CBE—Life Sciences Education
Vol. 15, 1–13, Spring 2016

Article

Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI)

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

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

Submitted June 21, 2015; Revised November 19, 2015; Accepted November 19, 2015
Monitoring Editor: Erin Dolan

We followed established best practices in concept inventory design and developed a 12-item inventory to assess student ability in statistical reasoning in biology (Statistical Reasoning in Biology Concept Inventory [SRBCI]). It is important to assess student thinking in this conceptual area, because it is a fundamental requirement of being statistically literate and associated skills are needed in almost all walks of life. Despite this, previous work shows that non–expert-like thinking in statistical reasoning is common, even after instruction. As science educators, our goal should be to move students along a novice-to-expert spectrum, which could be achieved with growing experience in statistical reasoning. We used item response theory analyses (the one-parameter Rasch model and associated analyses) to assess responses gathered from biology students in two populations at a large research university in Canada in order to test SRBCI's robustness and sensitivity in capturing useful data relating to the students' conceptual ability in statistical reasoning. Our analyses indicated that SRBCI is a unidimensional construct, with items that vary widely in difficulty and provide useful information about such student ability. SRBCI should be useful as a diagnostic tool in a variety of biology settings and as a means of measuring the success of teaching interventions designed to improve statistical reasoning skills.

INTRODUCTION

In undergraduate science education, interpreting graphical figures and summaries of data and statistically analyzing the results of experiments are fundamental to the development of scientific literacy (Roth *et al.*, 1999; Samuels *et al.*, 2012). One aspect of scientific literacy is statistical reasoning, which can be defined as the ways in which we reason with and make sense of numerical data and apply statistical theories (Garfield, 2002). Students require grounding in key statistical concepts, such as sampling, distribution, randomness, and uncertainty (Garfield, 2002); to make sound judgments about data, they must make connections among related concepts, such as center and spread, and variability and significance (Ben-Zvi and Garfield, 2004). The phrase "statistical reasoning" implies a combination of data description with the concept of probability (and statistical significance), which is needed to interpret statistics and make objective inferences.

The development of statistical reasoning skills is widely recognized as a critical component of undergraduate statistics education (Garfield, 2002); students undertaking any science degree will encounter experiences with experimental design and the subsequent analysis of data and must thus develop sufficient ability in statistical reasoning to succeed in their courses. Not only is a good grasp of statistical reasoning a hallmark of a successful science degree (Coil *et al.*, 2010), but these skills are also necessary for the logical interpretation of the many varieties of statistical data we all encounter in everyday life. Students need to interpret basic graphs and many other forms of data (Silverman, 2011; Howell, 2014) and evaluate the reliability and validity of the information they encounter (Glazer, 2011). Employers value

CBE Life Sci Educ March 1, 2016 15:ar5
DOI:10.1187/cbe.15-06-0131
Address correspondence to: Kathy Nomme (nomme@zoology.ubc.ca).

these skills in potential employees, even if jobs do not explicitly require the analysis of data (Durrani and Tariq, 2008). In essence, we can make better-informed decisions if we possess such skills (Watson, 2011; Blagdanic and Chinnappan, 2013). Additionally, these skills also help to develop critical thinking, another universal goal of undergraduate education (Osborne, 2010; Kim *et al.*, 2013).

Students typically hold a plethora of non–expert-like conceptions regarding statistical reasoning (Batanero *et al.*, 1994; Garfield, 2002; Forster, 2004; Kubliansky and Eschach, 2014), and they struggle to develop related skills, even following instruction (Bowen *et al.*, 1999; delMas *et al.*, 2007); this is true both at the undergraduate (Gormally *et al.*, 2012) and postgraduate (Zaidan *et al.*, 2012) level. To give two common examples, students often ignore the spread of data around the population mean and/or misinterpret data distribution when interpreting statistical significance (Garfield, 2003; Lem *et al.*, 2013), and they often have a fundamental misunderstanding of the purpose of hypothesis testing (Peskun, 1987; delMas *et al.*, 2007).

Moving students past non–expert-like conceptions (or naïve or alternative conceptions; see Maskiewicz and Lineback, 2013) toward more expert-like conceptions requires conceptual change. Although some non–expert-like conceptions may be productive in helping students to form expert-like conceptions (Smith *et al.*, 1993; Maskiewicz and Lineback, 2013; Crowther and Price, 2014), individuals with fragmented knowledge and non–expert-like conceptions may struggle to develop the expert-like conceptions we wish them to attain (Sinatra *et al.*, 2014). It should thus be a goal to promote deeper conceptual understanding of the concepts that are especially challenging and to initiate mini paradigm shifts in thinking when required (Price, 2012). Pedagogical interventions are often used to drive conceptual change (Smith, 2007), but before targeted interventions can be designed, it is important to understand which non–expert-like conceptions are most common and widespread within the demographic of interest.

The need to accurately assess the level of understanding and to quantitatively characterize specific examples of non–expert-like thinking in the statistical reasoning of our biology students led us to search for an appropriate assessment tool. We wanted a multiple-choice tool that would be suitable for undergraduates and simple to administer in large classes and that could provide results that were quick and easy to analyze so as to enable instructors to target specific concepts in their teaching based on these results. While open-ended assessments can offer great scope for students to express their knowledge, they require considerable time to grade, whereas multiple-choice questions are much easier to score (Wooten *et al.*, 2014). As a result, multiple-choice tests are often preferred in large classes, especially when set in case study contexts to encourage deeper learning (Donnelly, 2014). An additional concern is with unsupervised testing (using online learning management systems, for example), which can open the door to possible cheating (Schultz *et al.*, 2008; Styron and Styron, 2010; Ladyshewsky, 2014). There is also the concern that students will take screenshots of items and pass these to their peers or even post them online, which would devalue their future use. As such, in-class testing is often preferred. While there are a number of useful tools that assess statistical reasoning (Table 1), none of these fulfilled all of the key requirements we sought.

In this article, we describe the development and validation of a 12-item statistical reasoning concept inventory (the Statistical Reasoning in Biology Concept Inventory [SRBCI]) that is specific to biology (individual items are contextualized in biological experiment scenarios). We followed well-established best practices in developing the SRBCI (Adams and Wieman, 2010). Our aim was for instructors to use the SRBCI to help characterize specific examples of non–expert-like thinking affecting students at different stages of individual courses and at different stages of a program of study in biology. Specifically, the SRBCI should be able to measure the impact of curricula in which students conduct their own experiments and analyze their results or, in general, apply statistical reasoning concepts to data. Similarly, measuring student conceptual gains following instruction could indicate how well curricular innovations are meeting their goals. Such innovations have been called for to ensure we teach biology and statistical reasoning more effectively (American Association for the Advancement of Science, 2011), so it is vital that we have the tools to measure their impact and success on student learning.

To the best of our knowledge, this is the only concept inventory dedicated to assessing student conceptions in statistical reasoning specific to biology experiments. This is an important distinction, because people from different disciplines may engage differently with experiments and data, and as a consequence, concepts in statistical reasoning. Garfield (2002) noted that statistical reasoning skills need to be used by people from very different disciplinary backgrounds, and Ben-Zvi and Garfield (2004) underlined that discipline-specific norms exist for what constitute acceptable data arguments. While variation and spread are universally accepted as being important concepts in statistical reasoning, physicists and chemists, for example, who are used to relatively high precision in their experiments (Hunter, 2010), may be less used to reasoning about wide variation in experimental data. On the other hand, biologists should be comfortable with wide variation, seeing as it is a central concept in the discipline (Hallgrimsson and Hall, 2011). Medical researchers meanwhile, are still evolving ways of incorporating biological variation into the assessment of their studies (Simundic *et al.*, 2015).

In this article, we also assess the robustness of the SRBCI by conducting item response theory (IRT) analyses as opposed to classical test theory (CTT) analyses, due to a growing call from researchers to make use of this more objective form of measurement (Boone *et al.*, 2010; Planinic *et al.*, 2010; Wallace and Bailey, 2010; Wang and Bao, 2010). Despite there being some examples of the use of IRT in the biology education literature (Donnelly and Boone, 2006; Nehm and Schonfeld, 2008; Battisti *et al.*, 2010), these analyses may be relatively new to many researchers, so we elaborate on many of these techniques in the *Methods*.

METHODS

Development of the Statistical Reasoning Concept Inventory (the SRBCI)

Other concept inventory developers informed the method we used to develop the SRBCI (Garvin-Doxas *et al.*, 2007; Smith *et al.*, 2008; Adams and Wieman, 2010), and our approach was

Table 1. Other educational instruments that assess statistical reasoning skills in students assess a range of different constructs

Instrument/study	Summary[a]	Concepts mapped
Statistics Concept Inventory (Allen, 2006)	• 25-item multiple-choice instrument • Focuses on mathematics and statistics (statistical thinking, rather than statistical reasoning)	Data summary and presentation, probability, random variables, discrete probability distributions, continuous random variables and probability distributions, joint probability distributions, parameter estimation, linear regression, time series, confidence intervals and hypothesis testing, single-factor experiments and multifactor designs
Statistical Reasoning Assessment (Chan and Ismail, 2014)	• Five tasks completed in software program • Aimed at high-school students	Reasoning about center (mean, mode, median), spread (range, interquartile range, variance, SD), and distribution (combination of center, spread, skewness, density, outliers, causality, chance, sampling)
Statistical Reasoning Assessment (Garfield, 2003)	• 20-item multiple-choice instrument • Weighted averages based on the sum of correct reasoning and misconceptions (proportion)	Interpreting probability, selecting an appropriate average, computing probability (and as a ratio), independence, sampling variability, correlation versus causation, interpreting two-way tables, importance of large samples
Assessment Resource Tools for Improving Statistical Thinking (ARTIST website: https://apps3.cehd.umn.edu/artist/index.html)	• 11 scales/topics (each with 7–15 multiple-choice items) • Administered online	Data collection, data representation, measures of center, measures of spread, normal distribution, probability, bivariate quantitative data, bivariate categorical data, sampling distributions, confidence intervals, significance tests
CAOS 4 (delMas *et al.*, 2007)	• 40-item multiple-choice instrument • Focuses on concepts students must master after an introductory statistics course	Includes data collection and design, descriptive statistics, graphical representations, box plots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, tests of significance
Statistical Reasoning with Everyday Problems (Lawson *et al.*, 2003)	• 10-item open-ended instrument • Graders must code student answers based on the reasoning used.	Probability/chance, law of large numbers, estimation/sample bias, correlation, regression toward the mean
Verbal-Numerical and Graphical Pilot Study (Agus *et al.*, 2013)	• 11 pairs of open-ended items (one verbal-numerical, one graph per pair) • Graders must code student answers based on the reasoning used	Reasoning on uncertainty, reasoning on association

[a]The summaries indicate how these inventories compare with SRBCI, which features 12 multiple-choice items.

very similar to the path followed by other concept inventory developers in our research group (Kalas *et al.*, 2013; Deane *et al.*, 2014). Initially we 1) consulted the literature (Garfield, 2002; Haller and Krauss, 2002; Ben-Zvi and Garfield, 2004; Chance *et al.*, 2004; delMas and Liu, 2005; Fidler, 2006; Sotos *et al.*, 2007; Huck, 2009; Glazer, 2011; Karpiak, 2011; Gormally *et al.*, 2012; Zaidan *et al.*, 2012), 2) studied recent student exams and assignments, and 3) spoke to faculty members with expertise in statistical reasoning to compile a list of specific examples of non–expert-like thinking that were affecting undergraduate students. The student exams and assignments we studied came from a first-year introductory undergraduate laboratory course and a third-year lecture and laboratory undergraduate course. We chose these courses to provide a representative sample of student work across the undergraduate biology program at a large Canadian research university and to help identify the specific examples of non–expert-like thinking that seemed to be prevalent in students at various stages of their academic experiences. As expected, we found many different examples that fell along a novice-to-expert gradient for many different statistical reasoning constructs. We then met with five faculty members and two graduate students (all of whom had taught or were at that time teaching the first-year introductory undergraduate laboratory course). We requested these experts to provide us with additional examples of non–expert-like thinking they had witnessed in their students. We asked our experts open-ended questions, such as: "What do students think is meant by a significant difference?" and "What do students say when you ask them why more than one trial of an experiment should be performed before they draw firm conclusions?," and noted their answers. Using a rudimentary rubric designed to help categorize examples of non–expert-like thinking into broad categories, we found that four such core conceptual groupings emerged as having vastly more examples of non–expert-like thinking than others. We therefore decided to focus on these four core conceptual groupings in developing the SRBCI: 1) variation in data, 2) repeatability of results, 3) hypotheses and predictions, and 4) sample size. We further categorized more specific, but still common, nonexpert conceptions of statistical reasoning that fell into each of these core conceptual groupings and designed one item for each of these; as a result, there are 12 items in the SRBCI that each test a different concept, with three falling into each of the four core conceptual groupings (see Table 2 for the 12 statistical reasoning constructs that SRBCI attempts to capture).

Table 2. The four core conceptual groupings assessed by SRBCI[a]

Core conceptual grouping	SRBCI probes student understanding that:	Question	Scenario
Repeatability of results	*If statistically significant patterns in data are replicated in independent trials of an experiment, they are likely to be replicated again in other future trials …*	1	A
	But, replicated patterns do not prove that a conclusion or prior prediction is correct …	2	A
	However, conclusions are always more robust if patterns are replicated.	6	B
Variation in data	*Not all treatment groups need to differ for a researcher to conclude that the manipulated variable(s) has/have a statistically significant effect on what is being measured …*	3	A
	And patterns in data do not have to be linear to show statistically significant differences …	9	C
	But differences in variation around the sample mean are key—not the absolute differences.	10	C
Hypotheses and predictions	*Controlled experiments are designed to allow researchers to pose hypotheses, which are tested independent of predictions …*	5	B
	Which means that predictions do not have to follow either null or alternate hypotheses …	7	B
	And a hypothesis can therefore be supported by patterns in data when a prediction is not.	12	C
Sample size	*Larger sample sizes tend to increase the range of measurement values that are used to calculate sample means …*	8	B
	While simultaneously reducing the average variation around sample means …	4	A
	So different sample sizes tend to result in different average variation around sample means.	11	C

[a]Each of these groupings feature three different questions, which each assess a subtly different related concept, and the groupings are contextualized in >1 experimental scenario (scenario A = salmon; B = squirrels and raccoons; C = sunflowers).

Some questions in the SRBCI require students to answer whether or not two sample means with nonoverlapping 95% confidence intervals are significantly different from one another. Sometimes 95% confidence intervals can overlap when statistical tests indicate differences do exist between sample means (Belia *et al.*, 2005), but SRBCI was designed so as never to ask students to answer questions in such cases. We chose to use 95% confidence intervals to represent estimate error rather than other frequently used measures, such as *p* values and SE bars. This is because *p* values alone are deficient in providing certain information (Thomas *et al.*, 1997; Blume and Peipert, 2003), and we wanted to provide simple visual indicators of error in our figures, rather than relying on written statistics that were less likely to test conceptual application of basic statistical reasoning skills. We preferred 95% confidence intervals to SE bars, because they are more conservative in allowing one to infer significant differences between sample means that do not overlap and because SE bars are more easily confused by students, who sometimes mix up SD, SEM, and other measures (Belia *et al.*, 2005; Cumming *et al.*, 2007). Many of the SRBCI items incorporate figures, and while we are aware that some students find visual interpretation difficult for a variety of reasons (Chinn and Brewer, 2001; Friel *et al.*, 2001; Cooper and Shore, 2010), interpreting figures is a fundamental requirement of scientific literacy (and statistical reasoning; Coil *et al.*, 2010; Glazer, 2011; Watson, 2011). It should also be noted that some items within each of the four core conceptual groupings assessed by the SRBCI incorporate figures and some do not.

We designed the first versions of our 12 SRBCI items and contextualized these by using scenarios involving experiments with salmon, squirrels and raccoons, and sunflowers. We did this because using plausible real-world examples can increase student interest (Claxton, 2007; Rivet and Krajcik, 2008) and because student reasoning is often connected to context (Mortimer, 1995). Basing multiple questions on the same scenario minimized reading for students, because the background information for each scenario applied to all four items in that scenario grouping. Nehm and Ha (2011) found that changing the organism used to provide experimental context influenced student responses in an open-ended setting, but there was no indication from our think-aloud interviews that this was an issue for students; the scenario settings were all very simple, and the background information only required students to read around 70 words. Additionally, we asked students whether they could picture the experiments and the organisms in them. No student said he or she preferred one scenario over the other two, and the simple, succinct language used in all the SRBCI answer choices never seemed to cause confusion.

We held a total of 23 one-on-one, think-aloud interviews with undergraduate students (15 female, 8 male), in which they were asked to answer each item and explain why they had selected their answers. Students came from a variety of programs within the Faculty of Science, but all had taken or were taking at least one biology course at the time. Incoming science students are accepted to this large Canadian research university with excellent secondary school grade averages (around 92.5% entrance average), and the first-year students (n = 17) we interviewed were from this pool. We used in-class and online announcements to attract participants, who were paid C$15 for participating in a 50–60 min interview. Our purpose in conducting the interviews was to ensure students were interpreting the items as intended, to use student language and suggestions to improve the items in terms of wording (enhancing comprehension and providing more realistic non–expert-like distractors), and to boost the clarity of figures.

During the interviews, we probed student thinking to determine where our students were situated on the novice-to-expert spectrum with regard to statistical reasoning. We found that students possessed a range of nonexpert conceptions, but that many of the answers we received provided insights into the development of students' skills along this spectrum. While students would not be expected to progress linearly or easily from novices to experts (some expert-like reasoning may be lost and replaced by non–expert-like reasoning, for example), the answers they provide can help to

monitor the progression of their thinking. Supplemental Table S1 shows student quotes for each answer of one item from the four core conceptual groupings SRBCI attempts to capture, highlighting some of the thinking underpinning statistical reasoning of these students. Supplemental Table S2 delineates further common examples of statistical reasoning that students used when answering items in the four core conceptual groupings, and how such reasoning may be placed on a novice-to-expert spectrum.

Briefly, we asked students to read each item and its four answers aloud before giving them time to decide on an answer. We then asked them to justify their answers and explain why they did not select one of the other three answers to ensure they had understood the item and that any non–expert-like thinking was being correctly diagnosed rather than a student selecting such an answer as a result of misinterpreting the item or answer. We made audio recordings of every interview and listened to these to make sure nothing was missed. When a student misinterpreted an item and/or answer, we paid special attention to it/them in subsequent interviews to see whether similar issues arose. In the event that we altered wording and/or a figure, we also paid special attention to the new version in subsequent interviews to see whether the troublesome element(s) had been removed by our alterations. The final five interviews did not present a single case of a student misinterpreting an item, figure, or answer.

Following these interviews, we asked 17 experts (12 faculty members and five graduate students at the large Canadian research university where this study was conducted) to provide expert-like answers to the items to confirm their validity. The faculty members comprised five zoologists, four botanists, and three physiologists, all of whom had PhD qualifications and were instructing undergraduate biology courses at this university at the time. We qualified the five graduate students as experts, because they were enrolled in either the MSc ($n = 3$) or PhD ($n = 2$) program in the Department of Zoology and had been teaching assistants in a first-year introductory laboratory course that specifically taught students skills in statistical reasoning at least once (three were currently teaching it again). The 12 faculty members all selected the same answers, while the graduate students agreed on 96.7% of their answers. The two non–expert-like responses chosen by the graduate students were in response to different items (one considering the core conceptual grouping of variation in data, and the other considering repeatability of results). At the end of this process, we were left with our 12-item SRBCI.

Deployment of the SRBCI in Two University-Level Biology Courses

To gather data that could be used to assess the robustness and sensitivity of the SRBCI, we administered it in two different undergraduate biology courses in the Fall term of 2013; one of these was an introductory first-year laboratory course (which we refer to as "Biology-first-year level") and the other was a third-year laboratory course (which we refer to as "Biology-third-year level"; see Table 3 for course details).

We wished to gather data and assess item suitability from these two biology classes to see whether the SRBCI was suitable for use in classes being taken by students with different preconceptions and abilities in statistical reasoning. This is an important consideration when developing concept inventories and other multiple-choice instruments, because they may operate differently and be more or less effective when used to assess different populations (Planinic *et al.*, 2010). We reasoned that students in the two courses in which the SRBCI was deployed represented different populations, because those in Biology-third-year level must have completed at least one laboratory course in which they were taught basic statistical reasoning skills by the time they reached their third year, whereas the same was not true of Biology-first-year-level students, who may have been previously exposed to these concepts at the high school level. The tests were conducted in the first week of each course.

We administered the SRBCI in the same way in all sections of these two courses; students were given participation marks (0.5% bonus mark for completing the pretests) but were otherwise neither rewarded nor penalized for their performance. We looked for suspicious patterns in student answers that would have suggested they were guessing (e.g., answering all "A's" or in apparently nonrandom patterns) but found no such examples. The SRBCI items were presented on individual PowerPoint slides projected at the front of the classroom. We provided students with handouts that incorporated background information and figures or data tables for the three experimental scenarios in which the SRBCI items were contextualized, so they could refer back to these if they wished while answering the items, but they did not have hard copies of the items. Students were given sufficient time to read the background information for each scenario when it was presented on an individual PowerPoint slide before the first item set within each scenario appeared. We then gave students 70 s to read and answer each item; we set these timings based on feedback from student

Table 3. Course descriptions and typical enrollment data for the two courses we sampled (Biology-first-year level and Biology third-year level) in these analyses

Course name and description	Typical course enrollment
Biology-first-year level: An inquiry-based lab course investigating the response of organisms to changes in their environment through research and experimentation. Students design their own experiments, implement two trials, and analyze their own data. This course is intended for all life sciences majors but is also open to any student with the prerequisite courses. Assessment is based on written reports, oral presentations, a lab exam, and completion of a series of small assignments.	~1600/yr, 67 lab sections/yr, term 1 and term 2
Biology-third-year level: A lab skills–based course, with lectures integrating many 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. BIOL 121 = prerequisite. Assessment is based on written lab reports plus oral presentations.	~100/yr, 4 lab sections and 1 lecture section/yr, term 1 only

interviews and from our own observations, noting that, when interviewed, students very rarely needed more than 60 s to choose an answer and then thoroughly explain to us their reasoning for selecting it (indicating that they understood the item and were not rushed into guessing). We also gave students a 10-s warning before each new slide (item) appeared on the screen. Each student recorded his/her responses on an optical answer sheet (Scantron). This protocol enabled us to administer the SRBCI in 17 min in each class.

Statistical Analyses: The Use of IRT (the One-Parameter Rasch Model)

Concept inventories and other multiple-choice surveys and instruments have traditionally been validated by using either CTT item analyses (Smith *et al.*, 2008; Kalas *et al.*, 2013) or IRT analyses such as the one-parameter Rasch model (Rasch, 1960; and for examples of its use, see Boone *et al.*, 2010; Planinic *et al.*, 2010; Wallace and Bailey, 2010; Wang and Bao, 2010). While both methods accept that the probability of getting any item on a test correct is directly influenced by the ability of the person and the difficulty of the item, Rasch model analyses assess each individual's ability and each item's difficulty in such a way that they can be scored on the same continuous scale, which confers many benefits.

CTT analyses typically produce metrics such as item difficulty, discrimination, and reliability, which are then used to assess the quality of the test instrument (Wallace and Bailey, 2010). However, these metrics are heavily sample dependent, whereas IRT metrics are not. Because IRT metrics can separately estimate abilities and item properties, useful judgments of individual-item and whole-test quality can still be made even if the surveyed students are not representative of the population (Hambleton and Jones, 1993).

CTT analyses also focus on using students' total scores across the whole instrument to provide information on student ability; these estimates of student ability may not be accurate when questions vary in terms of difficulty. For example, a student who scores 6 out of 12 on an instrument does not necessarily possess twice as much ability in the latent trait that the instrument has been designed to assess when compared with a student who scores 3 out of 12, because not all questions reflect ability to the same extent. In IRT analyses, a standardized linear scale is created for both student ability and item difficulty (Furr and Bacharach, 2014). Figure 1 shows an example of how the linear scale can be used to more easily make meaningful quantitative comparisons between students and items based on their abilities and difficulties respectively.

Finally, CTT analyses provide descriptive statistics, but they do not propose or test a hypothesis model. On the other hand, IRT analyses propose a hypothesis (the Rasch model) and then assess the fit based on the data provided by students who respond to the instrument (the SRBCI), which allows us to test whether the model accurately describes how students respond to individual items (Embretson and Reise, 2000). If the data do not fit the model, this indicates that the instrument and/or the individual items on it may not be fit for assessing the intended trait. As a result, IRT analyses are more useful for deciding whether each item on an instrument should be retained, refined, or discarded.

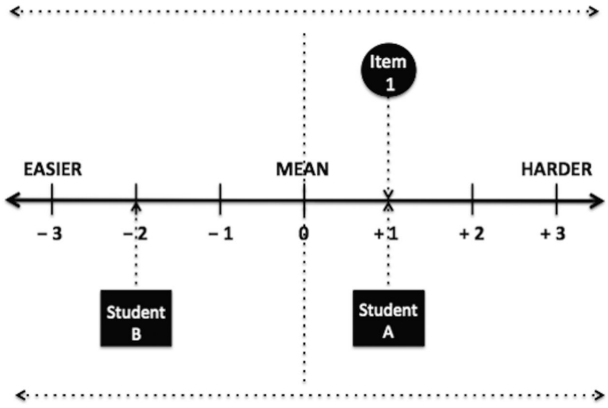

Figure 1. The same linear scale provided by Rasch model analyses can be used to make quantitative assessments of person ability and item difficulty after test data have been collected and fitted to the model. A number of principles apply to such assessments: 1) Items of mean difficulty, and persons with estimated mean ability in the trait being assessed, have difficulty and ability estimates of 0. 2) Items that are more difficult than this mean difficulty and persons who have greater abilities than this mean ability have positive values (>0). 3) The converse is true for items that are easier than the mean difficulty or for persons who have lower abilities than the mean ability (<0). 4) Persons with the same ability estimates as an individual item's difficulty have a 50% chance of answering that item correctly (e.g., Student A has a 50% chance of answering item 1 correctly, while Student B will have a much lower chance of answering it correctly). 5) The linear scale is set in SDs for ability and difficulty (e.g., Student A has an ability in the trait being assessed that is precisely three SDs greater than that of Student B).

It is also possible to analyze data with IRT approaches that include two other parameters (a discrimination parameter and a guessing estimate) in the models. There are trade-offs associated with each approach (Wallace and Bailey, 2010), but we chose to use the one-parameter (Rasch) model due to our relatively small sample size (Harris, 1989) and because others have recently shown the suitability of this method in assessing concept inventory data (Wallace and Bailey, 2010; also see Wright, 1997).

For a more comprehensive treatment of IRT, its conceptual basis, and alternative options for the assumption and model fit–testing procedures, see DeMars (2010) and Embretson and Reise (2013).

Rasch Model Analysis of Biology-First-Year-Level and Biology-Third-Year-Level Data

We performed separate one-parameter IRT (Rasch model) analyses on the SRBCI by using two populations of students at a large research institution in Canada (Biology-first-year level: n = 371; and Biology-third-year level: n = 86; see Table 3). We wished to see whether the SRBCI would operate differently in these two populations, as we expected first-year and third-year undergraduates to perform differently on a concept inventory that assesses statistical reasoning skills. We used the software package R (R Development Core Team, 2013) to perform all analyses, specifically using the package eRm (Mair and Hatzinger, 2007; Mair *et al.*, 2014).

Checking the Two Assumptions of Rasch Model Analysis (Unidimensionality and Local Independence)

Before using difficulty estimates and person abilities provided by the Rasch model analyses, it is important to check that the assumptions of unidimensionality and local independence are met (Wallace and Bailey, 2010; Slocum-Gori and Zumbo, 2011). A unidimensional instrument is one in which individual items all contribute to the underlying construct (statistical reasoning in the SRBCI's case) and are not calibrated too close together or too far apart in terms of difficulty (Planinic *et al.*, 2010). Local independence is achieved when there is no significant correlation between responses to items once the effects of difficulty and ability of the test taker have been taken into account. In other words, a correct (or incorrect) answer to one item should not correlate with a subsequent correct (or incorrect) answer to another item when the residuals are compared (observed responses compared with model-predicted responses). Nonindependence would suggest that responses—and therefore model-produced ability and difficulty estimates—are affected by something other than what the instrument is designed to test (Wang and Wilson, 2005).

We checked for unidimensionality by using Bejar's (1980) method. This method requires the Rasch model to estimate item parameters as normal for the whole SRBCI and then estimate a subset of these item parameters before plotting the resultant item estimates against one another. If the item is unidimensional, then the slope of the line should be ~1 and have an intercept of ~0. This is because the subset item parameters should be very close to the same item parameters derived from the whole analysis. Bejar (1980) recommended selecting a subset of items that might probe different content areas, so we selected six of the 12 SRBCI items that ensured representation of all four core concept groupings (variation in data, repeatability of results, hypotheses and predictions, and sample size). We also ensured our six items fell equally into the three different scenarios (two each from the salmon, squirrel and raccoon, and sunflower scenarios).

We checked for local independence of the SRBCI items by calculating Yen's (1984) Q3 statistic. This statistic looks at the difference between the observed responses and those predicted by the Rasch model before using these residuals to calculate item-by-item correlations. If the items are independent, the correlation statistics for each item pair should be ~0, but Yen and Fitzpatrick (2006) suggest a correlation less than or equal to ±0.20 is low enough to treat as indicating local independence.

Checking Goodness of Data Fit with the Rasch Model

We computed a series of statistics to check that the data from our two sampled populations fit the Rasch model. Specifically, we used Andersen's likelihood ratio (LR) test (Andersen, 1973; Mair *et al.*, 2014) to check whether there was any item bias present; such bias can occur when individuals of the same ability (as characterized by the model) react significantly differently in the way they answer an item (or items). In the event that this occurs, it can signal an item that might be problematic and that should be revised or removed. We then produced goodness-of-fit plots with 95% confidence ellipses to confirm that individual items were unbiased.

As a further check that items were behaving in a homogeneous way, we computed two chi-square ratio metrics, called infit and outfit mean-square (MSQ) statistics, for each item; these statistics are commonly calculated in Rasch analyses (e.g., Planinic *et al.*, 2010; Wallace and Bailey, 2010) and effectively assess the level of randomness in the data (Linacre, 2002) in terms of the way students of distinct abilities answer each item. These statistics are thus useful in indicating whether each item is a useful inclusion in the test instrument. For a given item, the expected value of both the infit and outfit MSQ is 1, but items for which values fall within the range of 0.5–1.5 are considered to be acceptable and productive for measuring a test-taker's ability in the trait the test is designed to measure (Linacre, 1991; although Green and Frantom (2002) suggest a slightly more conservative range of 0.6–1.4). The infit MSQ statistics are information-weighted sums that pay less attention to outliers in the data, whereas the outfit MSQ statistics are calculated from the mean sum of squared residuals (the difference between the observed student score on an item and the score predicted by the model). Large outfit values suggest persons whose ability differs considerably from that of an item's ability level have performed in a way that is unexpected (perhaps students of low ability have answered a difficult item correctly far more frequently than probability anticipated). In contrast, large infit values suggest persons whose ability barely differs from that of an item's ability level have performed in a way that is unexpected (e.g., students with middling abilities have answered a medium-difficulty question correctly far more or less frequently than anticipated). In either case, a statistic outside the range of 0.6–1.4 is indicative of a problematic test item.

Visualizing Item Difficulty and Person Ability

We produced item characteristic curves (ICCs) and item-person maps for all the SRBCI items based on the Rasch model results to characterize the spread of difficulty of each item and to visualize the relative abilities of our Biology-first-year-level and Biology-third-year-level students. ICCs provide a visual assessment of each item's difficulty, and the predicted probability that students of varying abilities will answer it correctly. One common use of these is to read the point on the *x*-axis (the ability level of a test taker) at which there is a 50% chance of the student answering the item correctly (for more information on how ICCs can be used, see Supplemental Figure S1). Item-person maps provide a visual assessment of how all the test items vary in terms of their difficulty levels and also show the frequency distributions of test-takers' abilities on the same map (Yu, 2013).

Assessing Suitability of Using Raw SRBCI Scores to Consider Student Ability

We computed Pearson product-moment correlations between student raw scores and their ability estimates according to the Rasch model results. A close correlation provides support for using the raw scores as a means of assessing student ability in the trait (conceptual ability in statistical reasoning) that the SRBCI was designed to measure (Wang and Bao, 2010).

Ethics Protocol Compliance and Accessing the SRBCI

This study complied with ethics requirements for human subjects as approved by the Behavioural Research Ethics Board at the large Canadian research university where this study was conducted (BREB H09-03080). 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 SRBCI and its accompanying materials for their own use.

RESULTS

Rasch Model Assumptions (Unidimensionality and Local Independence)

We found good evidence that the SRBCI is a unidimensional instrument (testing only one underlying trait of conceptual ability—statistical reasoning), and that the individual items displayed local independence. This same general conclusion held for the Rasch analysis conducted on both the Biology-first-year-level and Biology-third-year-level data. Briefly, the slopes and intercepts of these correlations were very close to the desired values for both data sets (Biology-first-year level: intercept = –0.19, slope = 1.04; Biology-third-year level: intercept = –0.13, slope = 1.01; $R^2 = 0.99$ and $p \leq 0.0001$ in both cases).

We show tables of all the pairwise item correlations conducted (Yen's Q3 statistics) to assess local independence (Supplemental Table S3, a and b). Only two pairwise correlations had Q3 statistics that were outside the recommended range of +0.20 to –0.20 for the Biology-first-year-level data (the item 2 and item 8 correlation was –0.23; the item 3 and item 6 correlation was –0.21), and only three for the Biology-third-year-level data (the item 1 and item 6 correlation was –0.21; the item 3 and item 11 correlation was –0.22; the item 7 and item 8 correlation was –0.21). All of these were only just outside the desired range, while the vast majority of pairwise item correlations were within this range for both data sets (64/66 or 97% for Biology-first-year level, 63/66 or 95.5% for Biology-third-year level).

Goodness of Data Fit with the Rasch Model

The data from both the Biology-first-year-level and Biology-third-year-level SRBCI deployments fit the Rasch model (Biology-first-year level: Andersen LR value = 12.47, $p = 0.33$, $df = 11$; Biology-third-year level: Andersen LR value = 10.02, $p = 0.53$, $df = 11$). We also calculated infit and outfit MSQ statistics for each item in both data sets and all individual SRBCI items fell well within the recommended range of 0.6–1.4 (Supplemental Table S4).

Item-Difficulty and Person Ability Estimates

Individual SRBCI items varied widely in their difficulty as estimated by the Rasch model for Biology-first-year level and Biology-third-year level (Figure 2, A and B, respectively). Item 2 (core conceptual grouping = repeatability of results) was the easiest for both populations, but there were differences in which items the two populations found most difficult; while Biology-first-year-level students found item 1 the most difficult (core conceptual grouping = repeatability of results), Biology-third-year-level students found item 9 (core conceptual grouping = variation in data) the most difficult. However, both populations found these items relatively difficult as evidenced by the items having positive difficulty values (e.g., to have a 50% chance of answering these items correctly, students in either population needed to have abilities > 0). The relative rank order of the items in terms of their difficulty was subtly different between populations. Biology-third-year-level students generally possessed higher abilities than Biology-first-year-level students (note the frequency distribution shifting to the right, i.e., toward higher ability estimates, in Figure 2B compared with Figure 2A).

The ICC curves for individual items indicate more clearly how the difficulty of the SRBCI items varied in the Biology-first-year-level and Biology-third-year-level populations (Supplemental Figures S1 and S2). Note that Supplemental Figure S1 shows two ways in which ICC curves are typically used to provide quantitative information about the relationship between student abilities and item difficulties, focusing on the ways students from each population responded to SRBCI item 1 in our study, while Supplemental Figure S2 shows all ICC curves for each population and each SRBCI item.

The ICC curve for item 1, which was the most difficult for Biology-first-year-level students, shows that those with a mean ability in statistical reasoning (ability estimate = ~0) had a probability of only ~16% of answering this correctly, and students from this population needed to have relatively high ability estimates (~1.85 SDs above a mean ability in statistical reasoning) to have a 50% chance of answering this item correctly (see Supplemental Figure S1). However, Biology-third-year-level students with a mean ability in statistical reasoning (ability estimate = ~0) had a much higher probability of answering this item correctly (~38%), and students from this population needed to have lower relative ability estimates (~0.6 SDs above a mean ability in statistical reasoning) to have a 50% chance of answering this item correctly (see Supplemental Figure S1).

Difficulty estimates for items falling into the four different core conceptual groupings probed by the SRBCI—1) variation in data = items 5, 9, and 10; 2) repeatability of results = items 1, 2, and 6; 3) hypotheses and predictions = items 4, 7, and 12; and 4) sample size = items 4, 8, and 11—also varied relatively widely in both student populations (Supplemental Figure S2).

Rasch model estimates of person ability based on total score were very similar for both Biology-first-year-level and Biology-third-year-level students achieving the same additive scores on the SRBCI (Table 4). There was wide variation of ability levels in the two populations analyzed, but the linear nature of the ability estimates produced by the Rasch model suggests that students scoring 11 out of 12 (the highest achieved in both populations) possessed ability in the trait the SRBCI assesses (statistical reasoning) of ~5.5 SDs in excess of those scoring 1 out of 12 (Biology-first-year level: +2.79 to –2.83; Biology-third-year level: +2.72 to –2.82; see Table 4).

Correlations between SRBCI Raw Scores and Rasch Model Ability Estimates

Pearson product-moment correlations between SRBCI raw scores and ability estimates provided by Rasch model

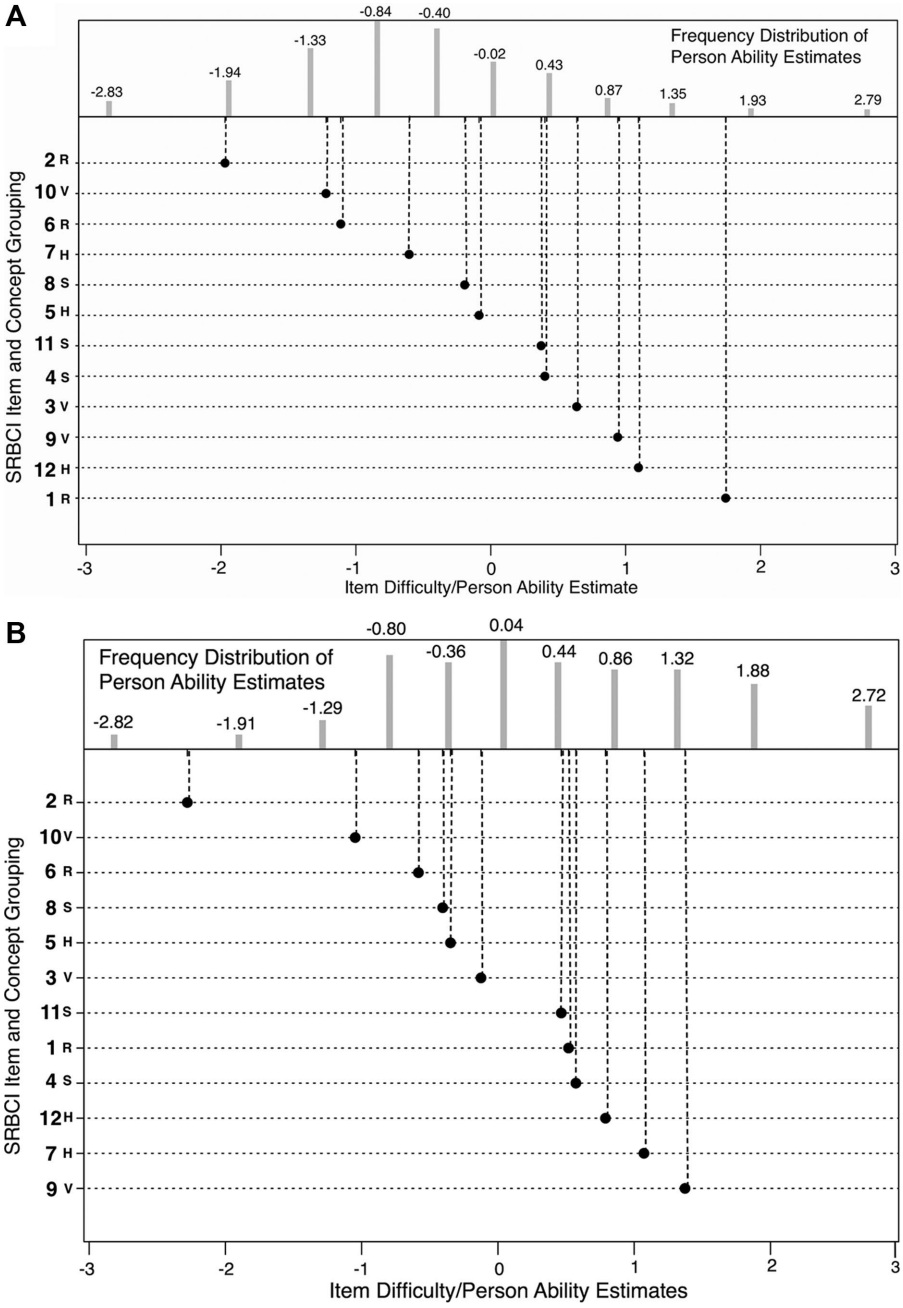

Figure 2. (A) The item-person map derived from Rasch model analysis of student responses to the SRBCI from the Biology-first-year-level population. SRBCI items appear on the *x*-axis in order of their difficulty (easiest at the top, hardest at the bottom). The frequency distribution of student abilities appears in the bars at the top (*y*-axis). The peak frequencies appear below the ability estimate of 0, suggesting that most students have relatively low ability in statistical reasoning at this stage of their education. (B) The item-person map derived from Rasch model analysis of student responses to the SRBCI from the Biology-third-year-level population. SRBCI items appear on the *x*-axis in order of their difficulty (easiest at the top, hardest at the bottom). The frequency distribution of student abilities appears in the bars at the top (*y*-axis). There are more individuals with abilities above the ability estimate of 0, suggesting that most students have relatively high ability in statistical reasoning at this stage of their education (especially in relation to the Biology first-year-level population). (A and B) Letters after items refer to the conceptual grouping to which these items belong (R, repeatability of results; V, variation in data; H, hypotheses and predictions; S, sample size).

analyses were almost linear (Biology-first-year level: R^2 = 0.98, $p \leq 0.0001$; Biology-third-year level: $R^2 = 0.99$, $p \leq 0.0001$).

DISCUSSION

The SRBCI Is a Sensitive Tool for Measuring Student Ability in Statistical Reasoning

Our analyses indicated that the SRBCI is an effective, sensitive tool for assessing student ability in the construct that it measures (conceptual ability in statistical reasoning) and that it is effective at assessing this construct in populations of students who are at different stages of their biological education at university. We found good evidence that all 12 SRBCI items test conceptual ability in a single construct (i.e., the SRBCI appears to be a unidimensional tool), and these 12 items seem to be locally independent. We found only two instances in the Biology-first-year-level analysis in which item–item comparisons were potentially nonindependent, and only three in the Biology-third-year-level analysis (both of out a possible 66 comparisons). The values were only marginally outside the recommended range suggested by Yen (1984) in these cases, and were less frequent than those appearing in a Rasch model analysis of the Star Properties Concept Inventory (Bailey, 2006) conducted by Wallace and Bailey (2010).

Table 4. Comparisons between SRBCI raw scores and estimates of student ability provided by Rasch model analyses of data provided by the two student populations assessed in these analyses

SRBCI raw score	Ability estimate	
	Biology-first-year level	Biology-third-year level
0	–	–
1	–2.83	–2.82
2	–1.94	–1.91
3	–1.33	–1.29
4	–0.84	–0.80
5	–0.40	–0.36
6	–0.02	0.04
7	0.43	0.44
8	0.87	0.86
9	1.35	1.32
10	1.93	1.88
11	2.79	2.72
12	–	–

Goodness-of-fit tests indicated that the data from both our student populations fit the Rasch model, and infit and outfit MSQ analyses indicated that each of the SRBCI's items also fit the model; as a result, we are confident that all 12 of the items are valuable in assessing student conceptual ability in statistical reasoning and, therefore, that none of them should be removed. There were a good spread of difficulty for the 12 items and a relatively good spread of difficulty for the four core conceptual groups of statistical reasoning—1) variation in data, 2) repeatability of results, 3) hypotheses and predictions, and 4) sample size. This adds further support to the assertion that all 12 SRBCI items, all of which test application of different concepts in statistical reasoning, are valuable in assessing student ability in this construct. That the ability estimates for students with the same raw scores were so similar for students in the different populations is another positive mark for the robustness of the instrument and its items.

Student Performance on the SRBCI

As expected, students in Biology-third-year level generally possessed higher abilities than those in Biology-first-year level; however, there was a good range of abilities in both populations. Interestingly, although both populations found the same item to be the easiest of the 12 SRBCI items, they found different items to be the most difficult. It was also interesting to note that the relative order of item difficulty was subtly different between the two populations, which has implications from an instructional viewpoint. Because Biology-third-year-level students have had more instruction involving statistical reasoning concepts at university, it is interesting to look at the items whose difficulties were higher in that population than in the Biology-first-year-level population. The concepts embedded in those items might be the ones that instructors should focus on teaching in greater depth: it is a reasonable expectation that students with more experience in statistical reasoning should find *every* item relatively easier unless non–expert-like aspects of their statistical reasoning have already become part of their long-term memory. Generally, the third-year students in our sample were further along the novice-to-expert spectrum than those in the first-year sample. However, the variation in performance seen within and between the populations and in the difficulty of individual items underlines the complexity of learning; students vary greatly in their ability to transition toward expert-like thinking in statistical reasoning, and relatively few have developed the conceptual abilities we would associate with experts after 3 yr of undergraduate education. A lot of meaningful learning activity (such as struggling, processing, sense-making) is known to go on in the intermediate space between novice and expert (Bass, 2012), while Smith *et al.* (1993) underlined the importance of acknowledging the existence of continuity between novices and experts. Given this, we would not expect learners to progress linearly from novice to expert following instruction in statistical reasoning concepts in biology, or even that some leaps in progression would be easier to make than others. That was broadly what we found in the preliminary interviews (see Supplemental Tables S1 and S2 for examples of student quotes and reasoning aligned to novice, intermediate, and expert answer choices). We now look forward to developing activities and learning interventions to help students develop their statistical reasoning skills in the areas they seemed to find most difficult; we also anticipate looking for patterns in progression to see whether some gains are easier to make than others or whether gains in some areas are linked to gains in others. As Ben-Zvi *et al.* (2015) notes, we still do not know much about the ways students develop statistical reasoning skills. We hope that the SRBCI will provide useful information that characterizes this process.

The Value of IRT Analyses

There has been a call from many researchers in both the sciences and social sciences to use IRT-based approaches rather than CTT approaches when assessing the suitability of instruments to diagnose what they have been designed to diagnose (Ding and Beichner, 2009; Boone *et al.*, 2010; Planinic *et al.*, 2010). With this in mind, we used one-parameter IRT (Rasch) modeling techniques to separately consider the suitability of the SRBCI as an instrument to assess conceptual ability in statistical reasoning of undergraduate students from two different populations. The great benefit of IRT analyses comes from the fact that student-ability and item-difficulty estimates are placed on the same linear scale of measurement, which makes it easier to draw meaningful comparisons between the groups of interest.

CTT-based analyses can still be valuable in assessing the relative ability of an individual (or a group of individuals) on a particular conceptual construct (see Hestenes *et al.*, 1992; Smith *et al.*, 2008; Schlingman *et al.*, 2012; Kalas *et al.*, 2013), while concept inventories in particular are also useful in diagnosing specific examples of non–expert-like thinking (highlighted by the distractors these students choose; see Smith and Knight, 2012). With particular reference to the use of non-IRT analyses, for those who may be deterred from using this method, we showed that there was a virtually linear correlation between student-ability estimates (from the Rasch model fits) and their raw SRBCI scores. This suggests that instructors using the SRBCI could simply use raw scores as a *somewhat* accurate guide to the conceptual ability in statistical reasoning of an individual (or a group of individuals).

While Embretson and Reise (2000) point out that raw scores are always sufficient statistics when analyzing responses with a one-parameter Rasch model, Wang and Bao (2010) reason that students with the same raw scores might have obtained these scores from getting different items correct; such variation is guaranteed in any test, but this highlights the value in considering student performance on individual items. For example, if high-ability students answer a given item correctly less frequently than predicted by that item's ICC, this might suggest a specific example of non–expert-like thinking has become embedded in their thinking and that instructors should focus on clarifying this concept.

Limitations of the SRBCI

The SRBCI's items were contextualized in simple experimental biology scenarios involving well-known plants and animals; however, the concepts being tested relate to simple statistical reasoning and figure/graph interpretation skills. As a result, it is most appropriate for use in biology and in biology-related fields as opposed to other science disciplines that approach and interpret results differently; for example, the level of uncertainty in experimental data that is acceptable is discipline specific. Ben-Zvi and Garfield (2004) underlined that discipline-specific norms exist for what constitute acceptable data arguments, which is why the SRBCI reflects the norms of acceptable practice in biological sciences.

We also note the limitation of our approach in using 95% confidence intervals as visual indicators of estimate sample error (see *Methods*, and Belia *et al.*, 2005; Cumming *et al.*, 2007), but we wanted to prioritize application of statistical reasoning skills based on visual comparisons of data in figures as opposed to providing written statistical information, which would be less likely to require conceptual application.

The SRBCI should be used for formative purposes, so instructors can identify which nonexpert conceptions their students hold at different stages of their teaching. It has not been designed for summative testing purposes; indeed, many would consider 12 items an insufficient number to reliably separate student abilities (Kruyen *et al.*, 2013). If SRBCI had more than 12 items, one would expect there to be less variability and smaller standard errors associated with individual item-difficulty estimations, but we settled on a 12-item instrument due to the need for rapid in-class testing and subsequent analysis.

Finally, we note that precise definitions of statistical reasoning and the concepts that comprise this construct vary (Ben-Zvi and Garfield, 2004). As such, there may be other conceptual groupings not assessed by SRBCI that others would argue are important components of statistical reasoning. Our goal was to produce a tool that focused on the key elements in most definitions and that we knew our students struggled with. We have designed SRBCI to be suitable for undergraduates, to be simple to administer in large classes, and to provide results that are quick and easy to analyze so as to enable instructors to better target their teaching as required by their students.

CONCLUSION

Various agencies have advocated for innovations in teaching that enhance engagement and deeper learning (Woodin *et al.*, 2010; American Association for the Advancement of Science, 2011), and it is vital that we measure the success of our own innovative teaching practices and interventions in a methodical way. The SRBCI is one such tool that could be used to see whether certain methods are resulting in positive learning changes in our biology students. In this paper, we have demonstrated that all of the SRBCI's 12 items provide useful information relating to student conceptual ability in statistical reasoning in two different populations of undergraduates (it can provide useful information in first-year and third-year biology students). The individual SRBCI items that students found most difficult would appear to represent the concepts around which we should design teaching interventions with the aim of improving conceptual thinking in statistical reasoning in our students.

ACKNOWLEDGMENTS

We thank the undergraduate students who took part in the think-aloud interviews and the in-class deployments of the finalized SRBCI and the instructors who kindly found time to incorporate the inventory into their classes. We also thank those instructors and graduate students who served as experts in validating the SRBCI questions. The Teaching and Learning Enhancement Fund of the research university sampled in this study provided the funds for the Questions for Biology project, and the university's Science Centre for Learning and Teaching provided additional funds.

REFERENCES

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

Agus M, Pero-Cebollero M, Guardia-Olmos J, Penna MP (2013). The measurement of statistical reasoning in verbal-numerical and graphical forms: a pilot study. J Phys Conf Ser *459*, 2023.

Allen K (2006). The Statistics Concept Inventory: development and analysis of a "cognitive assessment instrument" in statistics (May 1, 2006). PhD Thesis, University of Oklahoma. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2130143 (accessed 7 July 2014).

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).

Andersen EB (1973). A goodness of fit test for the Rasch model. Psychometrika *38*, 123–140.

Bailey JM (2006). Development of a concept inventory to assess students' understanding and reasoning difficulties about the properties and formation of stars. PhD dissertation, Tucson: University of Arizona.

Bass R (2012). Disrupting ourselves: the problem of learning in higher education. Educause Rev *47*, 21–14.

Batanero C, Godino JD, Vallecillos A, Green DR, Holmes P (1994). Errors and difficulties in understanding elementary statistical concepts. Int J Math Educ Sci Tech *25*, 527–547.

Battisti BT, Hanegan N, Sudweeks R, Cates R (2010). Using item response theory to conduct a distracter analysis on conceptual inventory of natural selection. Int J Sci Math Educ *8*, 845–868.

Bejar II (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. J Educ Meas *17*, 283–296.

Belia S, Fidler F, Cumming G (2005). Researchers misunderstand confidence intervals and standard error bars. Psychol Methods *10*, 389–396.

Ben-Zvi D, Bakker A, Makar K (2015). Learning to reason from samples. Educ Stud Math *88*, 291–303.

Ben-Zvi D, Garfield J (eds.) (2004). The Challenge of Developing Statistical Literacy, Reasoning and Thinking, New York: Kluwer Academic.

Blagdanic C, Chinnappan M (2013). Supporting students to make judgements using real-life data. Aust Math Teach *69*, 4–12.

Blume J, Peipert JF (2003). What your statistician never told you about P-values. J Am Assoc Gynecol Laparosc *10*, 439–444.

Boone WJ, Townsend JS, Staver J (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: an exemplar utilizing STEBI self-efficacy data. Sci Educ *95*, 258–280.

Bowen GM, Roth W-M, McGinn MK (1999). Interpretations of graphs by university biology students and practicing scientists: toward a social practice view of scientific representation practices. J Res Sci Teach *36*, 1020–1043.

Chan SW, Ismail Z (2014). Developing statistical reasoning assessment instrument for high school students in descriptive statistics. Procedia Soc Behav Sci *116*, 4338–4343.

Chance B, delMas R, Garfield JB (2004). Reasoning about sampling distributions. In: The Challenge of Developing Statistical Literacy, Reasoning and Thinking, ed. D Ben-Zvi and JB Garfield, Dordrecht, Netherlands: Kluwer, 295–323.

Chinn CA, Brewer WF (2001). Models of data: a theory of how people evaluate data. Cogn Instr *19*, 323–393.

Claxton G (2007). Expanding young people's capacity to learn. Brit J Educ Stud *55*, 1–20.

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.

Cooper LL, Shore FS (2010). The effects of data and graph type on concepts and visualizations of variability. J Stat Educ *18*, 1–16.

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

Cumming G, Fidler F, Vaux DL (2007). Error bars in experimental biology. J Cell Biol *177*, 7–11.

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*, 540–551.

delMas RC, Garfield J, Ooms A, Chance B (2007). Assessing students' conceptual understanding after a first course in statistics. Stat Educ Res J *6*, 28–58.

delMas RC, Liu Y (2005). Exploring students' conceptions of the standard deviation. Stat Educ Res J *4*, 55–82.

DeMars C (2010). Item Response Theory, New York: Oxford University Press.

Ding L, Beichner R (2009). Approaches to data analysis of multiple-choice questions. Phys Rev Spec Top Phys Educ Res *5*, 1–17.

Donnelly C (2014). The use of case based multiple choice questions for assessing large group teaching: implications on student's learning. Irish J Acad Pract *3*, 112.

Donnelly LA, Boone WJ (2006). Biology teachers' attitudes toward and use of Indiana's evolution standards. J Res Sci Teach *44*, 236–257.

Durrani N, Tariq V (2008). Employers' and students' perspectives on the importance of numeracy skills in the context of graduate employability. CETL-MSOR Conference 2008, Birmingham, UK: The Maths, Stats & OR Network, 18–24.

Embretson SE, Reise SP (2000). Item Response Theory for Psychologists, Mahwah, NJ: Erlbaum.

Embretson SE, Reise SP (2013). Item Response Theory, Mahwah, NJ: Erlbaum.

Fidler F (2006). Should psychology abandon *p*-values and teach CIs instead? Evidence-based reforms in statistics education. In: Proc Seventh International Conference on Teaching Statistics, Salvador, Brazil, Int Assoc Stat Educ.

Forster P (2004). Graphing in physics: processes and sites of error in tertiary entrance examinations in Western Australia. Res Sci Educ *34*, 239–265.

Friel SN, Curcio FR, Bright GW (2001). Making sense of graphs: critical factors influencing comprehension and instructional implications. J Res Math Educ *32*, 124–158.

Furr RM, Bacharach VR (2014). Psychometrics: An Introduction, 2nd ed., Thousand Oaks, CA: Sage.

Garfield J (2002). The challenge of developing statistical reasoning. J Stat Educ (Online) *10*(3), www.amstat.org/publications/jse/v10n3/garfield.html (accessed 7 July 2014).

Garfield JB (2003). Assessing statistical reasoning. Stat Educ Res J *2*, 22–39.

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.

Glazer N (2011). Challenges with graph interpretation: a review of the literature. Stud Sci Educ *47*, 183–210.

Gormally C, Brickman P, Lutz M (2012). Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE Life Sci Educ *11*, 364–377.

Green KE, Frantom CG (2002). Survey development and validation with the Rasch model. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, held 14–17 November in Charleston, SC. portfolio.du.edu/kgreen/Survey, Data Quality_Rasch.pdf (accessed 5 October 2014).

Haller H, Krauss S (2002). Misinterpretations of significance: a problem students share with their teachers? Methods Psychol Res *7*, 1–20.

Hallgrimsson B, Hall BK (eds.) (2011). Variation: A Central Concept in Biology, Burlington, MA: Elsevier Academic.

Hambleton RK, Jones RJ (1993). Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract *12*, 253–262.

Harris D (1989). Comparison of 1-, 2- and 3-parameter IRT models. Educ Meas Issues Pract *8*, 35–41.

Hestenes D, Wells M, Swackhammer G (1992). Force Concept Inventory. Phys Teach *30*, 141–158.

Howell DC (2014). Fundamental Statistics for the Behavioral Sciences, 8th ed., Belmont, CA: Cengage Learning.

Huck SW (2009). Statistical Misconceptions, New York: Taylor & Francis.

Hunter P (2010). Biology is the new physics. EMBO Rep *11*, 350–352.

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

Karpiak CP (2011). Assessment of problem-based learning in the undergraduate statistics course. Teach Psychol *38*, 251–254.

Kim K, Sharma P, Land SM, Furlong KP (2013). Effects of active learning on enhancing student critical thinking in an undergraduate general science course. Innov High Educ *38*, 223–235.

Kruyen PM, Emons WHM, Sijtsma K (2013). On the shortcomings of shortened tests: a literature review. Int J Test *13*, 223–248.

Kubliansky I, Eschach H (2014). Evaluating a contextual-based course on data analysis for the physics laboratory. J Sci Educ Technol *23*, 108–115.

Ladyshewsky RK (2014). Post-graduate student performance in "supervised in-class" vs. "unsupervised online" multiple choice tests: implications for cheating and test security. Assess Eval High Educ *40*, 883–897.

Lawson TJ, Schwiers M, Doellman M, Grady G, Kelnhofer R (2003). Enhancing students' ability to use statistical reasoning with everyday problems. Teach Pschol *30*, 107–110.

Lem S, Onghena P, Verschaffel L, van Dooren W (2013). External representations for data distributions: in search of cognitive fit. Stat Educ Res J *12*, 4–19.

Linacre JM (1991). A User's Guide to WINSTEPS. http://hbanaszak.mjr.uw.edu.pl/TempTxt/Software/winsteps.pdf (accessed 3 March 2014).

Linacre JM (2002). What do infit and outfit, mean-square and standardized mean? Rasch Meas Trans *16*, 878.

Mair P, Hatzinger R (2007). Extended Rasch modeling: the eRm package for the application of IRT models in R. J Stat Softw *20*, 1–20.

Mair P, Hatzinger R, Maier MJ (2014). eRm: Extended Rasch Modeling. R Package, version 0.15–4. http://erm.r-forge.r-project.org (accessed 11 December 2014).

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

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

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

Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: A comparison of the CINS, and open-response instrument, and an oral review. J Res Sci Teach *45*, 1131–1160.

Osborne J (2010). Arguing to learn in science: the role of collaborative, critical discourse. Science *328*, 463–466.

Peskun P (1987). Constructing symmetric tests of hypotheses. Teach Stat *9*, 19–23.

Planinic M, Ivanjek L, Susac A (2010). Rasch model based analysis of the Force Concept Inventory. Phys Rev Spec Top Phys Educ Res *6*, 1–11.

Price RM (2012). How we got here: an inquiry-based activity about human evolution. Science *338*, 1554–1555.

Rasch G (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research.

R Development Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org (accessed 15 August 2014).

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.

Roth W-M, Bowen GM, McGinn MK (1999). Differences in graph-related practices between high school biology textbooks and scientific ecology journals. J Res Sci Teach *36*, 977–1019.

Samuels M, Witmer J, Schaffner A (2012). Statistics for the Life Sciences, 4th ed., Upper Saddle River: NJ: Pearson Education.

Schlingman WM, Prather EE, Wallace CS, Rudolph AL, Brissenden G (2012). A classical test theory analysis of the Light and Spectroscopy Concept Inventory national study data set. Astron Educ Rev *11*, 010107.

Schultz M, Schultz J, Round G (2008). Online non-proctored testing and its affect on final course grades. Business Rev, Cambridge *9*, 11–16.

Silverman D (2011). Interpreting Qualitative Data, 4th ed., Thousand Oaks, CA: Sage.

Simundic AM, Bartlett WA, Fraser CG (2015). Biological variation: a still evolving facet of laboratory medicine. Ann Clin Biochem *52*, 189–190.

Sinatra GM, Kienhues D, Hofer BK (2014). Addressing challenges to public understanding of science: epistemic cognition, motivated reasoning, and conceptual change. Educ Psychol *49*, 123–138.

Slocum-Gori SL, Zumbo BD (2011). Assessing the unidimensionality of psychological scales: using multiple criteria from factor analysis. Soc Indic Res *102*, 443–461.

Smith C (2007). Bootstrapping processes in the development of students' commonsense matter theories: using analogical mappings, thought experiments, and learning to measure to promote conceptual restructuring. Cogn Instr *25*, 337–398.

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

Smith MK, Knight JK (2012). Using the Genetics Concept Assessment to document persistent conceptual difficulties in undergraduate genetics courses. Genetics *191*, 21–32.

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.

Sotos AEC, Vanhoof S, Van den Noortgate W, Onghena P (2007). Students' misconceptions of statistical inference: a review of the empirical evidence from research on statistics education. Educ Res Rev *2*, 98–113.

Styron J, Styron R (2010). Student cheating and alternative Web-based assessment. J Coll Teach *7*, 37–42.

Thomas R, Braganza A, Oommen LM, Muliyil J (1997). Confidence with confidence intervals. Indian J Ophthalmol *45*, 119–123.

Wallace CS, Bailey JM (2010). Do concept inventories actually measure anything? Astro Educ Rev *9*, 010116.

Wang J, Bao L (2010). Analyzing Force Concept Inventory with item response theory. Am J Phys *78*, 1064–1070.

Wang W, Wilson M (2005). Exploring local item dependence using a random-effects facet model. Appl Psychol Meas *29*, 296–318.

Watson JM (2011). Foundations for improving statistical literacy. Statistical J Int Assoc Off Stat *27*, 197–204.

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.

Wooten MW, Cool AM, Prather EE, Tanner KD (2014). Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory. Phys Rev Spec Top Phys Educ Res *10*, 020103.

Wright BD (1997). A history of social science measurement. Educ Meas Issues Pract *16*, 33–45.

Yen WM (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas *8*, 125–145.

Yen WM, Fitzpatrick AR (2006). Item response theory. In: Educational Measurement, 4th ed., ed. R Brennan, Westport, CT: American Council on Education, 111–153.

Yu CH (2013). A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling. www.creative-wisdom.com/computer/sas/IRT.pdf (accessed 3 December 2014).

Zaidan A, Ismail Z, Yusof YM, Kashefi H (2012). Misconceptions in descriptive statistics among postgraduates in social sciences. Procedia Soc Behav Sci *46*, 3535–3540.