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

Katharine Semsar  
*University of Colorado Boulder*

Janet Casagrand  
*University of Colorado Boulder*

Follow this and additional works at: https://scholar.colorado.edu/iphy_facpapers

Recommended Citation  
https://scholar.colorado.edu/iphy_facpapers/19

This Article is brought to you for free and open access by Integrative Physiology at CU Scholar. It has been accepted for inclusion in Integrative Physiology Faculty Contributions by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
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


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).


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 THIS 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 CATEGORIZED THE COGNITIVE LEVEL OF COURSE EXAMS AND OTHER COURSE MATERIALS BEFORE AND AFTER REFORM AS AN INDIRECT, RETROSPECTIVE MEASURE OF CHANGES IN STUDENT UNDERSTANDING. THIS, 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 MATERIAL, 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 KNOWLEDGABLE 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 discus-
sions 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 com-
plete, 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 incon-
sistencies 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, answer-
ing the question only requires recall, not a higher level of
understanding (see example 1 in Fig. 2). The remaining in-
consistencies all centered around raters focusing on different
information within the question. For example, as 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 ques-
tion. 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 catego-
rizes 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,

Table 1. Percent rater agreement

<table>
<thead>
<tr>
<th>Number of Exam Questions</th>
<th>Agreement With Authors, %</th>
<th>At Least Two Raters Agree, %*</th>
<th>All Three Raters Agree, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubric</td>
<td>26</td>
<td>46</td>
<td>88</td>
</tr>
<tr>
<td>BDK (exams)</td>
<td>26</td>
<td>67</td>
<td>88</td>
</tr>
<tr>
<td>BDK</td>
<td>155</td>
<td>Not applicable</td>
<td>81</td>
</tr>
</tbody>
</table>

BDK, Bloom’s dichotomous key. *Not always the same two raters.
<table>
<thead>
<tr>
<th>Bloom’s Taxonomy Challenge</th>
<th>Example Question</th>
<th>Bloom’s Level/Mistep Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1: Bloom’s level</td>
<td>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.</td>
<td><strong>Case 1:</strong> If the instructor talked about the effects of TEA in detail and showed a before/after graph: Remember… <strong>Case 2:</strong> If the instructor talked about the effects of TEA in detail: Comprehend <strong>Case 3:</strong> If the instructor just taught action potential basics: Synthesize/create</td>
</tr>
<tr>
<td>depends on what was taught</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example 2: rating based</td>
<td>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?</td>
<td>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.</td>
</tr>
<tr>
<td>on perceived item difficulty rather than cognitive skill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example 3: rating based</td>
<td>In a cell in which membrane resistance (R_m) is high relative to axial resistance (R_a), the length constant would be ___ (larger or smaller) and you would expect the action potential to be ___ (faster or slower) than in a cell in which R_a is low relative to R_m? A. Small, faster B. Small, slower C. Large, faster D. Large, slower</td>
<td>Authors’ Bloom’s level: Comprehend. Although these concepts have been taught in class, they have not been parsed 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.</td>
</tr>
<tr>
<td>on perceived item difficulty rather than cognitive skill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example 4: need to evaluate separate parts of a question</td>
<td>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.</td>
<td>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⁺ so before they’ll have to calculate this.</td>
</tr>
<tr>
<td>Example 5: use of higher-order cognitive skill language for lower-order cognitive skill questions</td>
<td>Which of the following factors BEST accounts for posttectonic 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 Ca²⁺ presynaptically D. A build up of Ca²⁺ postsynaptically E. Increased sensitivity of the postsynaptic membrane</td>
<td>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 judgment 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).</td>
</tr>
<tr>
<td>Example 6: use of higher-order cognitive skill language for lower-order cognitive skill questions</td>
<td>Predict how blood pressure changes with increasing heart rate. (But instructor taught this.)</td>
<td>Authors’ Bloom’s level: Remember or comprehend, depending on what exactly was taught. Rater (from the workshop) misstep: “Predict” goes with apply.</td>
</tr>
<tr>
<td>Example 7: multiple answers but only one solution</td>
<td>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, C, and D but not B F. B and C but not A G. A, B, and C</td>
<td>Authors’ Bloom’s level: Remember. Rater (from the workshop) misstep: Raters would reach the current question 4, which reads (“Is there more than one valid solution?”) 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).</td>
</tr>
<tr>
<td>Example 8: multiple answers but only one solution</td>
<td>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.</td>
<td>Authors’ Bloom’s level: Remember Rater (from the workshop) misstep: Same logic as example 7.</td>
</tr>
</tbody>
</table>

For the table above, 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).

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: go to 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: go to 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: go to 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: go to See Synthesize/Create
No: go to question 7.

Question 7. Are students being asked to compare/contrast information?
Yes: go to 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: go to See Evaluate
No: go to question 10.

Question 10. Are students building up a model or novel hypothesis from the data?
Yes: go to 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: go to See Analyze
No: go to question 12.

Question 12. Are students using the data to calculate the value of a variable?
Yes: go to See Apply
No: go to question 13.

Question 13. Are students redescribing the data to demonstrate they understand what the data represent?
Yes: go to 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: go to 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: go to 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: go to 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 Bloom’s 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-
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

<table>
<thead>
<tr>
<th>Prior Experience with Bloom’s Taxonomy</th>
<th>Extensive</th>
<th>Limited</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of workshop attendees with Bloom’s experience</td>
<td>4</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Mean reported difficulty level in using the BDK (1 = very easy and 5 = very difficult)</td>
<td>1.75</td>
<td>2.22</td>
<td>4</td>
</tr>
<tr>
<td>Number of attendees who would use the BDK in the future to help “Blooming” materials</td>
<td>3</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Would you recommend the BDK to others who are new to “Blooming?” Answer: yes</td>
<td>3</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>
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 commit-
ment necessary to categorize 394 homework and clicker ques-
tions 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 de-
velopment and publication of taxonomies and similarly styled
frameworks to categorize and assess student work and educa-
tion 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 assess-
ment 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 inher-
ent 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 knowl-
dge (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. How-
ever, for the two people who had had no background at all with
Bloom’s taxonomy, “Blooming” questions was still challeng-
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, indepen-
dent 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 categoriza-
tion 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 dif-
ferent 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 lev-
eels and help instructors refine the language of exam ques-
tions 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 “Bloomming” 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