INTRODUCTION

Teaching and assessing experimental-design skills

Designing experiments and applying the process of science are core competencies for biology students to develop during their undergraduate education (1, 2). To support students’ learning, educators are increasingly incorporating research experiences and other opportunities for authentic scientific skill-building into undergraduate courses, many at the introductory level. Such experiences help students learn to think like scientists, develop science-specific skills, and learn the various aspects of experimental design (3–5).

Experimental design encompasses a suite of scientific process skills, such as hypothesis generation, data collection decisions, and data analysis. Additionally, all components of an experimental design are interrelated, and internal consistency among the various components is critical. Because of the complexity of experimental design, it is no surprise that many undergraduate students, and particularly students in introductory courses, often have a surface-level understanding of experimental-design concepts and hold misconceptions about the process (reviewed in 6; 7–9). Students’ inaccurate or incomplete knowledge about experimental design can hinder their learning and limit the potential benefits of course-based skill-building and research experiences.

If students’ experimental-design abilities can be diagnosed early in their undergraduate education, then course experiences can be better designed to support student development in target areas (8). Therefore, we need to develop relevant and effective assessments of students’ experimental-design skills that align with the specific experimental-design learning objectives of a given course context (5, 10). Such assessments will allow instructors to evaluate the effectiveness of the course research and learning experiences and to offer the most effective instruction and feedback to support student development.
Authentic assessments of experimental design

Biology, similar to other scientific disciplines, is moving towards providing students with more experiences with the “process of science” and with authentic assessments as part of their undergraduate science education (1, 11). Authentic assessments are meaningful opportunities for students to integrate and apply their knowledge to a novel, complex, and/or realistic situation that simulates what individuals in the profession (in this case, scientists) may do in their work or life (12). Examples of authentic assessments used in undergraduate courses include written grant proposals or journal articles (13, 14), scientific poster presentations (15, 16), analysis of real scientific data sets (17) or case studies (18), and generating data to deposit in official scientific databases (19).

Authentic assessment is particularly important for evaluating students’ abilities in designing experiments due to the interconnected and complex nature of the components of experimental design. One approach to authentic assessment of experimental-design skills is to give students a biological scenario and ask them to design an experiment related to the scenario. An open-response experimental-design assessment format can allow instructors to gather detailed information about students’ thought processes and integrated understanding of the complex experimental-design process (6, 7, 20, 21).

Effective scoring tools for measuring experimental-design skills

The effectiveness and usefulness of any assessment of student learning depends upon the precision and transparency of the method by which it is scored. Scoring tools are more effective when they include clear and detailed criteria or expectations for students’ work (22, 23). A higher level of detail increases consistency, reliability, and objectivity with grading (23). A more detailed scoring tool also offers clearer feedback about students’ level of knowledge or skills, the specific strengths and weaknesses of the work, and areas for improvement (22, 24, 25). A clear and detailed scoring tool that is reliable across multiple scorers is particularly relevant for undergraduate introductory biology courses, as they are often taught in multiple sections by different instructors.

Scoring tools that aim to identify specific strengths and weaknesses in students’ experimental designs should score individual aspects or components of students’ learning rather than simply report a single score for overall performance (22). Scoring experimental-design skills independently, while also assessing the interconnections between skills, helps instructors to make data-driven decisions about their teaching and future instructional approaches (1, 25). Data from individual component scores can allow instructors to identify specific areas where skill-building is incomplete and also allows for tracking how students’ skills develop over time and in relation to each other.

Finally, effective scoring tools clearly differentiate levels of achievement or performance among a group of students (22). Students in introductory biology courses display a wide range of abilities based on their background experience in the sciences. Thus, any scoring tool used needs to be sensitive to variation in performance, detecting both foundational knowledge (e.g., writing a clear hypothesis statement) and recognizing the higher-level thinking required of effective experimental design (e.g., controlling for extraneous or confounding variables).

Based on the research on effective assessment, a scoring tool for introductory experimental design should do the following:

- Include clear, detailed criteria to increase interrater reliability between multiple scorers and scoring efficiency.
- Calculate separate scores for different components of experimental design (e.g., hypothesis, experimental/control groups) to provide detailed information about students’ skills in the various aspects of experimental design.
- Quantify the developmental progression of skills within the components of experimental design to assess changes in students’ skills over time.
- Assess the interrelatedness of experimental-design components to evaluate students’ ability to design an aligned experiment.

While a number of scoring tools for experimental design have been developed and shown to be effective in various settings, we were unable to identify an existing, published tool that meets all of the above desired criteria (see Appendix 1 for a comparison of scoring tools for experimental design across these criteria, and the Development of the Scoring Tool section in Methods, below, for additional details about how these tools informed our design). Consequently, we designed a novel scoring tool for experimental design that meets all of the criteria and builds on the strengths of existing tools.

Study aims

Our aim was to design a checklist tool to examine undergraduate biology students’ experimental-design skills in the context of an introductory laboratory course. Specifically, we aimed for the tool to 1) have high interrater reliability among multiple instructors teaching the same course, 2) evaluate separate components of experimental design as well as their interrelatedness, and 3) detect the range of skills among introductory students and the change in student skills that occurs during the semester. In this paper, we describe the process used to develop the Tool for Interrelated Experimental Design (TIED).

After designing the TIED, we hypothesized that it would 1) show high interrater reliability among instructors; 2) capture variability in students’ skills overall and in specific
areas of experimental design across the broad range of the scale, from scores of zero to the highest score; and 3) capture growth in some students by reporting statistically higher scores on the post-assessment compared to the pre-assessment. We used the TIED to analyze two semesters of assessment data to test these hypotheses. Finally, we describe the strengths and adaptability of this tool to other contexts and how this tool may be used to enhance biology teaching and learning.

METHODS

Course context and development of the assessment task

The context of this study was the Introductory Organismal Biology Laboratory course at Wellesley College, which is required for all biology and health science majors. Laboratory sections were limited to 16 students maximum, and one member of the Biological Sciences faculty taught each section. Elements of experimental design are introduced and practiced throughout the semester-long course. Laboratory instructors in the course were interested in developing an assessment to more authentically, consistently, and effectively measure experimental-design skills among the students and to measure any growth in skills that might occur over the course of the semester.

In January 2016, instructors began the development process by brainstorming components of introductory experimental design. Their list included hypotheses, biological rationale, experimental and control groups, data collection, data analysis, and drawing conclusions from the data collected. Instructors generated a prompt that asked students to design an experiment related to an observation about a process by brainstorming components of introductory experimental design in biology. Development of the scoring tool involved an iterative process with five phases, which are described below.

Phase 1: Reading students’ responses to detect issues with the assessment task. We removed all identifying information from students’ responses and then read through a selection of responses to get a sense of students’ level of understanding and to determine whether any of the assessment-task items were consistently misunderstood by students and should be excluded from the scoring tool. Responses to assessment-task items 3 and 9 consistently demonstrated a limited grasp of the null hypothesis, and responses to assessment-task item 7 demonstrated limited ability to state specific and appropriate statistical analyses. After consulting with course instructors, we concluded that students’ limited knowledge of null hypotheses and statistical test choices was not due to misconceptions or forgetfulness, but rather to a lack of explicit teaching on those topics in the introductory course. As a result, students were not expected to show noticeable gains in those areas and those assessment-task items were not included in the development of the scoring tool in Phase 3.

Phase 2: Gathering examples of existing tools. We searched for existing scoring tools, including checklists, rubrics, and rating scales, that assessed experimental design in any scientific discipline, and that were targeted

1. Develop and state a hypothesis for your experiment.
2. Suggest a biological rationale that supports your hypothesis.
3. State the null hypothesis.
4. What are the control group(s)?
5. What are the experimental group(s)?
6. What data will you collect, and how will you collect it?
7. What statistical analysis will you perform to compare the groups?
8. What observations would support your hypothesis?
9. What would support the null hypothesis?
for undergraduate or upper-level high school students (see Appendix I for a review of published scoring tools). While several “off-the-shelf” tools have been published (reviewed in 5), they did not meet all of our desired criteria, as previously discussed. Furthermore, the existing scoring tools were not fully aligned with our assessment-task items; they either did not contain all of the experimental-design components addressed in our assessment task or they contained additional skills that were not relevant to our task.

- Dirks and Cunningham’s (20) rubric for experimental design scenarios did not include criteria related to data or outcomes, which were essential to our assessment tool.
- The Rubric for Experimental Design (RED; 6) diagnoses specific errors that undergraduate biology students make when designing an experiment by coding whether a students’ response is correct or incorrect and, if incorrect, which type of error the student made. (This focus on specific errors was not appropriate to our task but inspired our eventual inclusion of sample correct and incorrect statements in our supplementary scoring materials.)
- The Experimental Design Ability Test (EDAT; 21) and Expanded Experimental Design Ability Test (E-EDAT; 7) focus predominantly on students’ descriptions of the variables in their experiment and include criteria that are not relevant to our assessment task, such as repetition and sample size.
- The Science Olympiad Experimental Design Checklist (26), a comprehensive checklist with 56 criteria, went beyond the scope of our assessment task.

However, we adapted some criteria and features from these tools, including several criteria from the Science Olympiad checklist (26) and the E-EDAT’s focus on evaluating student reasoning (7).

Finally, some existing scoring tools had a limited point range, which restricted the amount of possible variation in scores and limited the ability to capture change in student learning over time. For example, while the EDAT (21) has a similar purpose as our assessment tool—to assess introductory biology students’ abilities to design an experiment related to life science and detect change over time—its 10-item checklist restricts the amount of variation and change that can be detected.

**Phase 3: Developing components and criteria.**

The assessment-task items were used to develop scoring-tool components. As noted above, we did not create scoring-tool components for assessment-task items 3, 7, or 9. There were five total scoring-tool components, corresponding to the remaining six assessment-task items (Table 1).

We developed criteria for each component that identified key expectations for a successful student response. The criteria formed a checklist for each component such that a high-quality response would satisfy all of the criteria. We decided on a checklist format—criteria were either met (1 point) or not met (0 points)—so that we could precisely identify students’ strengths and gaps in their skills, thereby increasing the transparency of the scoring compared to tools where scorers select a score within a range of points. Each component had between three and five criteria.

During this phase, we recognized the importance of including criteria that focused on the interrelatedness or alignment between different experimental-design components. Other scoring tools for experimental design have included similar criteria (e.g., 6, 20, 26). For example, students’ descriptions of the control group should be evaluated in the context of their chosen experimental group(s). Therefore, we combined the experimental and control group assessment-task items into a single component and included criteria that emphasized the relationship between the groups or conditions. We made subsequent revisions to the entire scoring tool to identify whether students appropriately connected relevant experimental-design components. For example, one criterion in the Data Collection (E) component is, “Description of data collection addresses all variables stated in the hypothesis and does not introduce new variables beyond those stated in the hypothesis” and a criterion in the Observations (F) component is, “Observations described could actually be gleaned from the proposed data collection.”

Additionally, while we did not expect students to competently select specific and appropriate statistical tests for their design (see Phase 1), students in this course did learn about the importance of statistics in evaluating a hypothesis in general. Therefore, we added a single criterion to the scoring tool about mentions of statistical analysis or related terminology in the Observations (F) component: “Statement incorporates statistical terminology or analysis (e.g., p values, “statistically significant,” “positive correlation,” “significantly more/less”).”

We also made decisions about how to score incomplete responses. Items that were left blank or responses that signaled that the student did not understand (e.g., “I don’t know”, “unsure”) were assigned a score of 0, because such responses provide no information by which to evaluate students’ skills.

Students who attempted a response to a given item received one point, even if the response was incomplete or inaccurate. Thus, if a student attempted a response for each component but did not earn additional points for meeting any of the criteria, they would achieve a total score of 5. This approach of including points for attempted responses (5 points possible) can allow evaluators to glean more detailed information about students at the low end of the score distribution (especially scores from 0 to 5) and to determine appropriate instructional interventions. Evaluators can examine student responses in the 0 to 5 point range to determine which assessment-task items were responded to least often, perhaps indicating that the given experimental-
design skill was foreign or difficult for students. Additionally, responses with no attempts on any items may indicate that students were minimally motivated to complete the task, and instructional interventions could address the learning environment to increase motivation and/or foster a growth mindset among students.

Students who failed to address the prompt and instead wrote about an experiment on a different topic were not included in the analyses. This was done to ensure that assessment of student skills was standardized so that scores could be compared in a group in a given semester and/or over several semesters. For example, on the post-assessment some students described experiments designed and performed in lab during the semester rather than designing an experiment in response to the given scenario. This experimental design should not be evaluated as it does not reflect a student’s ability to independently design a novel experiment.

Phase 4: Testing and refinement. We used the draft scoring tool to score pre- and post-assessment responses from three randomly selected students (6 responses total). The authors scored the responses independently and then came together to discuss their scores on individual criteria and their rationale for the scores. The tool was improved through an iterative process of independently reading and scoring, discussing discrepancies in scores, refining tool language to make it clearer and more specific, and using students’ errors to make criteria more specific. During this process, we also tested whether the scoring tool was capturing differences in achievement levels between students at this introductory level, ensuring that the highest level of achievement (20 total points) was attainable but challenging.

After the first round of revisions, we scored an additional 11 randomly selected student responses to discuss scoring decisions and make further revisions to the tool. During this phase, we made a key decision about our approach: In our roles as course instructors, we sometimes award partial points for having some, but not all, elements of a correct response. We recognized that this was not an appropriate approach for this assessment, as our goal was to differentiate between students who demonstrated an accurate and complete understanding and students who did not. Thus, students only received a point for a given criterion if it was clearly, completely, and unquestionably achieved; zero points were awarded in all other cases. This decision increased the interrater reliability of the scoring process.

Phase 5: Training two additional scorers. We trained two additional instructors to use the scoring tool. These instructors had previously taught this course and were thus familiar with experimental-design abilities and expectations at the introductory level. Training involved introducing the instructors to the scoring tool, asking the instructors to independently score student responses, and meeting as a group with the authors to discuss consistencies and differences in scores. The new scorers first practiced scoring using the responses scored during Phase 4, after which minor revisions were made to the scoring tool. To ensure consistency in scoring, the two authors and two additional scorers then scored an additional 10 student responses that reflected a wide range of students’ experimental-design skills.

Finalized scoring tool for assessment of interrelated experimental design (TIED). The final scoring tool is provided in Figure 1. All components, with the exception of the first component (A), contain criteria that require interrelatedness or alignment with responses in a previous component. Thus, this is a tool for interrelated experimental design (TIED)—it requires experimental-design components to be “tied” together. The scoring tool contains brief instructions for scorers as well as notes within some criteria to clarify decision-making related to awarding points. It is important to note that the assessment-task prompt is interchangeable. The six assessment-task items and scoring-tool criteria are specifically designed to accommodate a variety of prompts or scenarios that set up an opportunity for students to design a biology experiment.

In addition to the scoring tool, we created two additional resources for scorers: a collection of sample student responses that did not earn points for given criteria and the rationale for the missed points (Appendix 2); and examples of a complete student response that earned full points (20 points) and a complete student response that earned points for only the “Attempts a response” criterion in each component (5 points) (Appendix 3).
Scoring and analysis of all participant responses with the TIED

Assessment responses collected from students in the Wellesley College Introductory Organismal Biology laboratory course in the spring 2016 and fall 2016 semesters were scored and analyzed. In total, 127 unique students responded to the assessment at the beginning of the semester (pre-assessment) and end of the semester (post-assessment) (66 in spring 2016 and 61 in fall 2016). All student responses were scored by the first author and two additional scorers using the TIED. Responses were scored by each instructor individually, and the scores were then compiled. Average scores were calculated to determine a total score, as well as component subscores, for each student response. Data from the spring and fall semesters were combined (all pre-assessment data combined) to increase the sample size.

JMP 13 statistical software was used to analyze distributions of pre- and post-assessment scores. Paired two-tailed t-tests were used to compare total scores, as well as component scores, between pre- and post-assessment responses.

Calculation of interrater reliability. Intraclass correlation coefficients (ICC) were used to measure interrater reliability. Selection of ICC was based on information provided by Hallgren (27). The conditions under which ICC is most appropriate were consistent with our study design. Namely, our study has two or more coders, and all student responses were rated by multiple coders (19). Additionally, the ICC is more appropriate when the data are ordinal, interval, or ratio, while kappa is better suited to nominal or categorical data (27, 28).
To calculate ICC, we used the R irr package. We used a two-way, mixed, average-measures ICC to assess the degree to which coders provided consistency in their ratings across students. We specified a two-way model because all subjects were rated by all coders, a mixed model because coders were not randomly selected, and average-measures because we were interested in the consistency among coders and because all student responses were rated by multiple raters (27). ICC values are on a scale of 0 to 1.00, with a score of 1.00 representing perfect agreement among raters and scores above 0.75 indicating excellent agreement (29).

RESULTS AND DISCUSSION

Does the TIED have high interrater reliability across different instructors?

The interrater reliability for the assessment total score was excellent, with an ICC of 0.945 (95% confidence interval [CI] 0.933–0.955) indicating a high degree of agreement across coders and a minimal amount of measurement error. There was also a high level of interrater agreement for each of the five individual components, with all ICC values at or above 0.839 (Table 2). High interrater reliability is particularly important for assessments in undergraduate introductory biology courses as these courses are often taught in multiple sections by different instructors. Additionally, if the tool is used for program evaluation over the course of several years, high interrater reliability is crucial.

TABLE 2. Average interrater reliability of three raters of the individual components of the TIED for evaluating students’ experimental-design skills.

| TIED Component                      | ICC  | 95% CI       |
|-------------------------------------|------|--------------|
| Total Score                         | 0.945| 0.933–0.955  |
| A. Hypothesis                       | 0.938| 0.925–0.950  |
| B. Biological Rationale             | 0.857| 0.827–0.883  |
| C/D. Experimental and Control Groups| 0.855| 0.823–0.881  |
| E. Data Collection                  | 0.839| 0.805–0.868  |
| F. Observations                     | 0.879| 0.853–0.901  |

Does the TIED capture variation in students’ abilities to design an experiment?

Students’ total scores ranged from 2.67 to 19.33 on the pre-assessment and 6.67 to 19.66 on the post-assessment (Fig. 2). Students’ responses earned the full range of possible scores for each component, with the exception of the Hypothesis component, in which there were no scores of 0 for any response. This means that all students in our sample attempted a response for the hypothesis component, while some students wrote “I don’t know” or left their response blank for the other components. Students who attempted a response for each component but did not earn additional points for meeting any of the criteria achieved a total score of 5 on the scoring tool (see Appendix 3 for example). Based on these results, the scoring tool is sensitive enough to capture variation in skills among students at the introductory level, with few students hitting the ceiling or floor on the assessment.

Students generally scored higher on the Hypothesis and Biological Rationale components compared to the other components (Table 3). Because the hypothesis was the first assessment-task item that students had to respond to in the assessment task, they did not have to connect their hypothesis to any other component of their experimental design. Students could therefore achieve a high score on this component with an unrealistic or biologically irrelevant hypothesis, as long as their response followed the correct format of a hypothesis. However, the variation in students’ scores on the Hypothesis component, despite the limited number of points possible, demonstrates that the scoring tool captures differences in students’ skills, even for this more simplistic component.

Subsequent sections of the scoring tool required students to appropriately align their responses to those in the previous components in order to achieve a high score. For example, students’ Observations (F) had to be connected back to their Hypothesis (A) and their description of Experimental and Control groups (C/D) in order to achieve full points on that component. The data showed the most variation in students’ scores in the Data Collection (E) and Observations (F) components, which contained more possible points and challenged students to design an aligned experiment.

FIGURE 2. Histograms representing the number of students achieving each total TIED score on the pre- and post-assessment.
Does the TIED capture change in students’ abilities to design an experiment?

The distribution of students’ total scores shifted towards the higher end on the post-assessment compared to the pre-assessment (Fig. 2). When comparing the pre- and post-assessment distributions, movement in scores is most evident in the lower end of the distribution, with a general right-shift in scores on the post-assessment. The median score on the pre-assessment was 14.67 and on the post-assessment was 15.33. Almost half of students (44.9%, \(n = 57\)) scored at or above 80% (≥16 points) on the post-assessment, compared to 35.4% (\(n = 45\)) of students who achieved such scores on the pre-assessment (Fig. 2). At the lower end of the scoring range, two students scored 5 or fewer points on the pre-assessment, while the lowest score on the post-assessment was 6.67 (Fig. 2).

Mean scores on the post-assessment were higher for all components, as well as for the overall total score, compared to the pre-assessment scores (Table 3). There was a significant difference between pre- and post-assessment scores for the total score (\(t(126) = 2.8, p = 0.006\)). There were also significant differences between pre- and post-assessment scores for the Data Collection component (E; \(t(126) = 2.26, p = 0.025\)) and the Observations component (F; \(t(126) = 3.18, p = 0.002\)) (Table 3).

Strengths of the assessment task and the TIED

In this study, we developed an assessment task and scoring tool (TIED) that successfully assessed undergraduate introductory biology students’ experimental-design skills. In particular, the structured, open-response format of the assessment revealed students’ thought processes and reasoning across key components of experimental design. The assessment tool was quick to implement in class, and because of the structured-response format and checklist-style scoring tool, the scoring of responses was also efficient.

The TIED captured variation in students’ skill levels in an introductory course and was sensitive to changes in students’ skills over a semester. The scoring tool was rigorous and appropriate for assessing the skills in our population of introductory biology students, with few students hitting either the floor or ceiling scores. Training instructors to use the scoring tool was relatively quick, and multiple scorers were able to reach consensus and high interrater reliability.

The TIED was also designed to work with any experimental design prompt in biology and is not in any way restricted to the prompt we used about guppies’ foraging behavior. Another strength of our scoring tool is that it requires students to align their decisions about their hypothesis, experimental and control groups, data collection, and data observations. Cohesive and aligned experimental design is essential for effective scientific practice, and our scoring tool assesses how well students are developing this particular ability in addition to the individual component skills involved in experimental design.

Considerations for adapting the TIED to other contexts

We have designed this assessment and scoring tool so that it can be used in a variety of introductory biology contexts. Adapting this assessment tool to other contexts may require additional effort and edits and we offer suggestions and considerations for how to do so.

### TABLE 3.
Mean scores on each TIED component and results of paired, two-tailed, \(t\)-tests comparing pre- and post-assessment responses for individual students.

| TIED Component                      | Possible Points | Pre-Assessment M (SD) | Post-Assessment M (SD) |
|-------------------------------------|-----------------|-----------------------|------------------------|
| Total Score                         | 20              | 13.76 (3.46)          | 14.82 (3.02)*          |
| A. Hypothesis                       | 4               | 3.34 (0.92)           | 3.40 (0.85)            |
| B. Biological Rationale             | 3               | 2.42 (0.74)           | 2.55 (0.55)            |
| C/D. Experimental and Control Groups| 3               | 2.01 (0.82)           | 2.19 (0.71)c           |
| E. Data Collection                  | 5               | 3.28 (1.05)           | 3.57 (0.89)a           |
| F. Observations                     | 5               | 2.71 (1.16)           | 3.12 (0.99)a           |

\* \(p < 0.01\).
\* \(p < 0.05\).
\* \(p < 0.08\).
This tool was designed to work with different assessment prompts. We selected a biological scenario that had not been previously introduced to students to allow for many possible hypotheses as a starting point for experimental design. We therefore did not penalize students who had a well-designed experiment about a hypothesis based on inaccurate knowledge of guppies themselves. However, some instructors may want to choose a scenario that is similar to one encountered by students during the semester or that requires students to demonstrate content knowledge from a course in their experimental design.

Alignment of the assessment task and scoring tool to the course learning outcomes and goals is essential (5, 10). For example, it was not a key learning outcome for our students to demonstrate knowledge about feasible data collection methods and specific data collection equipment, tools, or reagents. Therefore, we did not penalize student responses if their stated data collection method would be inefficient, wasteful, or impractical based on time, money, or personnel. However, if some courses emphasize feasibility of data collection or understanding of specific techniques, the scoring criteria should be modified to assess those learning outcomes. Additionally, instructors may be interested in assessing other experimental-design skills, such as students’ knowledge about sample size, the null hypothesis, statistical choices, or replication. If additional components are needed, we recommend reviewing other existing experimental-design assessment tools that contain these components (e.g., 7, 26), and adding criteria or components to the TIED as needed.

The TIED was also specifically designed for introductory undergraduate students. Because we focused on introducing students to the practice of designing an aligned experiment, students who were able to design a simple but consistent experiment across all components achieved a higher score than students who attempted to design a more complex experiment (e.g., with multiple independent variables in their hypothesis) but were unable to align each component within their experiment (e.g., to successfully test multiple variables). As a result, the highest score possible on this scoring tool does not represent expert-level experimental design skills, but rather the skills we can expect of a high-performing novice undergraduate biology student. The scoring-tool criteria may need to be adapted for use with students at a different educational level from whom more complex or sophisticated experimental designs may be expected.

Implications for teaching and learning

This tool can be used by instructors and departments to inform both instruction and curriculum. For example, instructors can use this tool as a pre/post assessment or as a standalone assessment of students’ experimental-design skills. Using this tool at the beginning of a course can provide useful information to instructors about students’ skills across the various components of experimental design, and about areas in which students require more instruction or practice. As a result, instructors can better focus their teaching efforts to support students in their areas for growth. When used at the end of a course or the end of a unit on experimental design, data can be used to evaluate the strengths and gaps in students’ understanding of experimental design. When administered at multiple times (e.g., pre/post), the data collected can be used to detect change in students’ skills over time and evaluate the effectiveness of the course with respect to teaching specific experimental-design skills. This information can help instructors make decisions about future iterations of the course, and in particular, which skills require more support.

This tool could also be used to provide feedback to students about their experimental-design skills. Scores could be communicated back to individual students to help them identify their strengths and areas to improve, as well as how much they have learned about experimental design during the course.

Departments and programs can also use this tool for curriculum review and improvement. This tool can be used to assess how well a program is supporting students’ learning of key knowledge and skills related to experimental design. Data can be used as a basis for departmental discussions about curricula, particularly with respect to where these skills are taught and reinforced throughout a curriculum, and how well students are learning these skills through the current curriculum. At the program level, this tool can be used in multiple courses to track students’ learning longitudinally, to identify how well program learning outcomes are being achieved, and to highlight gaps in students’ development of these skills.

SUPPLEMENTAL MATERIALS

Appendix 1: Comparisons of the TIED with other published scoring tools for experimental design

Appendix 2: Examples that do not earn credit for a given TIED criteria

Appendix 3: Examples of complete student responses that earned full points or only “Attempts a response” points based on TIED criteria

ACKNOWLEDGMENTS

Our research protocols were reviewed and deemed Exempt by the Wellesley College IRB under §46.101b Exemption 1 and Exemption 2. We thank Julie Roden, Christa Skow, Jocelyne Dolce, Janet McDonough, and Jeff Hughes for their support as instructors in designing and implementing the assessment task. We thank Julie Roden and Leah Okumura for serving as scorers. We thank anonymous reviewers for their valuable feedback on previous versions of this manuscript. This project was supported by an internal institutional grant from the Andrew W. Mellon Foundation for scholarship of teaching and
learning and assessment projects. Funding supported stipends for faculty scorers and a stipend to TK for development of the scoring tool and dissemination of the findings. The authors declare that there are no conflicts of interest.

REFERENCES

1. American Association for the Advancement of Science. 2011. Vision and change in undergraduate biology education: a call to action: a summary of recommendations made at a national conference organized by the American Association for the Advancement of Science, July 15–17, 2009. Washington, DC. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf.

2. Coil D, Wenderoth MP, Cunnigham M, Dirks C. 2010. Teaching the process of science: faculty perceptions and an effective methodology. CBE Life Sci Educ doi:10.1187/cbe.10-01-0005.

3. Jeffery E, Nomme K, Deane T, Pollock C, Birol G. 2016. Investigating the role of an inquiry-based biology lab course on student attitudes and views toward science. CBE Life Sci Educ 15:ar61.

4. Shortlidge EE, Bangera G, Brownell SE. 2016. Faculty perspectives on developing and teaching course-based undergraduate research experiences. BioScience 66:54–62.

5. Shortlidge EE, Brownell SE. 2016. How to assess your CURE: a practical guide for instructors of course-based undergraduate research experiences. J Microbiol Biol Educ 17(3):399–408.

6. Dasgupta AP, Anderson TR, Pelaz N. 2014. Development and validation of a rubric for diagnosing students’ experimental design knowledge and difficulties. CBE Life Sci Educ 13:265–284.

7. Brownell SE, Wenderoth MP, Theobald R, Okoroafor N, Koval M, Freeman S, Walcher-Chevillet CL, Crowe AJ. 2014. How students think about experimental design: novel conceptions revealed by in-class activities. BioScience 64(2):125–137.

8. Deane T, Nomme K, Jeffery E, Pollock C, Birol G. 2014. Development of the biological experimental design concept inventory (BEDCI). CBE Life Sci Educ 13(3):540–551.

9. Shi J, Power JM, Klymkowsky MW. 2011. Revealing student thinking about experimental design and the roles of control experiments. Int J Scholarsh Teach Learn 5:1–16.

10. Cooper KM, Soneral PAG, Brownell SE. 2017. Define your goals before you design a CURE: a call to use backward design in planning course-based undergraduate research experiences. J Microbiol Biol Educ 18(2):1–7.

11. National Research Council. 2003. BIO2010: Transforming undergraduate education for future research biologists. The National Academies Press, Washington, DC. doi:10.17226/10497.

12. Wiggins, G. 1998. Ensuring authentic performance, p 21–42. In Educative assessment: designing assessments to inform and improve student performance. Jossey-Bass, San Francisco, CA.

13. Oh DM, Kim JM, Garcia RE, Krilowicz BL. 2005. Valid and reliable authentic assessment of culminating student performance in the biomedical sciences. Adv Physiol Ed 29 (2):83–93.

14. Kloser Mj, Brownell SE, Chiariello NR, Fukami T. 2011. Integrating teaching and research in undergraduate biology laboratory education. PLOS Biol 9(11):e1001174.

15. Dogan A, Kaya ON. 2009. Poster sessions as an authentic assessment approach in an open-ended university general chemistry laboratory. Procedia Soc Behavior Sci 1(1):829–833.

16. Laungani, R., Tanner C, Brooks TD, Clement B, Clouse M, Doyle E, Dworak S, Elder B, Marley K, Schofield B. 2018. Finding some good in an invasive species: introduction and assessment of a novel CURE to improve experimental design in undergraduate biology classrooms. J Microbiol Biol Educ 19(2):19.2.68. doi:10.1128/jmbe.v19i2.1517.

17. Makarevitch I, Frechette C, Wiatros N. 2015. Authentic research experience and “big data” analysis in the classroom: maize response to abiotic stress. CBE Life Sci Educ 14(3):ar27.

18. Dolan EL, Collins JP. 2015. We must teach more effectively: here are four ways to get started. Mol Biol Cell 26(12):2151–2155. doi:10.1091/mbc.e13-11-0675.

19. Jordan TC, Burnett SH, Carson S, Caruso SM, Class K, Delong Rj, Dennehy Jj, Denver DR, Dunbar D, Elgin SCR, Findley AM, Gissendanner CR, Golebiewska UP, Guild N, Hartzog GA, Grillo WH, Hollowell GP, Hughes LE, Johnson A, King RA, Lewis LO, Li W, Rosenzweig F, Rubin MR, Saha MS, Sandoz J, Shaffer CD, Taylor B, Temple L, Vazquez E, Ware VC, Barker LP, Bradley KW, Jacobs-Sera D, Pope WH, Russell DA, Cresaw SG, Lopatto D, Bailey CP, Hatfull GF. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio 5(1):e01051-13.

20. Dirks C, Cunningham M. 2006. Enhancing diversity in science: is teaching science process skills the answer? CBE Life Sci Educ 5:218–226.

21. Sirum K, Humburg J. 2011. The experimental design ability test (EDAT). BioScience 37(1):8–16.

22. Brookhart SM. 2013. How to create and use rubrics for formative assessment and grading. ASCD, Alexandria, VA.

23. Peat B. 2006. Integrating writing and research skills: development and testing of a rubric to measure student outcomes. J Public Aff Educ 12(3):295–311.

24. Andrade HG. 2000. Using rubrics to promote thinking and learning. Educ Leadersh 57:13–18.

25. Kishbaugh TLS, Cessna S, Horst SJ, Leaman L, Flanagan F, Neufeld DG, Siderhurst M, Sandoz J, Shaffer CD, Taylor B, Temple L, Vazquez E, Ware VC, Barker LP, Bradley KW, Jacobs-Sera D, Pope WH, Russell DA, Cresaw SG, Lopatto D, Bailey CP, Hatfull GF. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio 5(1):e01051-13.

26. Science Olympiad. 2017 Experimental design checklist for B/C. Retrieved from https://www.soinc.org/sites/default/files/uploaded_files/ExDesChecklist17Final.pdf.

27. Hallgren KA. 2012. Computing interrater reliability for observational data: an overview and tutorial. Tutor Quantitat Meth Psychol 8:23–34.

28. Mandrekar J. 2011. Measures of interrater agreement. Biostat Meas Meth Psychol 6:23–34.

29. Cicchetti DV. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6:284–290.