CBE—Life Sciences Education
Vol. 13, 540–551, Fall 2014

Article

Development of the Biological Experimental Design Concept Inventory (BEDCI)

Thomas Deane,* Kathy Nomme,* Erica Jeffery,* Carol Pollock,† and Gülnur Birol‡

*Departments of Botany and Zoology, Biology Program, Faculty of Science, †Department of Zoology, Faculty of Science, and ‡Science Centre for Learning and Teaching, Faculty of Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Submitted November 8, 2013; Revised June 16, 2014; Accepted June 19, 2014
Monitoring Editor: Mary Pat Wenderoth

Interest in student conception of experimentation inspired the development of a fully validated 14-question inventory on experimental design in biology (BEDCI) by following established best practices in concept inventory (CI) design. This CI can be used to diagnose specific examples of non–expert-like thinking in students and to evaluate the success of teaching strategies that target conceptual changes. We used BEDCI to diagnose non–expert-like student thinking in experimental design at the pre- and posttest stage in five courses (total $n = 580$ students) at a large research university in western Canada. Calculated difficulty and discrimination metrics indicated that BEDCI questions are able to effectively capture learning changes at the undergraduate level. A high correlation ($r = 0.84$) between responses by students in similar courses and at the same stage of their academic career, also suggests that the test is reliable. Students showed significant positive learning changes by the posttest stage, but some non–expert-like responses were widespread and persistent. BEDCI is a reliable and valid diagnostic tool that can be used in a variety of life sciences disciplines.

INTRODUCTION

Fundamental to all the sciences is the underlying process of scientific investigation, often referred to as "the scientific method" in educational settings. Students may recognize the general steps involved in scientific investigation, but often have only a surface understanding of the process; this is not surprising, given students are taught to mimic the steps of the scientific method in elementary and secondary school science classes as if following a recipe (Aikenhead and Ryan, 1992). Most students have experience performing experiments and some experience in designing experiments by the time they arrive at college or university, yet many students beginning an undergraduate program have a poor understanding of the basic, even fundamental concepts that must be applied to design reliable, valid experiments (Hiebert, 2007; Pollack, 2010). As a result, first-year college and university students often believe that, by following instructions, the experiments they perform should provide "right" answers that will reinforce the information and theory presented in class (Sere *et al.*, 2001).

Moving students past this sort of dualistic thinking, in which there is either a "right" or "wrong" answer, can be especially challenging (Perry, 1981). Their prior knowledge about experimentation, even if it is flawed, will become the basis for constructing further knowledge. It is now recognized, however, that many non–expert-like conceptions can play a potentially productive role in forming expert-like thinking and should not necessarily be regarded as "misconceptions" that need to be displaced (Smith *et al.*, 1993, Maskiewicz and Lineback, 2013; although see Crowther and Price [2014] and Leonard *et al.* [2014] on whether the term "misconceptions" is problematic). If incorrect or naïve conceptions are held strongly by students, it is not likely their non–expert-like thinking will evolve into expert-like thinking without focused instruction. In our experience, many undergraduates

DOI: 10.1187/cbe.13-11-0218
Address correspondence to: Kathy Nomme (nomme@zoology.ubc.ca).

demonstrate non–expert-like thinking regarding experimental design when they first arrive at our university, and this is consistent with findings at other research universities (Dasgupta *et al.*, 2014). Much of this thinking is persistent and may impede student learning well into third and fourth year and beyond. Indeed, non–expert-like thinking can even persist in graduate students, instructors, and researchers (Zolman, 1999; Festing, 2003). This can lead to poorly designed experiments, which in turn devalue the conclusions made in many peer-reviewed articles. For example, the appropriate application of independent sampling continues to be a very difficult concept for researchers (Hurlbert, 1984; Lazic, 2010). This example underscores the importance of diagnosing non–expert-like conceptions about experimental design and creating learning opportunities to shift student thinking toward more expert-like conceptions.

Hands-on experience in the design of experiments has been widely recognized as an effective means of teaching experimentation and as a critical component of undergraduate science education (Cummins *et al.*, 2004; Adams, 2009; Roy, 2013). In the early stages of the biology program at our university, students complete at least one laboratory skills course in which they design and conduct their own experiment before analyzing the data and synthesizing their work by writing a report in the form of a journal article. As they progress through their studies, there are opportunities to design experiments in subsequent courses. The need to reveal non–expert-like conceptions relating to experimental design led us to the search for an appropriate assessment tool. Other researchers have investigated non–expert-like thinking in biology using various means. Shi *et al.* (2011) used a series of multiple-choice and open-ended questions to investigate student understanding of experimentation and found that undergraduates often struggled to correctly design and understand the importance of controls in experiments. Following a redesign of a cell biology course, they found evidence that student learning of these concepts had improved. Sirum and Humburg (2011) developed the Experimental Design Ability Test, which measures the understanding and application of fundamental concepts of experimental design by scoring open-ended responses to a prompt, while Dasgupta *et al.* (2014) developed a rubric that can be used to measure specific difficulties students have with experimental design. These tools are useful in the characterization of student thinking in conceptual areas related to experimental design. However, it was the need to assess student thinking on the essential elements of experimental design in large or multisection courses that primarily led us to develop a comprehensive set of multiple-choice questions. The Biological Experimental Design Concept Inventory (BEDCI) included questions related to: 1) the need for controls, 2) the philosophy of hypothesis testing, 3) the importance of limiting biological variation, 4) the importance of accuracy in affecting conclusions, 5) the need to control for extraneous factors, 6) the importance of independent replication, 7) the purpose of random sampling, and 8) the purpose of doing experiments. Each of these interrelated concepts provides valuable information regarding the student learning experience in our courses. By using the BEDCI, instructors should be able to identify concepts their students need to master, as well as the specific examples of non–expert-like thinking that are affecting their understanding in a negative way.

METHODS

Development of the BEDCI

We followed a multistep process similar to ones used by other concept inventory (CI) developers (Adams and Wieman, 2010) in designing and developing the BEDCI; we first sought information from the literature (see Anderson-Cook and Dorai-Raj, 2001; Green and Bozzone, 2001; Hiebert, 2007; Coil *et al.*, 2010; Shi *et al.*, 2011; Sirum and Humburg, 2011) in drawing up a list of common examples of non–expert-like thinking relating to experimental design in biology. As the next step, we consulted faculty members and graduate students who were associated with biology lab courses and asked for examples of common non–expert-like thinking based on their experiences. We also looked for examples of non–expert-like thinking in previous student assignments and exams from a first-year lab course taken at our university. The exam is made up of questions requiring students to design experiments based on fictional scenarios and to answer related questions requiring short answers. In addition students are required to answer questions involving simple statistical analysis and interpretation of data generated from the fictional experiments. Based on the literature search, consultations with faculty members, and a sample of 60 exams, a rudimentary rubric was developed to categorize examples of non–expert-like thinking (errors in student answers) into eight distinct core concepts (see Table 1). We then met with five faculty members (all of whom had taught, or were teaching, the first-year lab course) and asked them to list examples of non–expert-like student thinking they had observed that pertained to the eight conceptual areas. For example, in relation to the core concept Controls, we asked questions such as: "What sort of mistakes are common when students are asked to select a control group?," "What do students think the purpose of a control group is?," and "How do students interpret the results of experiments with no control group?" We used student responses to open-ended interview questions framed around the concepts highlighted in Table 1, as well as the incorrect answers from the previous student assignments and exams, to draft multiple-choice questions. These questions included the expert-like thinking as one answer along with three distractors that featured different non–expert-like thinking. We initially had unequal numbers of questions spanning the eight core concepts (three questions for each of core concepts numbers 1–3 and 5, and two questions for each of core concepts 4 and 6–8). This was because we created more questions for the concepts that were either 1) linked more frequently to non–expert-like thinking in student exams or 2) identified by faculty members as concepts that were more important for students to master if they were to gain a fundamental understanding of experimental design.

Contextualizing questions is thought to help test conceptual application of knowledge by fostering greater student interest (see Rivet and Krajcik, 2008), and student reasoning is often connected to context (Mortimer, 1995). For this reason, we developed three simple, easy-to-understand scenarios; each scenario ("Growth of Rainbow Trout," "Tomato Plant Fertilizers," and "Invasive Cheatgrass Management") set the context for multiple questions, thereby minimizing reading for students (see Table 1). We note that Nehm and Ha (2011) found that the specific organism used to contextualize

Table 1. The distribution of concept questions and experimental design scenarios among the eight core concepts of the final, validated BEDCI

Concept	Questions examine student understanding/thinking of:	BEDCI question	BEDCI scenario[a]	Expert agreement (%)[b]
Controls	How to design suitable controls, and why controls are required in experiments.	1	A	100
		5	B	94.4
Hypotheses	How to design suitable hypotheses, and how many hypotheses can be assessed in different experiments.	2	A	100
		9	C	100
Biological Variation	Which factors are expected to vary between and within individuals, and how these affect experiments.	3	A	100
		10	C	94.4
Accuracy	How the accuracy of results can be improved.	4	A	94.4
Extraneous Factors	Which factors should be controlled, and how noncontrolled factors affect conclusions.	6	B	88.9
		14	C	100
Independent Sampling	How to design sampling techniques for experiments so that individual replicates are only sampled once.	7	B	88.9
		12	C	83.3
Random Sampling	Why replicates should be sampled randomly, and how other factors affect the suitability of the technique.	8	C	94.4
		13	C	100
Purpose of Experiments	Why we conduct experiments and/or what makes them successful/useful.	14	C	100

[a] Scenario A: "Growth of Rainbow Trout"; B: "Tomato Plant Fertilizers"; C: "Invasive Cheatgrass Management."

[b] The expert agreement shows how many of our experts ($n = 18$ total, $n = 12$ faculty members, $n = 6$ graduate students) answered this question in an expert-like way at the final stage of validation.

open-ended assessment items influenced the way students approached questions of natural selection, so, when constructing BEDCI scenarios, we purposefully chose simple experimental settings and also provided students with handouts of background information about the organisms.

BEDCI (Construct) Validation: Additional Student Interviews, Online Implementation, and Final Faculty Validation

We held one-on-one, think-aloud interviews with a total of 29 undergraduate students (18 interviews using the original 20-question instrument; 11 interviews with the revised, 15-question instrument). All students ($n = 19$ female; $n = 10$ male) were enrolled in first- to fourth-year biology courses when interviewed, and the vast majority were studying in the Faculty of Science ($n = 25$). We wanted to develop BEDCI with the capability to assess a range of undergraduates, so we interviewed a range of students at different stages of their undergraduate studies ($n = 18$ first years, $n = 6$ second years, $n = 4$ third years and $n = 1$ fourth year) and with different levels of experience with English as a first language. We wanted to hear from students with a range of knowledge, views, and experiences that might affect the way they interpreted and answered the questions, but all of the students had completed at least one undergraduate biology course at our university by the time they were interviewed, so the subject material was not unfamiliar. Incoming science students at this university are accepted with very high secondary school averages (92.5%) and our first-year participants (the majority of the sample, $n = 18$) represented this pool. We used in-class and online announcements as a means of attracting participants, who were paid $15 for participating in a 50–60 min interview. We ensured that each question was answered and discussed by at least 20 students from its creation to final version (the BEDCI in its final form was worked through by at least eight students). The most common improvements made to BEDCI questions involved rephrasing questions to eliminate jargon and rewording answers using vocabulary that students expressly stated they would use (thus enabling us to create more effective distractors; see Supplemental Table S1).

The validation step is crucial in the design of robust CIs, and we closely followed the advice of Adams and Wieman (2010). Briefly, we asked students to read aloud each question and its four answers before asking them to explain the reasoning for their choice so as to confirm they were interpreting the answer as we intended. We made audio recordings of student interviews, so we could revisit them afterward. Student interview responses led to us frequently improving questions and/or answers in the early stages (see Table S1).

We administered a version of the original 20-question instrument as a pretest (January 2012) and a posttest (April 2012) in one section of a first-year lecture course required for biology majors and other life sciences programs. To encourage students to take the tests seriously, we offered a 0.5% bonus mark for completing the pre- and posttest (for a total of 1% that would be added to their course grade). Respondents answered the questions in sequence and were not restricted by a time limit, but we reasoned that the 20 questions should be completed within 15–60 min, and so discarded responses from outside that range ($n = 11$). We also incorporated a dummy question to ensure students were reading the questions and not just submitting random answers to procure bonus marks. After removing any such respondents from the sample ($n = 4$), we had a total of 96 students responding to and completing both tests, and their matched responses were then analyzed as a way of assessing and guiding the development of BEDCI.

Questions were eliminated if they were not answered in an expert-like way by 25–75% of students, because those with responses in this range are best able to show learning changes when pre- and posttest performance is compared (Kline, 2005).We eliminated two questions that were very easy (i.e., ≥ 85% expert-like student responses on the pretest; Smith *et al.*, 2008). One discarded question regarded core concept 3 (the importance of limiting biological variation), which 95.8% of students answered correctly on the pretest, and the other question regarded core concept 5 (the need to control for extraneous factors), which 89.6% of students answered correctly on the pretest. We also eliminated questions that did not discriminate between the high-performing and low-performing students, as indicated by overall BEDCI score. Three questions that did not discriminate sufficiently between high and low-scoring students: one for core concept 1 (the need for controls), one for core concept 2 (the philosophy of hypothesis testing), and one for core concept 4 (the importance of accuracy in affecting conclusions) were removed. After removing these questions, we were left with a 15-question version of BEDCI.

Before experts were solicited to review the remaining questions, we conducted a further 11 student interviews and made a few subtle improvements to the questions based on student responses. Experts consisting of 12 faculty members and six graduate students with various backgrounds in experimental design agreed to answer the BEDCI questions. The participating research faculty members were experts in zoology (5), botany (4), and physiology (3). All of them had PhD qualifications and were instructing undergraduate biology courses at the time. The six graduate students who worked through the BEDCI questions qualified as experts, because they were enrolled in either the MSc ($n = 4$) or PhD ($n = 2$) program in the department of zoology and had been teaching assistants in a first-year lab course at least once. These experts were asked to choose the answer they considered expert-like and were also encouraged to provide alternative answers and comment on the suitability of all answers.

Any questions that had ≤ 80% answer agreement among our experts, including suggested changes to wording or answers, were removed. This level of expert agreement is consistent with Smith *et al.* (2008), who reported that >80% of their experts agreed that all but one of their 25 Genetics Concept Assessment (GCA) questions tested the learning goal it was designed to assess. They concluded that this was sufficient validity evidence and supported the accuracy of their assessment's items. Naegle (2009) reported that experts scored 85.5% on a phylogenetic tree–thinking CI, suggesting that a similar level of agreement should be sought in other CIs. In the event that experts proposed even slightly different answers to any of those provided, we asked them to comment on the suitability of the question (and its available answers). In all but one case, experts did not think there were any associated problems. We were forced to discard one question (core concept Purpose of Experiments), which was answered in an expert-like way by 13 of the 18 experts (72.2%); in addition to five of our experts choosing what we considered a non–expert-like answer, another three suggested alternative wording to the answer we considered expert-like. Thirteen of the other 15 questions we gave our experts had ≥ 88.9% expert agreement (i.e., they were answered correctly by 16–18 of our experts), and the other question was answered correctly by 15 of the 18 experts (83.3%). Two experts suggested slight wording changes to this question, but 16 did not, so we did not revise it. Following this final step, we had our fully validated 14-question BEDCI, which had a mean expert agreement of 95.6% ($n = 18$).

Implementation of the BEDCI

To assess the suitability and sensitivity of the BEDCI with students at various stages of their undergraduate careers, we administered it in five biology classes at our university (Table 2). The number of classes available to participate in the study was limited, because not all instructors were willing to sacrifice class time for both the pre- and posttests. The classes that participated included four sections of a first-year biology lecture course and one section of a third-year biology laboratory course (third-year laboratory). Students in the first-year biology classes were in one of two specialized programs (either the first-year coordinated program or the first-year science program), or they were in either term 1 or term 2 of the first-year general program. This assortment of classes provided a degree of diversity in year of study and university experience among the students we sampled.

The delivery of learning objectives relating to experimental design varied among the five classes included in our study. Additionally, the eight concepts incorporated in BEDCI are not the domain of any one specific biology course, but subsets of the concepts addressed by each course. A review of course materials from the five classes included in this study indicates that many of the eight concepts addressed by the BEDCI were either taught implicitly or taught and assessed explicitly in course homework, quizzes, and exams (Table 3).

Pooling all pre- and posttest respondents across the five classes, we had a total matched sample size of 580 students who had given their consent to be part of the study. The BEDCI was administered in the same way in all five classes; students were given a bonus participation mark (0.5% bonus mark for completing each of the pre- and posttests) but were otherwise neither rewarded nor penalized for their performance. Students were shown BEDCI questions in class on a series of PowerPoint slides, which instructors deployed consistently (timing was set for each question and students were warned before the next one appeared). Following this protocol meant that we administered the BEDCI in 18 min and 30 s in each of the five classes. We chose to administer BEDCI in each class, rather than online, to ensure a more controlled environment and to emphasize the importance of the exercise to students.

Statistical Analyses

We developed BEDCI to detect shifts in conceptual learning across a range of courses and students. In this study, we analyzed student responses from five classes but look forward to comparing results with other courses at other institutions.

Table 2. Course descriptions for the five classes used in the BEDCI pre- and posttests[a]

Course name and description	Typical course enrollment	Term	*n*
First-year general: a multisection, lecture-only course that introduces ecology, evolution, and genetics to first-year students; it is a required course in the biology program and in some other faculties (e.g., forestry). Assessment is based on midterm and final exams, plus assignments and participation in clicker quizzes.	∼ 2000 per year; ≤ 225 per section; 10 sections per year; offered term 1 and term 2	1* 2*	170 160
First-year coordinated: students self-select for this cohort program in which they attend the same sections of core science courses (lab, lecture, workshop) in biology (first-year general), chemistry, physics, and math. Assessment is based on midterm and final exams, plus assignments and participation in clicker quizzes.	∼ 160 per year; 1 section per year; offered term 2 only	2*	110
First-year science: a cohort program similar to first-year coordinated, except that students apply and are screened for admittance based on high performance in high school science and English courses. The biology lessons are integrated with other science courses, and small-class tutorials reinforce integrated concepts. Assessment is based on midterm and final exams, plus weekly tutorial assignments.	∼ 75 per year; 1 section per year; offered term 1 and term 2	1*	57
Third-year laboratory: a lab skills–based course, with lectures integrating topics from the ecosystem-level investigation of organisms to molecular techniques and model organism studies. Not intended for biology majors but open to students with a third-year standing or higher in the Combined Major in Science program. First-year general is a prerequisite. Assessment is based on written lab reports plus oral presentations.	∼100 per year; 4 lab sections and 1 lecture section per year; offered term 1 only	1****	83

[a] Data were collected from two different classes of first-year general students. Term 1: Fall term (September–December 2012); term 2: Winter term (January–April 2013). *, Data collected from one section; ****, data collected from four sections.

Each class in this current study differed either substantially or subtly. Of the five different classes, four were first-year classes, and one was a third-year class. The four first-year classes were lecture based while the third-year class was laboratory based. The first-year classes differed in that we sampled two populations of first-year general students (terms 1 and 2) and two cohorts of students in specialized programs (first-year coordinated and first-year science). Although subtle differences existed between classes, we suspected that students in term 1 would hold more non–expert-like conceptions about experimental design, while term 2 students would have had some time to adjust to university and be influenced by their learning experiences during the previous term. Additionally, all five classes were taught by different instructors. To evaluate the precision and sensitivity of the BEDCI, we calculated 1) a measure of item difficulty to ensure that questions represented a variety of difficulty levels (i.e., no question was too easy or too difficult), 2) a measure of discrimination to ensure the tool was sensitive enough to diagnose non–expert-like thinking in students of different abilities, and 3) a measure of reliability to ensure responses were broadly stable among similar student populations (see Ding *et al.*, 2006; Smith *et al.*, 2008; Adams and Wieman, 2010). Because we were interested in the extent to which the BEDCI questions were able to discriminate the full ability range of the target population (first- to third-year undergraduate biology students), we conducted the majority of our analyses using the pooled data set ($n = 580$).

Table 3. Differences in the way concepts in experimental design were taught and assessed in the courses surveyed with the BEDCI[a]

	Experimental design-related concept assessed by BEDCI							
Class	Controls	Hypotheses	Biological variation	Accuracy	Extraneous factors	Independent sampling	Random sampling	Purpose of experiments
First-year general	∼	✔	✔	✘	✔	✘	✘	∼
First-year coordinated	∼	✔	✔	✘	∼	✘	✘	∼
First-year science	✔	✔	✔	✘	✔	✔	∼	✔
Third-year laboratory	✔	✔	✔	✔	✔	✔	∼	✔

[a] ✔ = taught explicitly (students gain practical experience in labs/tutorials/field trips and are assessed on their understanding in at least one aspect (homework, oral presentation, lab report, midterm, final, etc.). ∼ = taught implicitly (students are assessed on their understanding in a nonspecific way; they learn about the concept in lectures/tutorials and must allude to it when answering assessed questions). ✘ = absent (students are not exposed practically, and it is rare for these concepts to be taught implicitly).

To assess question difficulty (P), we calculated the percentage of students answering each question correctly. To assess discrimination, we calculated the discrimination index (D) for each question by first dividing the 580 students equally into top, middle, and bottom groups based on their total scores for the entire BEDCI (note: when multiple students with the same overall score straddled the division between two groups, we randomly divided these students between the groups to provide groups with equal numbers). We then calculated the discrimination index (D) for each question using the following formula: $D = (f^{\mathrm{TOP}} - f^{\mathrm{BOTTOM}})/(n/3)$, where f^{TOP} = the frequency of expert-like answers provided by students in the top third, f^{BOTTOM} is the frequency of expert-like answers provided by students in the bottom third, and n is the total number of responses from all students (Baker, 2004; Morrow *et al*., 2005; Smith *et al*., 2008; Streveller *et al*., 2011). This index provides an indication of how well each question discriminates between students that performed well and those who performed poorly across the BEDCI as a whole instrument. To provide a measure of the reliability of BEDCI, we needed to compare student responses from the most similar courses we had sampled. We chose to compare pretest responses (proportion of students choosing each available multiple-choice answer across all BEDCI questions) from the first-year general classes (T1 and T2). While noting that these two classes were independent, and that a test–retest reliability calculation would have been preferable if we had been able to collect such data, we assumed these two classes would show the least variation in incoming student populations, because they were in the same academic year and class sizes were large and similar (n = 170 students and n = 160 students, respectively). Following the protocol of Smith *et al*. (2008), we used the same test–retest correlation analysis (Pearson product-moment correlation) to calculate an r value as a tentative indicator of the reliability of pretest scores.

Although our primary purpose in collecting data from the five classes was to assess the sensitivity and reliability of the BEDCI, we also compared matched pre- and posttest performance for our 580 students. Because our data were nonnormally distributed, we used two-tailed Wilcoxon rank-sum tests, with $\alpha \leq 0.05$, to compare pre- and posttest performance by students within each class. We also pooled students from all classes (n = 580) to perform pre- and posttest analyses of each question separately; these analyses allowed us to identify questions for which students performed significantly better on the posttest compared with the pretest. Pre- and posttest scores for this question-by-question analysis were compared using chi-square goodness-of-fit tests, where expected frequency = the frequency of correct pretest responses, and observed frequency = frequency of correct posttest responses.

Not all of the concepts assessed in the BEDCI were taught explicitly in the five classes in our study (Table 3), but some of the core concepts were, and others may have been taught in other courses the students were taking concurrently. As such, we did not necessarily expect high positive-learning changes but still hypothesized that we would see improvement in scores between the pre- and posttests. We calculated normalized change (c) for each student (Marx and Cummings, 2007) and an overall c-average for each class (i.e., the average normalized change across all students in that class). We also calculated the probability of superiority (*PSdep*) to provide a nonparametric assessment of the effect size for each class when comparing student BEDCI scores in the posttest with the same student scores in the pretest. The probability of superiority (Grissom and Kim, 2012) provides an estimate of the probability (0–1) that a randomly chosen student in the posttest would perform better than the same student in the pretest.

Ethics Protocol Compliance

This study complied with ethics requirements for human subjects as approved by the Behavioural Research Ethics Board at the research university used in this study (BREB H09-03080).

RESULTS

Sensitivity and Reliability of the BEDCI

Questions varied widely in their difficulty (P) in both the pretest and posttest (Table 4), as indicated by the proportion of students who provided expert-like answers. In the pretest, question difficulty ranged from 24.1 to 80.0%, and in the posttest it ranged from 37.9 to 85.0% (Table 4). For all but three of the questions (i.e., questions 2 and 9 regarding Hypotheses, and question 8 regarding Random Sampling), a higher proportion of students answered in an expert-like manner in the posttest compared with the pretest. Difficulty was very similar for both Hypotheses questions (questions 2 and 9) in the pre- and posttests but was informative, as further investigation showed that students answering one question in an expert-like way were not necessarily answering the other one in an expert-like way.

Discrimination index (D) values ranged from 0.15 to 0.52 in the pretest and from 0.24 to 0.52 in the posttest (Table 4). In the pretest, six questions were good discriminators on the basis that they had D values between 0.25 and 0.39, and six were excellent ($D > 0.4$). In the posttest, eight were good discriminators, and five were excellent. The one question with a D value below 0.25 at the posttest stage was question 1, Controls ($D = 0.24$).

We compared the pretest scores on the BEDCI for the two classes of first-year general students (T1 and T2), using the Pearson product-moment correlation as an estimate of the BEDCI's reliability, and found scores to be well correlated ($r = 0.84$, $p \leq 0.0001$).

Student Performance on the BEDCI

Posttest performance was higher than pretest performance in all five classes. In the case of first-year general (T2), first-year coordinated, and third-year laboratory students, the differences between pre- and posttest scores were significant (Table 5). Both pre- and posttest performances were higher for students in first-year science and first-year coordinated than in the other classes at these two test stages. Third-year laboratory students scored lower than those in any other class at the pretest stage, and second lowest at the posttest stage.

All five classes showed positive average normalized learning changes, although the first-year general T1 class showed minimal gains (4.1 ± 2.4%; Table 5). The first-year coordinated class showed the highest learning gains, at 23.8 ± 2.7%. The average normalized learning change for the pooled sample

Table 4. Posttest precision and sensitivity analyses (difficulty: *P*; discrimination: *D*) for the 14 BEDCI questions (all classes pooled, $n = 580$), split into the associated concepts they address

		Difficulty (*P*) (% expert-like)		Discrimination (*D*) (−1 to +1)	
Core concept	BEDCI question number	Pre	Post	Pre	Post
Controls	1	80.0	85.0	0.21	0.24
	5	55.3	63.3	0.46	0.35
Hypotheses	2	40.2	38.6	0.34	0.25
	9	42.2	38.4	0.46	0.31
Biological variation	3	69.0	78.8	0.31	0.30
	10	37.4	43.6	0.15	0.35
Accuracy	4	58.6	58.5	0.50	0.35
Extraneous factors	6	49.7	69.0	0.49	0.51
	14	49.5	52.9	0.45	0.37
Independent sampling	7	50.3	52.6	0.52	0.52
	12	24.1	37.9	0.25	0.49
Random sampling	8	51.7	46.9	0.35	0.47
	13	63.6	70.5	0.31	0.32
Purpose of experiments	11	32.4	50.2	0.38	0.46

($n = 580$) was 11.5 ± 1.2%. Similar patterns were found in the probability of superiority (*Psdep*) analyses; the first-year general class effect size was negligible (*Psdep* = 0.5), but the first-year coordinated class effect size was quite high (*Psdep* = 0.7). The average probability of superiority for the pooled sample was 0.56 (Table 5).

The percentage of students answering in an expert-like way in the posttest was significantly higher than in the pretest for eight of the 14 BEDCI questions when considering the pooled sample: questions 1 and 5, Controls (Q1, $\chi^2 = 8.75, p \leq 0.001$, Q5, $\chi^2 = 14.44$, $p \leq 0.0001$), questions 3 and 10, Biological Variation (Q3, $\chi^2 = 25.71$, $p \leq 0.0001$; Q10, $\chi^2 = 9.28$, $p \leq 0.01$), question 6, Extraneous Factors (Q6, $\chi^2 = 157.13$, $p \leq 0.0001$), question 12, Independent Sampling (Q12, $\chi^2 = 59.5$, $p \leq 0.001$), question 13, Random Sampling (Q13, $\chi^2 = 11.62$, $p \leq 0.001$), and question 11, Purpose of Experiments (Q11, $\chi^2 = 82.68, p \leq 0.0001$); see Figure 1.

There was no significant difference between pre- and posttest performance in five of the BEDCI questions: questions 2 and 9, Hypotheses, question 4, Accuracy, question 14, Extraneous Factors, and question 7, Independent Sampling (all $p \geq 0.05$). For one question, question 8, Random Sampling, a significantly lower percentage of students answered in an expert-like way in the posttest than in the pretest (Q8, $\chi^2 = 5.22, p \leq 0.022$).

DISCUSSION

We have designed and tested a CI to diagnose specific examples of non–expert-like thinking by undergraduate students regarding experimental design in biology (Table 1), filling an existing gap in CI literature. We followed best practices in designing and validating the inventory (Garvin-Doxas *et al.*, 2007; Smith *et al.*, 2008; Adams and Wieman, 2010). The BEDCI was implemented as a pre- and posttest in five undergraduate biology classes and proved to be both sensitive and reliable in diagnosing specific examples of non–expert-like thinking in student perceptions of experimental design.

Sensitivity and Reliability of the BEDCI

Questions varied in difficulty (*P*) for all cohorts of students. Doran (1980) recommends that difficulty should fall within a range of 0.30–0.90 in order to effectively capture changes from the pre- to posttest stage. At the pretest stage, only

Table 5. Summary of mean pre- and posttest scores for individual courses[a]

Course (term)	*n*	Pretest mean % ± SE	Posttest mean % ± SE	Test statistic (significance)	Average normalized learning change (*c*-average) % ± SE	Probability of superiority (*PSdep*)[b]
General (T1)	170	47.5 ± 1.3	50.2 ± 1.3	W = 12984.5 ($p = 0.1033$)	4.1 ± 2.4	0.5 [$n = 85$]
General (T2)	160	50.1 ± 1.3	56.0 ± 1.4	W = 20197 ($p = 0.0015$*)	11.9 ± 2.7	0.54 [$n = 87$]
Coordinated (T2)	110	53.8 ± 1.6	65.6 ± 1.4	W = 3568 ($p \leq 0.0001$*)	23.8 ± 2.7	0.70 [$n = 77$]
Science (T1)	57	59.6 ± 1.9	62.7 ± 1.6	W = 1382.5 ($p = 0.1631$)	10.0 ± 3.3	0.53 [$n = 30$]
Biology 3 (T1)	83	45.4 ± 1.8	52.2 ± 1.9	W = 2659.5 ($p = 0.0130$*)	10.8 ± 3.1	0.59 [$n = 49$]
All (T1 and T2)	580	50.3 ± 0.7	56.2 ± 0.7	W = 2873 ($p = 0.0062$*)	11.5 ± 1.2	0.56 [$n = 327$]

*Posttest scores were significantly higher than pretest scores (W: Wilcoxon sum-rank tests).

[a] T1: Fall term (September–December 2012); T2: Winter term (January–April 2013). Also shown are average normalized learning changes and probability of superiority (*Psdep*) effect size values.

[b] [$n =$] represents the number of students whose posttest BEDCI score was greater than their pretest score.

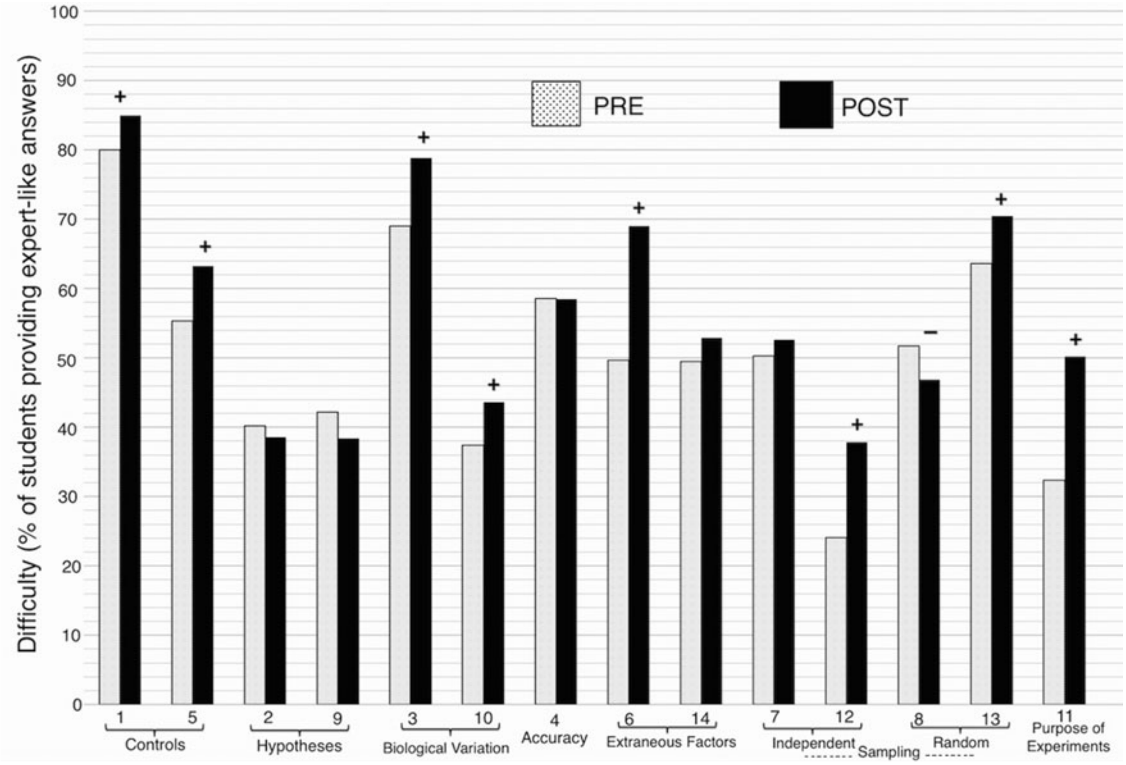

Figure 1. Comparisons of pre- and posttest student performance on each of the BEDCI questions (n = 580 student). Numbers on the x-axis are BEDCI question numbers, organized by core concepts. Statistical significance was calculated using chi-square goodness-of-fit tests. Plus symbol (+): significantly *more* students provided expert-like answers in the posttest; minus symbol (–): significantly *fewer* students provided expert-like answers in the posttest.

one BEDCI question (question 12, Independent Sampling) fell outside this range (0.24), and by the posttest stage, all questions were within this suggested range. These values suggest that the BEDCI is an appropriate tool for measuring change in learning.

The discrimination index (D) values also suggest that BEDCI is an effective tool for measuring change in learning; individual questions with relatively high D values indicate that students who scored well on the BEDCI as a whole also scored well on that individual question. This is important, because if a question has a low D value, it suggests it is ineffective at distinguishing high-scoring students from low-scoring students (Baker, 2004; Streveller *et al*., 2011). Thorndike (1997) suggests questions with D values between 0.25 and 0.39 are "good" discriminators, whereas those of 0.40 or above are "excellent." Eight of the BEDCI questions fell into the "good" category based on this scale in the posttest, while a further five questions were "excellent" discriminators; only one question was just below the 0.25 mark (question 1, Controls, $D = 0.24$). This was the easiest question on the BEDCI, which meant high- and low-scoring students across the whole inventory scored relatively well on it.

We determined a measure of reliability of BEDCI as a research instrument by comparing student pretest scores between the two sections of first-year general that were included in this study (Smith *et al*., 2008; Semsar *et al*., 2011). With a reliable instrument, similar populations of students should score similarly on individual questions as well as on the whole instrument when tested at the same stage of their university education. Semsar *et al*. (2011) suggest that an r value of $\geq$0.80 indicates high reliability; for the BEDCI; student scores had a correlation coefficient of $r = 0.84$. If a CI is to be used as an indicator of which specific examples of non–expert-like thinking are present in student populations and reveal how they change over time, high reliability is very important. Such an instrument would allow assessors to be more confident that any changes in student performance (between the pre- and posttest stages, for example) indicate genuine shifts in learning. Other CIs, such as the meiosis CI, have been used by Kalas *et al*. (2013) to evaluate the effectiveness of teaching strategies. We anticipate that BEDCI will be effective in measuring the impact of new teaching activities especially developed to target the non–expert-like thinking identified at the pretest stage.

Student Cohort Performance on the BEDCI

Student understanding of concepts relating to experimental design appears to have improved over the duration of the term; when all students were pooled, the mean posttest score was significantly higher than the mean pretest score. When the five classes surveyed in this study were analyzed individually, three showed significant improvement at the posttest stage. The two other classes (first-year general [T1] and first-year science) also showed improvements but the changes were not statistically significant. Why these two classes did not show significant improvement is unclear. In the case of the first-year general (T1) class, we hypothesize that first-year students in the first term of their undergraduate studies may not be motivated to modify their learning strategies to suit instruction that is likely more conceptually based than what they have encountered in high school. The performance of these T1 students on the BEDCI may reflect their ineffective learning strategies or indicate their reluctance to revise preheld conceptions of experimental design (Butler *et al*., 2008). In the case of first-year science students, scores were high in the pretest, at least relative to other classes' pretest scores, and thus there was less room for improvement on the posttest. However, we note that there were still many questions on which students could have improved, indicating concepts that are especially difficult for students to grasp. This is a subject for detailed, future investigation.

That we observed variation in scores among different classes suggests the BEDCI is able to distinguish between students from slightly different educational backgrounds and with slightly varied experience in experimental design. For example, first-year science students, who are admitted to their cohort program based on high school performance, scored much higher on the BEDCI (pre- and posttest) than the two classes of first-year general students (T1 and T2). Given that these classes all featured first-year students, this variation in scores suggests the BEDCI should be sensitive in discriminating between students whose expertise in experimental design falls at different points on the nonexpert to expert spectrum. We were surprised that students in the third-year laboratory class (the only third-year class to which BEDCI was administered) scored lower than any other class at the pretest stage and second lowest at the posttest stage. We expected that they would answer a higher proportion of questions in an expert-like way than students from the first-year classes, given they would have had more instruction on experimental design and would have already been taught either explicitly or implicitly the eight core concepts assessed by the BEDCI (Table 1). One potential explanation for the relatively low pretest scores in third-year laboratory is that there are no experimental, laboratory-based biology courses in the biology program for second-year students studying at our university. Students in this class therefore have not had explicit instruction regarding experimental design since first year. Other possible explanations for the low posttest score are that the third-year laboratory class is not intended for biology majors, this cohort of third-year students may not have been academically strong, and non–expert-like conceptions had become deeply entrenched at this stage of their studies. Either way, this unexpected result warrants further investigation into whether students at this stage of their biology education require more frequent explicit instruction on experimental design to prevent regression to non–expert-like conceptions.

The average normalized learning change (*c*-average), which is a measure of the normalized learning change for all students (Marx and Cummings, 2007), was relatively low. The learning changes for each class as measured by BEDCI were also smaller than those reported for other biology CIs, such as the GCA (Smith *et al*., 2008). The positive learning changes in large, genetics-based undergraduate courses as measured with the GCA were in excess of 50%. However, we feel that the small learning changes measured by the BEDCI are not unexpected. In contrast to the genetics course, which specifically addressed concepts directly tested by the GCA, four of the five classes surveyed in this study focused on concepts relating to ecology, genetics, and evolution in a lecture-based course, and the elements of experimental design were not all taught explicitly but may instead have been incorporated to provide context for the intended course content. In our study, students, (except those in the third-year laboratory class), were not actively applying experimental design in their courses. Nehm and Reilly (2007) and Sirum and Humburg (2011) reported that, when biology students engaged with concepts through active learning, their understanding of natural selection and the important factors in designing a good experiment, respectively, improved significantly; students taught the same material in a traditional lecture setting did not show significant improvement. The first-year students in our study were enrolled in lecture courses and did not have the benefit of explicit instruction and practice applying experimental design concepts. Our study also differed in the manner in which we administered our CI. Smith *et al*. (2008) administered the GCA posttest as part of a final exam, which ensured that students took the CI questions very seriously, had studied the concepts beforehand, and arrived ready to be tested. In contrast, the BEDCI posttest was administered in either the last or second to last week of classes and without the recommendation that students should study related material beforehand. We hypothesize that students would have achieved higher learning gains had they been given explicit instruction and practice in experimental design and had extra motivation been provided by incorporating BEDCI as part of a final exam.

We conducted the probability of superiority (*Psdep*) analyses (Grissom and Kim, 2012) to provide a measure of effect size for the different classes and the pooled sample, and these showed the same general patterns indicated by the average normalized change analyses. Students in the first-year general (T1) class were no more likely to have higher BEDCI scores in the posttest than the pretest, but those in the first-year coordinated class were 70% more likely to have scored higher in the posttest. The instructor in this class ran a small group-based activity toward the end of the term, asking students to critique an experiment, so it is perhaps not surprising that this class performed better on the BEDCI and showed the greatest normalized learning change and probability of superiority.

Analysis of Individual Questions

When we considered the BEDCI questions individually, comparing pre- and posttest scores, significant shifts in student understanding were evident. Students showed significant positive normalized learning changes in eight of the 14 questions, falling into the following core concepts: Controls, Biological Variation, Extraneous Factors, Independent Sampling, Random Sampling, and Purpose of Experiments. Although the instructors did not target their instruction regarding the use and effective design of control groups in experiments or use any additional learning activities regarding controls, we found that performance improved on BEDCI questions relating to the core concept Controls. This positive shift suggests that Controls might be one of the easier concepts to master in experimental design or that students are more familiar with the concept as it is commonly used in the media and may have had this concept reinforced in lecture courses. An alternate explanation is that students could be taking other courses that reinforce learning regarding the use of controls in experiments.

In our study, not all shifts in understanding were significant or positive (i.e., toward more expert-like understanding). There was no significant change toward expert-like thinking in five of the BEDCI questions at the posttest stage. In some cases, a majority of students still chose answers that reflected non–expert-like thinking at the posttest stage. Less than 50% of students gave expert-like responses to the two Hypotheses questions on both the pretest and posttest. Three questions that test the understanding of Biological Variation, Independent Sampling, and Random Sampling were also answered correctly by <50% of students on the posttest.

In the case of the Random Sampling question, students shifted significantly away from the expert-like response and toward non–expert-like responses. Open-ended questions in think-aloud interviews in the first stages of designing the BEDCI indicated that virtually all undergraduates knew that it was important to sample randomly but that they did not know why or how to do this. The fact that they took a backward step at the posttest stage suggests confusion in students' minds as they grapple with new conceptions of random sampling. Theory suggests that learning is not a unitary process, and there are different stages involved, such as accruing knowledge and interpreting, conceptualizing, and then modifying it (Qian and Alvermann, 1995; Pearsall *et al.*, 1997). It may be that the confusion with this particular concept requires greater modification of a student's cognitive framework, which would take longer for the student to process. The backward conceptual shift evident in the posttest highlights the value of using CIs in identifying naïve conceptions (Garvin-Doxas *et al.*, 2007); while other tests or forms of student evaluation might indicate reasonably good understanding of this concept (students can recite the need to sample randomly), the BEDCI indicated that many students could not apply this conceptual knowledge. Instructors recognizing that students have difficulty with these concepts can take steps to implement teaching strategies that can directly address the non–expert-like conceptions.

Limitations of the BEDCI

Two of the core concepts assessed by BEDCI—Accuracy and Purpose of Experiments—have only one question to capture student thinking. In both cases, the second question was discarded over the course of the development process (see *Methods*). It is preferable to have at least two questions testing the same concept to confirm that learning changes measured using the two (or more) questions are similar for given students and classes. However, during the development process, we used student interviews to assess the accuracy of all the questions, including the two addressing these two core concepts as assessment items. We believe that we have sufficient validity evidence that these questions will provide meaningful data on student thinking in these conceptual areas of experimental design.

There is some evidence that when different organisms are used to contextualize the same basic open-ended questions, student responses can be affected (Nehm and Ha, 2011). We did not assess whether changing the organisms used in our scenarios would affect student responses to BEDCI questions, but we purposefully chose organisms that were easy to imagine, and we also provided simple background information with each scenario to reduce the chance that a lack of familiarity with the organism would impact interpretation. When we interviewed students, none stated that they preferred a particular scenario because of the organism or the experimental setup described in it. By choosing simple, organismal-level scenarios, we could potentially paint a simplified picture of experimental design. However, incorporating nonorganismal scenarios was beyond the scope of this instrument, which was designed to be useful for undergraduate students, from first year to upper levels. More technically sophisticated scenarios might have proved to be too conceptually difficult for first-year students to grasp.

CONCLUSION

In this paper, we have demonstrated that the BEDCI is a reliable and valid diagnostic tool, able to measure non–expert-like student thinking regarding experimental design. BEDCI is sensitive in detecting differences among student populations, as well as in measuring individual and cohort shifts in student learning. We hope that the BEDCI, in addition to diagnosing non–expert-like student understanding, can be used as a means of assessing the effectiveness of new teaching strategies in improving student learning.

Innovations in teaching, especially strategies that enhance engagement and deeper learning, have been advocated by various agencies (Woodin *et al.*, 2010; American Association for the Advancement of Science, 2011). Gathering evidence on student learning and our teaching effectiveness provides important feedback as we modify our teaching practice. In many ways, this process of investigation into teaching mirrors scientific investigation and is often referred to as scientific teaching (Handelsman *et al.*, 2004).

While the BEDCI was developed with specific relevance to experimental design in biology, we predict that it will have wider applications in related disciplines, such as in the medical and life sciences. Because the fundamentals of good experimental design apply to any discipline in which experiments are performed to produce reliable data, the BEDCI should be useful in diagnosing specific examples of non–expert-like thinking by medical science students and professionals who are conducting research. Lambert and Black (2012) documented many experiments that were poorly designed by medical researchers. Assessing the validity and quality of such studies is crucial (Craig *et al.*, 2008; Röhrig *et al.*, 2009) and underscores the need for better instruction regarding experimental design and robust evaluation tools in the medical and other life sciences disciplines.

We encourage instructors to either visit our Questions for Biology website (http://q4b.biology.ubc.ca) or contact the corresponding author for more information about gaining access to the BEDCI and its "inventory package."

ACKNOWLEDGMENTS

The Teaching and Learning Enhancement Fund of the research university sampled in this study provided the funds for the Questions for Biology project. Additional funds were provided by Skylight, through the university's Science Centre for Learning and Teaching. We specifically thank the undergraduate students who took part in the think-aloud interviews, pilot assessment, and in-class pre- and posttest deployments of the finalized BEDCI, as well as the instructors who incorporated the inventory into their classes. We also thank the instructors and graduate students who served as experts in validating BEDCI questions. We thank Abdel Azim Zumrawi of our Centre for Teaching, Learning and Technology for his help with the statistical aspects of this paper. Finally, we acknowledge the two anonymous reviewers for providing insightful feedback that improved the manuscript.

REFERENCES

Adams DJ (2009). Current trends in laboratory class teaching in university bioscience programmes. Biosci Educ *13*, 1–13.

Adams WK, Wieman CE (2010). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ *33*, 1289–1312.

Aikenhead GS, Ryan AG (1992). The development of a new instrument: "Views on Science–Technology–Society" (VOSTS). Sci Educ *76*, 477–491.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf (accessed 9 June 2013).

Anderson-Cook CM, Doraj-Raj S (2001). An active learning in-class demonstration of good experimental design. J Stat Educ *9*(1). http://amstat.org/publications/jse/v9n1/anderson-cook.html (accessed 6 November 2013).

Baker FB (2004). Item Response Theory: Parameter Estimation Techniques, Cleveland, OH: CRC.

Butler DL, Pollock C, Nomme KM, Nakonechny J (2008). Promoting authentic inquiry in the sciences: challenges faced in redefining university students' scientific epistemology. In: Inquiry in Education, vol. II, Overcoming Barriers to Successful Implementation, ed. BM Shore, MW Aulls, and MAB Delcourt, New York: Taylor & Francis, chap. 14.

Coil D, Wenderoth MP, Cunningham M, Dirks C (2010). Teaching the process of science: faculty perceptions and an effective methodology. CBE Life Sci Educ *9*, 524–535.

Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ *337*, a1655.

Crowther J, Price RM (2014). Re: misconceptions are "So yesterday!" CBE Life Sci Educ *13*, 3–5.

Cummins RH, Green WJ, Elliott C (2004). "Prompted" inquiry-based learning in the introductory chemistry laboratory. J Chem Educ *81*, 239–241.

Dasgupta AP, Anderson TR, Pelaez N (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. CBE Life Sci Educ *13*, 265–284.

Ding L, Chabay R, Sherwood B, Beichner R (2006). Evaluating and electricity and magnetism tool: brief electricity and magnetism assessment. Phys Rev *2*, 010105.

Doran R (1980). Basic Measurement and Evaluation of Science Instruction, Washington, DC: National Science Teachers Association.

Festing MFW (2003). Principles: the need for better experimental design. Trends Pharmacol Sci *24*, 341–345.

Garvin-Doxas K, Klymkowsky M, Elrod S (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: report on a National Science Foundation–sponsored conference on the construction of concept inventories in the biological sciences. CBE Life Sci Educ *6*, 277–282.

Green DS, Bozzone DM (2001). A test of hypotheses about random mutation: using classic experiments to teach experimental design. Am Biol Teach *63*, 54–58.

Grissom RJ, Kim JJ (2012). Effect Sizes for Research: Univariate and Multivariate Applications, New York: Routledge, ebook.

Handelsman J *et al.* (2004). Scientific teaching. Science *304*, 521–522.

Hiebert SM (2007). Teaching simple experimental design to undergraduates: do your students know the basics? Adv Physiol Educ *31*, 82–92.

Hurlbert SH (1984). Pseudoreplication and the design of ecological field experiments. Ecol Monogr *54*, 187–211.

Kalas P, O'Neill A, Pollock C, Birol G (2013). Development of a meiosis concept inventory. CBE Life Sci Educ *12*, 655–664.

Kline TJB (2005). Psychological Testing: A Practical Approach to Design and Evaluation, Thousand Oaks, CA: Sage.

Lambert CG, Black LJ (2012). Learning from our GWAS mistakes: from experimental design to scientific method. Biostat *13*, 195–203.

Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci *11*, 5.

Leonard MJ, Kalinowski ST, Andrews TC (2014). Misconceptions yesterday, today, and tomorrow. CBE Life Sci Educ *13*, 179–186.

Marx JD, Cummings K (2007). Normalized change. Am J Phys *75*, 87–91.

Maskiewicz AC, Lineback JE (2013). Misconceptions are "So yesterday!" CBE Life Sci Educ *12*, 352–356.

Morrow J, Jackson A, Disch J, Mood D (2005). Measurement and Evaluation in Human Performance, Champaign, IL: Human Kinetics.

Mortimer E (1995). Conceptual change or conceptual profile change. Sci and Educ *4*, 267–285.

Naegle E (2009). Patterns of Thinking about Phylogenetic Trees: A Study of Student Learning and the Potential of Tree-Thinking to Improve Comprehension of Biological Concepts, Pocatello: Idaho State University.

Nehm RH, Ha M (2011). Item feature effects in evolution assessment. J Res Sci Teach *48*, 237–256.

Nehm RH, Reilly L (2007). Biology majors' knowledge and misconceptions of natural selection. BioScience *57*, 263–272.

Pearsall NR, Skipper JEJ, Mintzes JJ (1997). Knowledge restructuring in the life sciences: a longitudinal study of conceptual change in biology. Sci Educ *81*, 193–215.

Perry WG, Jr. (1981). Cognitive and ethical growth: the making of meaning. In: The Modern American College, ed. AW Chickering and associates, San Francisco: Jossey-Bass, 76–116.

Pollack AE (2010). Exploring the complexities of experimental design: using an on-line reaction time program as a teaching tool for diverse student populations. J Undergrad Neurosci Educ *9*, A47–A50.

Qian G, Alvermann D (1995). Role of epistemological beliefs and learned helplessness in secondary school students' learning science concepts from text. J Educ Psychol *87*, 282–292.

Rivet A, Krajcik J (2008). Contextualizing instruction: leveraging students' prior knowledge and experiences to foster understanding of middle school science. J Res Sci Teach *45*, 79–100.

Röhrig B, du Prel J-B, Blettner M (2009). Study design in medical research. Dtsch Artzebl Int *106*, 184–189.

Roy NM (2013). Using RNAi in *C. elegans* to demonstrate gene knockdown phenotypes in the undergraduate biology lab setting. Bioscene *39*, 16–20.

Semsar K, Knight JK, Birol G, Smith MK (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in biology. CBE Life Sci Educ *10*, 268–278.

Sere M-G, Fernandez-Gonzalez M, Gallegos JA, Gonzalez-Garcia F, De Manuel E, Perales FJ, Leach J (2001). Images of science linked to labwork: a survey of secondary school and university students. Res Sci Educ *31*, 499–523.

Shi J, Power JM, Klymkowsky MW (2011). Revealing student thinking about experimental design and the roles of control experiments. Int J Schol Teach Learn *5*, 1–19.

Sirum K, Humburg J (2011). The Experimental Design Ability Test (EDAT). Bioscene *37*, 8–16.

Smith J, diSessa A, Rochelle J (1993). Misconceptions reconceived: a constructivist analysis of knowledge in transition. J Learn Sci *3*, 115–163.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ *7*, 422–430.

Streveller RA, Miller RL, Santiago-Roman AI, Nelson MA, Geist MR, Olds BM (2011). Rigorous methodology for concept inventory development: using the "assessment triangle" to develop and test the Thermal and Transport Science Concept Inventory (TTCI). Int J Eng Educ *27*, 968–984.

Thorndike RM (1997). Educational Tests and Measurements: Psychological Tests, Upper Saddle River, NJ: Merrill.

Woodin T, Carter VC, Fletcher L (2010). Vision and Change in Biology Undergraduate Education, a call for action—initial responses. CBE Life Sci Educ *9*, 71–73.

Zolman JF (1999). Teaching experimental design to biologists. Adv Physiol Educ *277*, 111–118.