes → See *Synthesize/create*
No: go to *question 15*.
Question 15. Are students predicting the outcome or trend of a fairly simple change to a scenario?
Yes → See *Apply*
No: go to *question 16*.
Question 16. Are students demonstrating that they understand a concept by putting it into a different form (new example, analogy, comparison, etc.) than they have seen in class?
Yes→ See *Comprehend*
No: go back through each category or refer to category descriptions to see which fits the best

*This question originally had the word "answer" in place of the word "solution." In subsequent use of the BDK, we found that the word solution led to less confusion about the application of this question. This was not an issue in our initial use of the BDK for this report. †Originally, if answering "no" to *question 7*, we had reviewers go back to *question 4* and if they were sure it was "yes," they should be able to answer "yes" to *questions 5*, *6*, or *7*. This did not lead to any difficulties in our initial use of the BDK for this report. However, in subsequent use of the key, we found examples of questions in which comprehension-level questions were also possible. Therefore, we revised the BDK to lead raters to *question 16* here to account for those question types.

istics of a particular group to reproducibly sort them into taxonomic groups. While experts can make these categorizations quickly using patterns of knowledge, novices can use this step-wise series of questions to focus on salient information and consistently make identifications. For example, an expert in phylogenetic identification can use salient features and patterns of knowledge that have become second nature to identify organisms without needing the help of taxonomic descriptions. Meanwhile, novice biologists can use dichotomous keys to help them develop recognition of salient features that lead to taxonomic identification. Thus, rather than sifting through taxonomic descriptions of each species and then trying to match their specimen to the descriptions, the novice looks at the specimen and answers a series of questions. For example, the key may start with the following query: "Does the organism have cell walls?" If yes, Kingdom Plantae, go to *question 2*; if no, Kingdom Animalia, go to *question 5*. From there, the dichotomous key follows a series of such salient features that help narrow down the classification choices. In this way, the dichotomous key scaffolds the pattern recognition of identification into specific steps, feature by feature. In this same manner, we created the BDK to scaffold the process of categorizing cognitive skill levels using Bloom's taxonomy.

When developing a dichotomous key, one first identifies classifying characteristics, those features of the items that create large distinctions among groups of items, until all items can be uniquely referenced. These classifying characteristics are then organized from the broadest to narrowest such that raters answer a series of "yes or no" questions that guide them through common elements of questions and ultimately to a Bloom's level for the question being categorized. Using our observations from the think-aloud interviews, we determined our three broadest classifying characteristics guiding Bloom's categorization decisions were *1*) whether or not the answer to the specific question could have been memorized, *2*) whether there was more than a single plausible/valid solution to a problem, and *3*) whether the problem contained data interpretation. Yes or no responses to the prompts associated with these characteristics then lead to further distinguishing features of specific Bloom's taxonomic groups. The BDK begins with the two broadest classifying characteristics: whether or not the answer could be memorized (*question 1*; if yes, then the classification is "remember") and whether there was more than a single plausible solution (*question 4*; nearly every time there is more than a single correct way to approach a problem, one will be working at a higher cognitive level, such as "analyze," "evaluate," or "create/synthesize"). These BDK questions sort most exam/homework questions that fall into the lower-order cognitive skills or higher-order cognitive skills of Blooms taxonomy. From there, the BDK moves to the third broad classifying characteristic: whether the question requires interpretation of data (*question 8*). If the rater answers yes to this question, the BDK guides the rater through the different cognitive skills that can be tested under the broader context of interpreting data [e.g., describing data ("comprehend") or using data to calculate an answer ("apply")]. The last few BDK prompts help sort the remainder of the question types we encountered.

In addition to using these classifying characteristics to aid the raters in their categorizations, we also designed the BDK to clarify other common sources of inconsistencies. First, to resolve the issue of raters categorizing questions based on perceived difficulty rather than the skills needed to solve a problem, all BDK prompts are specifically worded to ask raters what a student is being required to do to answer a question (e.g., recall a fact, calculate a number, or interpret data). Second, to address the fact that raters had sometimes stopped at the lower Bloom's levels when using the conventional rubric and did not take into account that higher-order questions include mastery of lower-order cognitive skills, we designed the BDK to guide raters to generally consider higher-order skills before lower-order skills within each section of the BDK. Third, some rater discrepancies were due to raters not taking into account all information that students had to work with. Thus, many of the BDK prompts specifically have the rater take into account whether students are considering only a single piece of information (generally lower-order cognitive skills) or multiple pieces of information (generally higher-order cognitive skills).

To ensure the prompts were being interpreted appropriately and consistently, we then performed additional think-aloud interviews as our 3 raters used the BDK to rerate the original 26 sample questions. (No feedback was given to raters between their use of the conventional rubric and BDK during think-alouds.) Based on the interviews and additional rater feedback, the wording of some of the BDK prompts was revised. For example, the first prompt was changed from "Have students seen the answer to this question in the course materials?" to "Could students memorize the answer to this specific question?" In addition, we changed the fourth prompt, "Is there potentially more than one valid answer?" to "Is there potentially more than one valid solution?" This distinction was necessary to avoid confusion between cases in which a single solution to a question included multiple components that appeared like separate answers (see *examples 7* and *8* in Fig. 2). Finally, from feedback we later received during workshop sessions using the BDK, we added a prompt (*question 16*: "Are students demonstrating that they understand a concept by putting it into a different form than they have seen in class?") to address question types that were not originally in our sample questions but were represented in other materials subsequently "Bloomed" with the BDK.

BDK Evaluation

Statistical analysis. To evaluate whether the BDK was meeting our goal of creating greater categorization similarity among raters, we statistically compared mean average deviation scores and SDs of deviation scores (23) between the use of the more conventional, BBT-styled rubric and BDK. Our raters produced significantly more similar categorizations when using the BDK than when using the conventional rubric to rate the same 26 sample questions. The mean average deviation score dropped from 0.65 to 0.48 (Fig. 1). In another measure of rater agreement, we looked at the percentage of questions for which multiple raters agreed on a categorization. Although at least two raters agreed on a categorization for 88% of the questions for both the conventional rubric and BDK, the percentage of questions for which all three raters agreed doubled to 40% when using the BDK (Table 1, comparable with Ref. 23). Furthermore, use of the BDK reduced the SD of average deviation scores by more than half, from 0.31 to 0.14, indicat-

ing that when a rater deviated from the average rating, he/she did not deviate as far from the average. While it is possible that these scores were simply getting more consistent with rater practice, we do not believe this is the case as the degree of variation was lower after use of the BDK than use of the rubric despite no additional discussions or training between these events. Finally, in addition to the raters becoming more consistent with each other, they also were more likely to match the authors' categorization of a question, with the average match between raters and the authors jumping from 46% to 67% (Table 1).

Think-aloud interview observations. The statistical improvement in consistency that we saw when raters used the BDK for categorization of course materials was supported by the think-aloud observations conducted when raters used the BDK. During these think-aloud interviews, we observed that the BDK specifically brought rater attention to what had been taught and helped raters consider what had been taught in relation to question categorization. Second, use of the BDK focused raters on considering the skills being used by a student in answering the question rather than other nonsalient features, such as perceived item difficulty. [For example, while one rater was settling on "analyze" based on *question 11* (rather than "comprehend," which "felt" more right based on perceived difficulty), they said "I'm not totally happy with this, it seems really simple. But they do have to decide what's important to solve the problem. That's more than just a calculation."] Third, use of the BDK improved the rating consistency of questions involving data. Pre-BDK, raters would often see data in the exam question and immediately jump to the Bloom's category of analyze because they conflated all data with "analyzing data." The BDK helped raters focus on what students were being asked to do with the data rather than just noting that data were involved in the question. (For example, while one rater was deciding on what was being done with information, they said "They're given information, but they aren't interpreting all of it [to answer the question].") Fourth, when using the BDK compared with the conventional rubric, raters focused on the highest cognitive level of the question rather than focusing on a lower-level component embedded within a higher-level question. In addition, all three raters reported that the dichotomous key was easier and faster to use than the conventional rubric. In addition to the observer's notes that raters spent much less time going back and forth between category descriptions, raters also said "This went a lot faster, and I'm more confident [in my answers] too and [Questions] seem to bin well, definitely quicker".

Utility. We examined the utility of the BDK in two ways. First, we had raters use the BDK to categorize course materials from a neurophysiology course to examine the effectiveness of course reforms (K. Casagrand and J. Semsar, (7a)). Briefly, our three raters used the BDK to categorize 155 exam questions, ascertaining that the number of Bloom's higher-order questions on exams more than doubled, from 24% to 67%, after the introduction of several research-based teaching methods. In addition, a single rater used the BDK to categorize 394 homework and clicker questions to demonstrate the degree of alignment of course materials.

Second, we presented the BDK in several Bloom's taxonomy workshops. In one such workshop, attendees were surveyed about their prior level of experience with Bloom's taxonomy and their opinions about the utility of the BDK (Table 3). Overall, of the 25 attendees, 19 attendees had limited previous experience with Bloom's taxonomy before the workshop. All but one of these attendees rated the BDK as easy to use, and all but one attendee would both use the BDK themselves in the future and recommend the BDK to people who are new to Bloom's taxonomy. Of the four people who had extensive experience with Bloom's taxonomy, three individuals also agreed that they would recommend the BDK to people who are new to using Bloom's taxonomy. Meanwhile, the two people who had no previous experience with Bloom's found the tool difficult to use.

Discussion

The BDK as a Bloom's taxonomy training tool. Learning to efficiently and fluently use Bloom's taxonomy is a challenging cognitive task. Thus, not surprisingly, many of the categorization inconsistencies demonstrated by both our rater team and others (5, 8, 23) are typical of those that novices face in general when performing cognitively complex tasks (7). While more conventional rubrics and guides like Anderson's guide to Bloom's taxonomy (2) and the BBT (8) can aid in learning the complexities of Bloom's taxonomy, they were not sufficient for our raters during training. In our development of a more structured training tool, the use of a dichotomous key (BDK) provided a more specific scaffolding that allowed raters to streamline their classification process and direct their attention toward the more salient features of a question (such as skill level rather than perceived difficulty or buzzwords), thus resolving many of the aforementioned discrepancies in reasoning and decision-making originally encountered with the more conventional rubric. The BDK also provided raters a starting point from which to start their classifying decisions, saving the raters time by having them search for specific characteristics of a question rather than spending time rereading Bloom's descriptions. These unique features of the BDK resulted in significantly more consistent and reliable Bloom's categorizations. As the BDK specifically addresses categorization difficulties common to novices, the

Table 3. *Feedback about the BDK from a Bloom's taxonomy training workshop*

	Prior Experience with Bloom's Taxonomy		
	Extensive	Limited	None
Number of workshop attendees with Bloom's experience	4	19	2
Mean reported difficulty level in using the BDK (*1* = very easy and *5* = very difficult)	1.75	2.22	4
Number of attendees who would use the BDK in the future to help "Bloom" materials	3	18	0
Would you recommend the BDK to others who are new to "Blooming?" Answer: yes	3	18	0

BDK will likely be a useful tool for anyone new to Bloom's taxonomy.

The greater consistency among raters may make the BDK a useful tool for education researchers as well. In particular, because the ratings were more tightly centered around the average rating when using the BDK, we were able to have more confidence in the categorizations by any single rater. This was especially important to us as the resources to have multiple raters are rare. For example, during our own analysis of the course reforms in a neurophysiology course, the time commitment necessary to categorize 394 homework and clicker questions meant that we could only use one rater. Thus, having additional tools to train independent raters on some of the more nuanced distinctions of Bloom's categorizations may help streamline the training process and make "Blooming" course materials more feasible for more people.

Beyond Bloom's taxonomy specifically, it appears the development and publication of taxonomies and similarly styled frameworks to categorize and assess student work and education materials is becoming more common (e.g., Refs. 18 and 19). However, to the authors' knowledge, we have not yet seen the development of other dichotomous keys to accompany such education-related frameworks. As the nature of dichotomous keys can be greatly beneficial, especially to novice users of any classification scheme, the development of dichotomous keys to accompany these new education-related frameworks might prove to be a useful approach. The keys can help to guide the new user to the most salient features of the classification system and provide guidance to common challenges in the categorization process, potentially improving both the ease and accuracy of how these new frameworks are used among a broad audience.

Limitations. While we were able to use the BDK to classify all of the questions on our course's exams and other course materials, we recognize that there will be question types not yet specifically covered by the BDK when it is used more broadly. For example, we used the BDK in a workshop in which a question was used that asked students to provide new examples of a concept (see *example 9* in Table 1). Although the question had multiple possible correct answers, it was still at the level of comprehend and thus did not have an appropriate place in that version of the BDK. While we have corrected this particular issue in the BDK by redirecting the answer for *question 7* to *question 16* (Table 2), we expect other examples will surface as the BDK is used in new contexts. Indeed, we hope that with more extensive use and feedback on the BDK, it may be possible to expand its ability to classify a broad range of biology questions.

Another limitation is that examining the cognitive skill level using Bloom's taxonomy is not the only way in which to categorize the challenge of a question. Item difficulty, time on task, etc. also factor into the level of challenge of an assessment item (7, 14, 15). These are all important judgments about assessment items regarding what students are learning and how they are demonstrating their knowledge. However, our goal here was simply to help increase categorization similarity of Bloom's measures among multiple raters. If the BDK can also help streamline a rater's ability to categorize questions based on Bloom's level, it might be useful in quickly assessing the Bloom's level of materials that can in turn be combined with other indicators of question challenge to get a more complete picture of how and what students are learning and what skills they are demonstrating during assessments.

In addition, another limitation of the BDK lies in the inherent ambiguities in Bloom's cognitive domain levels in any context. Even among experts in the cognitive domain, not everyone will agree on what skills are being used by a student 100% of the time (e.g., Refs. 15 and 23). This is typical of any evaluative process that requires judgment. Furthermore, for complex problems, students may use different cognitive skills than experts to arrive at their answers (4, 7, 9). For example, across multiple disciplines, novices tend to solve problems using superficial cues rather than organized conceptual knowledge (4). These differences in the cognitive skills used to solve problems are not something that can be easily discerned by experts "Blooming" questions without attention to specific student thought processes.

Using the BDK. For anyone wishing to use the BDK, obtaining background on Bloom's taxonomic categories and basic theory is still necessary. For example, when one of the authors (J. Casagrand) gave workshops on using the BDK, she started with an explanation of Bloom's levels and gave examples of buzzwords and a few exemplars before she handed out the BDK and had people use it to "Bloom" questions. However, for the two people who had had no background at all with Bloom's taxonomy, "Blooming" questions was still challenging even with the BDK, suggesting that more discussion about what Bloom's categories are might be helpful.

Once raters become familiar with Bloom's taxonomy theory, if the raters are also familiar with the course and its subject matter, they can begin to use the BDK. If, however, independent raters are not familiar with the course, we have a few additional suggestions. First, as independent raters typically do not know whether material has been taught or used as an example in a particular course, we suggest that the course instructor answer yes or no for *question 1* on the BDK, "Could students memorize the answer to this specific question?" as we did when we categorized neurophysiology course materials. Second, it is important that raters have knowledge of the subject matter to accurately understand what skills students need to answer questions. Finally, as a check for anyone using the BDK, we recommend they refer back to a conventional rubric (e.g., BBT) to confirm that the final rating makes sense.

Outside its use as a specific training tool for categorization of previously generated material, the BDK may also be helpful for refining one's own thought process about and design of course materials and assessments. As instructors trying to incorporate higher-order cognitive skills in their assessments sometimes fall short of these goals (16, 20), the BDK may be useful in drawing attention to common errors people make in this process. For example, if you want students to use data, the BDK prompts that are related to the interpretation of data can provide guidance about the different cognitive skills associated with data interpretation (e.g., describing data or coming to a conclusion). As a general Bloom's taxonomy training tool, the BDK may also help clear up misunderstandings surrounding the use of buzzwords associated with the different cognitive skill levels and help instructors refine the language of exam questions or learning objectives so students can better identify what they are being asked to demonstrate they can do. For example, we have seen in practice, both in our own assess-

ments and in others during Bloom's taxonomy workshops we have hosted, that these buzzwords are not always used correctly, leading to misclassification of the cognitive skills required for a given question. This is especially true when questions are phrased as higher-order Bloom's levels but are really lower-order Bloom's levels (see *examples 5* and *6* in Fig. 2). By focusing raters on cognitive skills rather than action verbs, which may or may not actually express what the learner needs to be able to do, the BDK provides another tool to help someone understand Bloom's taxonomy that may help new users to Bloom's taxonomy better develop their understanding of this theoretical construct.

Conclusions. Using the BDK as a scaffolding structure to help guide decision making during the "Blooming" process, categorization similarity among raters greatly improved. Because the BDK worked well for our raters in evaluating all the neurophysiology questions in pre- and post-reform semesters and was well received at Bloom's workshops, we believe that the current BDK is suitable for use with a wide range of question types. While the BDK does not remove all ambiguities inherent in working with Bloom's taxonomy, we believe that it is a valuable training tool that will *1*) make it quicker and easier for novice raters to use Bloom's taxonomy to determine the cognitive level for exam questions and other course materials and *2*) help instructors new to evidence-based science education bridge the gap between theory and practice, facilitating the use of Bloom's taxonomy to conduct scholarly assessment of course reforms.

ACKNOWLEDGMENTS

The authors thank Françoise Benay-Bentley, Dr. Teresa Foley, Dr. Jeffrey Gould, Dr. Dale Mood, Dr. Jennifer Avena, Dr. Carl Wieman, and the University of British Columbia CWSEI reading group for their assistance with this project.

GRANTS

Financial support was provided by the President's Teaching and Learning Collaborative of the University of Colorado and the University of Colorado Science Education Initiative.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

K.S. conceived and designed research; K.S. and J.C. performed experiments; K.S. and J.C. analyzed data; K.S. and J.C. interpreted results of experiments; K.S. and J.C. prepared figures; K.S. and J.C. drafted manuscript; K.S. and J.C. edited and revised manuscript; K.S. and J.C. approved final version of manuscript.

REFERENCES

1. **Allen D, Tanner K.** Approaches to cell biology teaching: questions about questions. *Cell Biol Educ* 1: 63–67, 2002. doi:10.1187/cbe.02-07-0021.
2. **Anderson LW, Krathwohl DR, Bloom BS (editors).** *A Taxonomy for Learning, Teaching and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives*. San Francisco, CA: Addison Wesley Longman, 2001.
3. **Anderson LW, Sosniak LA (editors).** *Bloom's Taxonomy: A 40 Year Retrospective. Ninety-Third Yearbook of the National Society for the Study of Education*. Chicago, IL: University of Chicago Press, part 2, 1994.
4. **Bassok M, Novick LR.** Problem solving. In: *Oxford Handbook of Thinking and Reasoning*, edited by Holyoak KJ, Morrison RG. New York: Oxford University Press, 2012, p. 413–432.
5. **Bissell AN, Lemons PP.** A new method for assessing critical thinking in the classroom. *BioSci* 56: 66, 2006. doi:10.1641/0006-3568(2006)056[0066:ANMFAC]2.0.CO;2.
6. **Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR.** *Taxonomy of Educational Objectives: the Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: David McKay, 1956.
7. **Bransford J, Brown AL, Cocking R.** *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academies, 2000.

7a. **Casagrand J, Semsar K.** Redesigning a course to help students achieve higher-order cognitive thinking skills: from goals and mechanics to student outcomes. *Adv Physiol Educ*. In press.

8. **Crowe A, Dirks C, Wenderoth MP.** Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sci Educ* 7: 368–381, 2008. doi:10.1187/cbe.08-05-0024.
9. **Docktor JL, Mestre JP.** A Synthesis of Discipline-Based Education Research in Physics. Second Committee Meeting on the Status, Contributions, and Future Directions of Discipline-Based Education Research, Washington, DC: 2011.
10. **Freeman S, Haak D, Wenderoth MP.** Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* 10: 175–186, 2011. doi:10.1187/cbe.10-08-0105.
11. **Fuller D.** Critical thinking in undergraduate athletic training education. *J Athl Train* 32: 242–247, 1997.
12. **Griffing LR.** Who invented the dichotomous key? Richard Waller's watercolors of the herbs of Britain. *Am J Bot* 98: 1911–1923, 2011. doi:10.3732/ajb.1100188.
13. **Handelsman J, Miller S, Pfund C.** *Scientific Teaching*. New York: Freeman, 2007.
14. **Koedinger KR, Booth JL, Klahr D.** Education research. Instructional complexity and the science to constrain it. *Science* 342: 935–937, 2013. doi:10.1126/science.1238056.
15. **Lemons PP, Lemons JD.** Questions for assessing higher-order cognitive skills: it's not just Bloom's. *CBE Life Sci Educ* 12: 47–58, 2013. doi:10.1187/cbe.12-03-0024.
16. **Momsen JL, Long TM, Wyse SA, Ebert-May D.** Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sci Educ* 9: 435–440, 2010. doi:10.1187/cbe.10-01-0001.
17. **National Research Council.** *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, edited by Singer SR, Nielsen NR, Schweingruber HA. Washington, DC: National Academies, 2012.
18. **Quillin K, Thomas S.** Drawing-to-learn: a framework for using drawings to promote model-based reasoning in biology. *CBE Life Sci Educ* 14: es2, 2015.
19. **Richmond G, Merritt B, Urban-Lurain M, Parker J.** The development of a conceptual framework and tools to assess undergraduates' principled use of models in cellular biology. *CBE Life Sci Educ* 9: 441–452, 2010. doi:10.1187/cbe.09-11-0082.
20. **Silverthorn DU, Thorn PM, Svinicki MD.** It's difficult to change the way we teach: lessons from the Integrative Themes in Physiology curriculum module project. *Adv Physiol Educ* 30: 204–214, 2006. doi:10.1152/advan.00064.2006.
21. **Simons KD, Klein JD.** The impact of scaffolding and student achievement levels in a problem-based learning environment. *Instr Sci* 35: 41–72, 2007. doi:10.1007/s11251-006-9002-5.
22. **Wood WB.** Innovations in teaching undergraduate biology and why we need them. *Annu Rev Cell Dev Biol* 25: 93–112, 2009. doi:10.1146/annurev.cellbio.24.110707.175306.
23. **Zheng AY, Lawhorn JK, Lumley T, Freeman S.** Assessment. Application of Bloom's taxonomy debunks the "MCAT myth". *Science* 319: 414–415, 2008. doi:10.1126/science.1147852.
24. **Zoller U.** Are lecture and learning compatible? Maybe for LOCS: unlikely for HOCS (SYM). *J Chem Educ* 70: 195–197, 1993. doi:10.1021/ed070p195.