ZDM Mathematics Education (2012) 44:883–898
DOI 10.1007/s11858-012-0447-5

ORIGINAL ARTICLE

Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course

Joan Garfield · Robert delMas · Andrew Zieffler

Accepted: 19 July 2012 / Published online: 4 August 2012

Abstract While models are an important concept in statistics, few introductory statistics courses at the tertiary level put models at the core of the curriculum. This paper reports on a radically different approach to teaching statistics at the tertiary level, one that uses models and simulation as the organizing theme of the course. The focus on modeling and simulation—along with inference—was facilitated by having students use TinkerPlots™ software for all modeling and analysis. Results from a 3-month teaching experiment suggest that a course focused on modeling and simulation through randomization and resampling methods in which students learn to think using a powerful and conceptual modeling tool can foster ways of thinking statistically. Furthermore, such an approach seems to help students develop experiences with and appreciation for the science and practice of statistics.

Keywords Statistics education · Modeling · Simulation · Random

1 Introduction

Despite repeated calls for change and attempts to change the content and pedagogy of the introductory statistics course at the tertiary level, there is little to no evidence that substantial changes have taken place and that student outcomes have improved. One piece of evidence to support this claim is an analysis of data gathered over a 6-year period using the comprehensive assessment of outcomes in statistics (CAOS) to measure important learning outcomes in tertiary level first courses in statistics (delMas, Garfield, Ooms & Chance, 2007). Data from 13,917 undergraduates enrolled in a first course in statistics in the United States show the average percent correct on this test when given at the end of a course has remained stable from 2005 to 2011 (see Fig. 1). Since these data indicate that previous reform efforts did not seem to lead to improved student outcomes, there appeared to be a need for a radically different curriculum for the introductory statistics course. The curriculum, called Change Agents for Teaching and Learning Statistics (CATALST), took its name from the 3-year grant from the National Science Foundation in the USA—the CATALST project—which funded the development and study of this curriculum. Inspired by ideas proposed by Cobb (2005, 2007), the CATALST curriculum uses the ideas of chance and models, along with simulation and randomization-based methods, to enable students to make and understand statistical inferences.

A simulation-based approach to inference requires students to create a model with respect to a specific context, repeatedly simulate data from the model, and then use the resulting distribution of a particular computed statistic to draw statistical inferences. CATALST immerses students in this process from the first day of the curriculum, initially having students make informal inferences and then moving to simulation based methods of formal statistical inference. This paper describes the CATALST curriculum, as well as the results of a 3-month teaching experiment (Steffe & Thompson, 2000) designed to study the implementation of the CATALST curriculum and evaluate important student learning outcomes. While the CATALST project has been collecting information to answer a large set of research

J. Garfield (✉) · R. delMas · A. Zieffler
Department of Educational Psychology, University of Minnesota
250 Education Sciences Bldg, 56 East River Road, Minneapolis, MN 55455, USA
e-mail: jbg@umn.edu

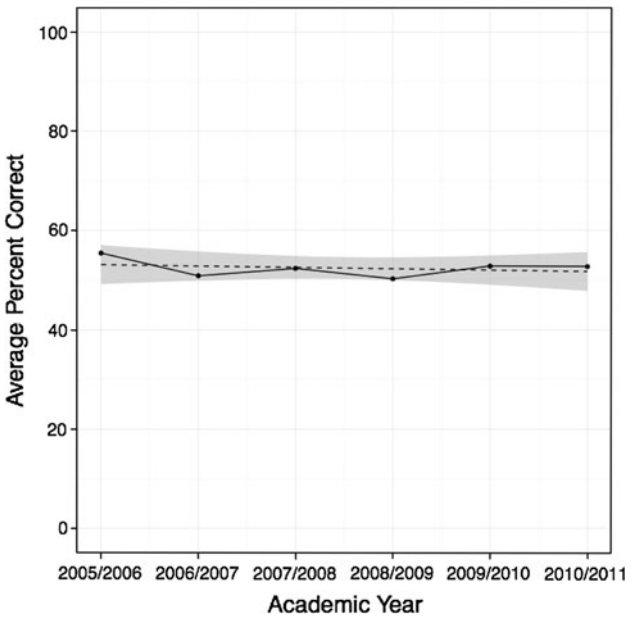

Fig. 1 Stability of performance on the CAOS posttest. The figure also includes the regression line (*dotted line*) and smoother (*light grey area*) of the average percent of CAOS items answered correctly over time

questions, this paper focuses on a subset of those questions related to student outcomes measured at the end of courses implementing the CATALST curriculum. These research questions included:

- How do students respond to the demands of the course and to the exclusive use of TinkerPlots^{TM} software?
- How well are students able to think statistically and really "cook" after taking this course?
- How well do students understand and reason about basic statistical ideas?
- How do students view the discipline of statistics and the nature of statistical problem solving?

2 Background and research foundations

In science, *catalysis* is the acceleration (i.e., increase in rate) of a chemical reaction by means of a substance called a catalyst. In more general terms anything that accelerates a process may be called a "catalyst." The authors of this paper associate catalysts with actions that lead to some profound or major change and use the acronym CATALST to represent the goal of accelerating change in the teaching and learning of statistics. The CATALST project was designed to create curricular materials based not only on Cobb's ideas regarding randomization-based inference (Cobb, 2005, 2007), but also using research in cognition and learning, and instructional design principles. Specifically, research in four areas were foundational to the project: (1) model-eliciting activities (e.g., Lesh & Doerr, 2003; Lesh, Hoover, Hole, Kelly, & Post, 2000); (2) inventing to learn and the role of prior knowledge (e.g., Schwartz, Sears, & Chang, 2007); (3) instructional design principles (e.g., Cobb & McClain, 2004); and (4) the modeling work (Konold et al., 2011) and software developed by Konold and Miller (2011). Finally, the work of mathematician Alan Schoenfeld (1998) on the importance of teaching students mathematics in a way that prepares them to think mathematically, using the metaphor of being able to "really cook" rather than just follow recipes, became a guiding standard for the development of statistical thinking within the CATALST curriculum. Each of these foundational components is briefly described.

2.1 Focus on modeling

A fundamental aspect of statistical practice involves the use of models, for comparison with empirical data or to simulate data to make an estimate or test a hypothesis (Garfield & Ben-Zvi, 2008). Part of developing statistical thinking is to develop ideas of statistical modeling and the importance of selecting appropriate models (see Wild & Pfannkuch 1999), and realizing why Box's statement that all models are wrong, but some are useful (Box & Draper, 1987) is so wise. Models are of particular importance when considering statistical inference. Inferences are made by using a model to compare observed results, typically producing a *p*-value.

Modeling approaches have been advocated in mathematics education, to shift attention from finding a solution to a problem to creating a model that can be generalized and used in another problem (e.g., Doerr & English, 2003). In another use of modeling, Konold, Harradine and Kozlak (2007) describe an instructional approach that helps students develop important statistical ideas of distribution and variability by creating models to simulate data to try to match observed data. Both types of modeling appear to have relevance to the learning of statistics and have traditionally been part of an introductory college course. The use of model-eliciting activities, described in the following section, provides a way to introduce these aspects of modeling into such a course.

2.2 Model-eliciting activities (MEAs)

MEAs are open-ended problems that are designed to encourage students to build mathematical models in order to solve complex problems, as well as provide a means for educators to better understand students' thinking. MEAs are

created to look like authentic, real-world problems and require students to work in teams of 3–4 to generate solutions to the problems via written descriptions, explanations and constructions by "repeatedly revealing, testing, and refining or extending their ways of thinking" (Lesh et al., 2000, p. 597). MEAs are based on six principles (Lesh et al., 2000). Among other requirements, these principles state that the problem posed during an MEA must motivate students to both construct a model in the solution, and assess how well their constructed model works. Aside from being meaningful and realistic to students, the MEA must also lead to a solution that can be used in another problem (i.e., is generalizable). The use of model-eliciting activities has been shown to lead to significant forms of learning (Lesh et al., 2000), and has led to dramatic and positive results in mathematics and engineering education (Moore, Diefes-Dux, & Imbrie, 2007, 2006; Diefes-Dux, Imbrie, & Moore, 2005; Zawojewski, Bowman, Diefes-Dux, 2011). MEAs appear to have promise in the statistics classroom, by exposing student to the kinds of messy, real-world problems that do not have one clear solution. The authors of this paper reviewed MEAs related to statistical content and created new MEAs that would fit the content of the course described in this paper.

2.3 Instructional design principles

Cobb and McClain (2004) offer a model for instructional design that is based on educational research about how students learn and how to design effective instruction to develop statistical reasoning. Their instructional design principles lead to activities that:

- Have students make conjectures about data that can be tested.
- Are focused on central statistical ideas.
- Are built on the investigative spirit of data analysis.
- Are developed to enable teachers to achieve their instructional agendas by building on the range of data-based arguments that students produce.
- Develop students' reasoning about data generation as well as data analysis.
- Integrate the use of technological tools that support students' development of statistical reasoning and allow them to test their conjectures.
- Promote classroom discourse that includes statistical arguments and sustained exchanges that focus on significant statistical ideas.

These principles were used to develop lesson plans and activities for the CATALST curriculum that encouraged students to make and test conjectures, work in groups while using technology-tools, and engage in whole class and small group discussions.

2.4 Learning to think statistically

Many current introductory statistics courses at the tertiary level present students with a wealth of material covering many topics and procedures, and do not appear to be leading to desired students outcomes. Students do not appear to remember what they have learned, and are generally not able to transfer their knowledge to more advanced topics or new material outside the class (see Garfield & Ben-Zvi, 2008). Students enrolled in such courses seem to be developing basic statistical literacy but do not appear to be developing the often desired goals of statistical reasoning and thinking. Using a metaphor introduced by Schoenfeld (1998), these types of statistics courses are teaching students how to follow "recipes", but not how to really "cook". That is, even if students leave these classes able to perform routine procedures and tests, they do not have the big picture of the statistical process that will allow them to solve unfamiliar problems and to articulate and apply their understanding. Harper and Edwards (2011) point out that students who learn mathematics in the cookbook way of following procedures do not have opportunities to develop their own methods of investigation or develop a full appreciation for the potential of the subject area, a concern that applies equally to learning statistics. On the other hand, someone who knows how to "cook" knows the essential things to look for and focus on, and how to make adjustments on the fly.

The CATALST curriculum was designed so that one of the outcomes would be that students would learn how to "cook" (i.e., do statistics and think statistically). The goal was for students to approach a statistical problem not by applying a formal procedure (e.g., carrying out a *t*-test) but instead, to consider: what is an appropriate model to use to generate data, what will be considered strong enough evidence in testing an observed result, and how data should be used to estimate a standard error for estimating a parameter or a difference in parameters. This type of "cooking" is basic—we did not prepare "gourmet chefs" in a 15-week, 3-h-a-week course—but instead developed the skills that could be used in subsequent courses, as well as in daily life. While students were not learning how to compute a *t*-test, they were learning how to examine and judge evidence gathered from two groups and how that evidence is compared to what would be expected if there really were no difference between the two groups.

The general "cooking" method taught in the class is the exclusive use of simulation to carry out inferential analyses. Activities were developed that require students to develop and apply this type of "cooking". Students practiced setting up models, used models to simulate data, examined distributions of simulated data, evaluated an observed result within a distribution, and used that

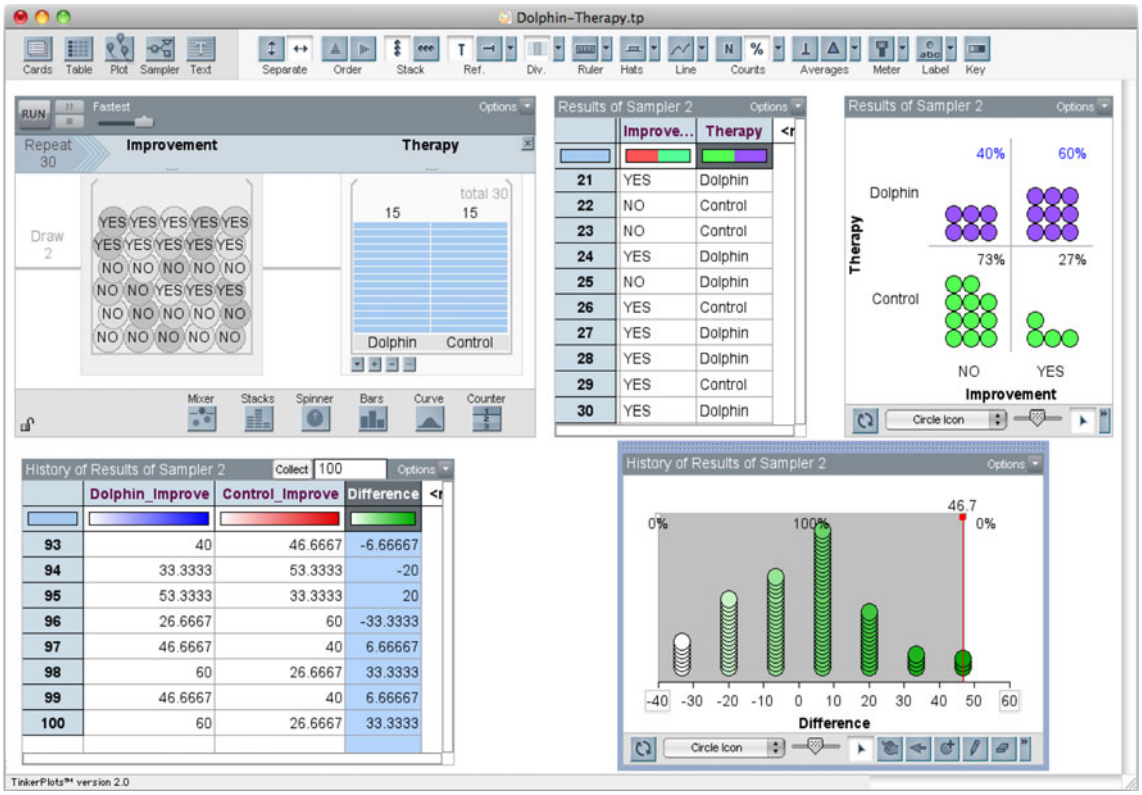

Fig. 2 A screenshot of the use of TinkerPlots™ in the CATALST curriculum. The example shows a simulation used to carry out a randomization test for categorical outcomes

distribution to estimate a standard error. This general cooking method allowed students to conduct a variety of tests (one sample and two sample tests of mean, medians and proportions and even standard deviations) and to estimate a variety of parameters.

2.5 Modeling and simulation

In order to perform the modeling and simulation tasks needed for the curriculum, a software tool was needed. Two tools used to simulate data in teaching introductory concepts of statistics are Fathom (Finzer, 2012) and TinkerPlots™ (Konold & Kazak, 2008; Konold et al., 2011; Konold & Miller, 2011). Biehler and colleagues (Maxara & Biehler, 2006, 2007; Biehler & Prömmel, 2010), used Fathom® for modeling and simulation and discussed some of the challenges in developing students' competency to use Fathom® for these methods. While Fathom® appeared to have the capability to perform the types of modeling and simulation needed, TinkerPlots™ software was chosen instead because of the unique visual capabilities it has, allowing students to see the devices they select (e.g., sampler, spinner) and to easily use these models to simulate and collect data, which allows students to examine and evaluate distributions of statistics in order to draw statistical inferences. Although the software was developed for use in elementary and secondary classes, its capabilities provide a unique and novel way for tertiary students to learn to think statistically as they consider, develop, and use models to draw inferences.

Figure 2 shows a screenshot of one example of how the TinkerPlots™ software is used in the CATALST curriculum. The simulation depicted in the screenshot was used to test the result of a published research study (Antonioli & Reveley, 2005) that investigated whether or not swimming with dolphins is therapeutic for patients suffering from clinical depression. In the study, 66.7 % improved in the dolphin therapy group compared to 20 % who improved in the control group. The figure shows key components from the TinkerPlots™ simulation.

The upper-left part of Fig. 2 shows the sampler. This particular sampler includes the improvement (*YES* or *NO*) of each of the 30 subjects (13 improved and 17 did not) in a mixer and the two conditions of the therapy (*Dolphin* or *Control*) as stacks. Under the null hypothesis of no difference between the two therapies, the sampler will randomly assign 15 improvement balls (representing the subjects) to the dolphin therapy condition and 15 improvement balls to the control condition. One of those possible random assignments is shown in the table and plot

on the upper-right side of Fig. 2. Students use TinkerPlots™ to collect summary measures from that plot. For example, in this simulation they would collect the percentage of subjects that have improved (*YES*) in each of the two therapy conditions. These values are collected into a *collection table* (lower-left side of Fig. 2; the 100th row corresponds to the percentages collected from the plot in the upper-right side of Fig. 2). In this example, students added a third variable to the collection table that computed the difference in the percentage who improved between the two therapy conditions. The *Collect* button in the *collection table* allows students to repeat the process multiple times. The lower-right side of Fig. 2 shows a plot of the differences that were collected across 100 randomizations and identifies the percent of the simulated percentage differences with a difference larger than the observed percentage difference (46.7 % points). Students can then evaluate the original result obtained from the study within this plot.

3 The CATALST course

During the first 2 years of the CATALST project, MEAs and subsequent activities were developed, pilot tested, and revised. The third year consisted of two sequential teaching experiments (Steffe & Thompson, 2000) in which the entire curriculum was taught, observed, studied, and modified. Taking a teaching experiment approach, artifacts of students understanding (e.g., homework assignments, assessments, class session field notes) were continuously examined at weekly meetings by the research team to identify aspects of the curriculum that were and were not effective. Modifications to the curriculum also followed aspects of design experiments (Cobb, Confrey, diSessa, Lehrer & Schauble, 2003): modifications were made to activities based on reasoned conjectures of design features that would promote targeted learning, and the effect of changes based on observation of the first teaching of the entire curriculum were observed and evaluated during the second teaching.

The current version of the CATALST curriculum consists of three units: (1) *Chance Models and Simulation*, (2) *Models for Comparing Groups*, and (3) *Estimating Models using Data*. The first two units begin with an MEA, which is used to create the prior knowledge for the following activities in the unit. These MEAs are based on a real statistical inquiry and use real data. Aside from engaging students in an interesting context, the MEA is designed to motivate and prepare students to learn the relevant content in each unit, and to stimulate statistical thinking. In order to solve the problem posed in the MEA, students invent and test models that promote statistical reasoning and thinking related to the topic of each unit.

Subsequent activities are built on ideas of modeling and simulation, with "the core logic of inference" as the foundation (Cobb, 2007, p. 13). When applied to randomized experiments and random samples, Cobb refers to this logic as the "three Rs": *randomize*, *repeat*, and *reject*. The CATALST project generalized this logic for a broader simulation-based approach to inference as follows:

- *Model.* Specify a model that will generate data to reasonably approximate the variation in outcomes attributable to the random process–be it in sampling or assignment. The model is often created as a null model that may be rejected in order to demonstrate an effect.
- *Randomize and repeat*. Use the model to generate simulated data for a single trial, in order to assess whether the outcomes are reasonable. Specify the summary measure to be collected from each trial. Then, use the model to generate simulated data for *many trials*, each time collecting the summary measure.
- *Evaluate*. Examine the distribution of the resulting summary measures. Use this distribution to assess particular outcomes, evaluate the model used to generate the data, compare the behavior of the model to observed data, make predictions, etc.

A final activity in the first two units presents an expert's solution to the initial MEA used in that unit and has students use what they have learned in the unit to use this approach with the original data presented in the MEA. Each unit has a set of learning goals and also includes a set of visual diagrams of the modeling used in the unit. The following sections provide more detail on each of the three units in the CATALST curriculum.

3.1 Chance models and simulation

The *Chance Models and Simulation* unit begins with the *iPod Shuffle MEA*, as a way to engage students in considering what randomly generated data would look like and confronts their intuitive ideas about random sequences of data. They are given a problem concerning an iPod user who believes that the playlists generated by the Shuffle feature on his iPod are not random. Students are given multiple playlists (data) that were randomly generated from the same music library (8 artists, 10 songs each) and are asked to come up with characteristics that describe the playlists. They then use these characteristics to create "rules" to determine whether a given playlist has NOT been randomly generated.

The most common rules that students created were similar to these:

- If an artist is repeated more than 3 times in a row, the playlist is not random.

- If an artist appears more than 6 times in a playlist, it is not random.
- If a playlist does not contain at least 7 of the 8 artists, it is not random.
- If all artists are represented proportionally, the playlist is not random.

The students test their "rules" and modify them (i.e., they modify their models), using additional randomly generated playlists. At the end of the activity, the students are given the three playlists that the iPod user claims are not random, and use their "rules" (i.e., their models) to decide what the evidence suggests. Most students judge that there is not convincing evidence that the playlists are not randomly generated.

In subsequent activities students encounter and build probability models. Initial models are based on simple random devices (e.g., coins, dice) and are used for simple modeling of real-life phenomena (e.g., birthrates, 'blind guessing'). From these models, students generate empirical data. As they progress through the unit, students also learn how to generate empirical data from a model in which each trial is dependent on a stopping rule. Throughout the unit, the overall purpose for modeling remains the same—students generate data from a model in order to judge whether a particular observed outcome is likely to have occurred by chance. In order to evaluate the outcome in question, students also learn about how simulation results are used to examine conjectures or hypotheses about real-world phenomena. They also learn to examine the 'unusualness' of an observed result under a particular model and assess the strength or degree of evidence against the conjectured/hypothesized model. However, the formal term "*p*-value" is not introduced in this unit.

Fundamental ideas related to probability are also introduced in the first unit of the *Chance Models and Simulation* unit. Aside from the primary focus of informally introducing ideas related to inference, the instructional activities in this unit were also purposefully designed to help students:

- Understand that human intuitions about randomness/probability may be faulty
- Understand that randomness/probability cannot be outguessed in the short term but patterns can be observed over the long term.
- Understand that simulation can be used to investigate probabilistic outcomes and model things that happen by chance
- Understand that simulation can be used to determine whether a particular result could have happened just by chance
- Understand that different chance models lead to different simulation results (coins vs. dice)
- Understand that there are predictable patterns/characteristics of simulation results based on repeatedly sampling/generating random data (e.g., a bell shape from a graph of sample statistics).

At the end of the unit, students again visit the *iPod Shuffle* problem, and use the modeling tools and methods they have learned to simulate and examine random playlists from the same music library, giving them more technical ways to confirm or contradict their original judgments about the playlists. This final activity also presents the logic of inference and provides a transition to the following unit where students learn about the *p*-value and how to use it as a way to quantify strength of evidence (in a qualitative way) against a particular model.

3.2 Models for comparing groups

The second unit, *models for Comparing Groups*, extends the ideas of modeling, simulation and hypothesis testing. As the name suggests, the activities in this unit focus on group comparison. Study design, random assignment and random sampling, and the role of variation play a central role in this unit. This unit again begins with an MEA that begins with a media article about problems with airline reliability in departure and arrival times. Students are then presented students with a small subset of real data from a much larger data set, for multiple airlines that fly between two cities. Students are asked to create a model and use a model (set of rules) it to judge which airline is more reliable.

Following the MEA, an activity is used to build informal ideas and vocabulary regarding the description and summarization of distributions (e.g., shape, center, variation). After this, several class days are spent on the randomization test (for more detail on this test see Zieffler, Harring & Long, 2011). Students use this method to model the variation in a statistic due to chance (in this case random assignment) under the assumption of no group differences (i.e., the null model). In the evaluation of the simulation results (i.e., the randomization distribution) students continue to develop ideas related to characterizing the variation (e.g., the distribution is symmetrically centered around 0). The quantification of how likely the observed result (i.e., *p*-value) is under the model of no group differences and assessment of the model continue in this unit in a more formal setting.

The randomization method is introduced with both quantitative and categorical outcomes. After using the method, students are more formally introduced to ideas about random assignment and its importance in drawing inferences regarding group differences. Basic ideas of design (e.g., random assignment and sampling) are also introduced in this unit. The focus is for students to consider the study design as they are drawing inferences. This point

is reiterated as students encounter comparisons involving randomly assigned, randomly sampled, and purely observational data. Lastly, issues related to the use of inferential test results in making decisions are explored. Students experience an activity in which they are introduced to ideas of type I and type II errors, as well as to specificity and sensitivity. Once again, the unit ends by revisiting the original problem from the Airline Reliability MEA, applying the ideas and methods used in the unit to produce a more expert solution.

3.3 Estimating models using data

In the last unit, the focus is on estimating models using sample data. This unit begins with an activity that explores how different sampling methods may affect the parameter estimates that are made. Students draw both a non-random and random sample from a known population and examine whether the estimates are biased or unbiased (accuracy). They also examine how sample size affects the estimates and are informally introduced to the idea of precision. Lastly, they examine how these properties change under different sample sizes.

The second activity in this unit is one that formalizes a summary measure of variation (standard deviation). This is introduced as a way of estimating variation, and also for providing a more complete summary of a distribution when paired with an estimate such as the mean. Students then use the standard deviation to compute the variation in a distribution of collected summary measures from a simulation (i.e., standard error). Students are introduced to the nonparametric bootstrap to obtain an estimate of this measure. The nonparametric bootstrap uses the observed sample data as a proxy for the population, and resampled data sets are then randomly drawn (with replacement) from the observed data. A summary measure is computed from this resampled data and the process is repeated many times. The variability of these summary measures tends to approximate the theoretical standard error. [For more theoretical and mathematical detail of this methodology, see Efron (1981) or Efron & Tibshirani (1993)].

The bootstrapped standard error (SE) is initially introduced as a measure of precision—how variable an estimate would be from sample to sample. Later in the activity, students compute a margin of error using ± 2SE to obtain an interval estimate that accounts for sampling variation. Finally, the idea of confidence is introduced by having students randomly sample from a known population and compute interval estimates for the many such samples. Each interval can be evaluated to determine if it includes the known parameter, and through this evaluation, confidence is related to the method's effectiveness across all possible samples.

In this unit, students also learn about effect size via the two-group comparison. Under the assumption of group differences, students learn to bootstrap using a model that resamples from each group separately. This is in contrast to the randomization test introduced in Unit 2 in which the data from each group is combined prior to re-sampling. This contrast is made explicit to students when they evaluate simulation results and examine their conjectures of where the resampled distribution is centered (at 0—no group difference; not at 0—group difference). As in the previous activity, students examine the precision of the estimate (variability from potential random samples or random assignments) and then compute an interval estimate for the true effect size.

This unit concludes with a transition to non-simulation based statistical methods and terms (e.g., *t*-test, confidence interval), so that students can see how the results from such methods are interpreted in a similar way to the results of the "no differences" tests and interval estimates they have learned to carry out in the CATALST curriculum.

3.4 Promoting statistical thinking

Students' statistical thinking is developed carefully throughout the three units in several ways. For example, while detailed instructions for using TinkerPlots™ software are provided at the beginning of the course, this scaffold is gradually removed throughout the subsequent activities. In this way, students are being taught how to "cook" rather than just "follow recipes". Most activities are built on real research studies and data to engage students with the content and to show the nature of real-world statistical problems. Actual research articles that provide the relevant data are included in out-of-class reading assignments.

The out-of-class activities (homework) were created to build on and extend the ideas that students experienced during the in-class activities. These activities also provide additional instruction and practice for using the TinkerPlots™ software. An illustrated summary and synthesis of the modeling and simulation process (see Fig. 3) used in each activity is provided to the students and discussed during almost every class session. These visual illustrations of the simulation process, inspired by earlier visual diagrams by Saldanha and Thompson (2003, p. 267) were designed to help students understand the commonalities across the different simulations carried out in the course, which in turn, would translate to a deeper understanding of the modeling and simulation process. For example, Fig. 3 shows a visual summary of the process students might use to solve the following problem:

> In each box of Munchy Crunch cereal, there is one of six possible prizes. Each box contains exactly one

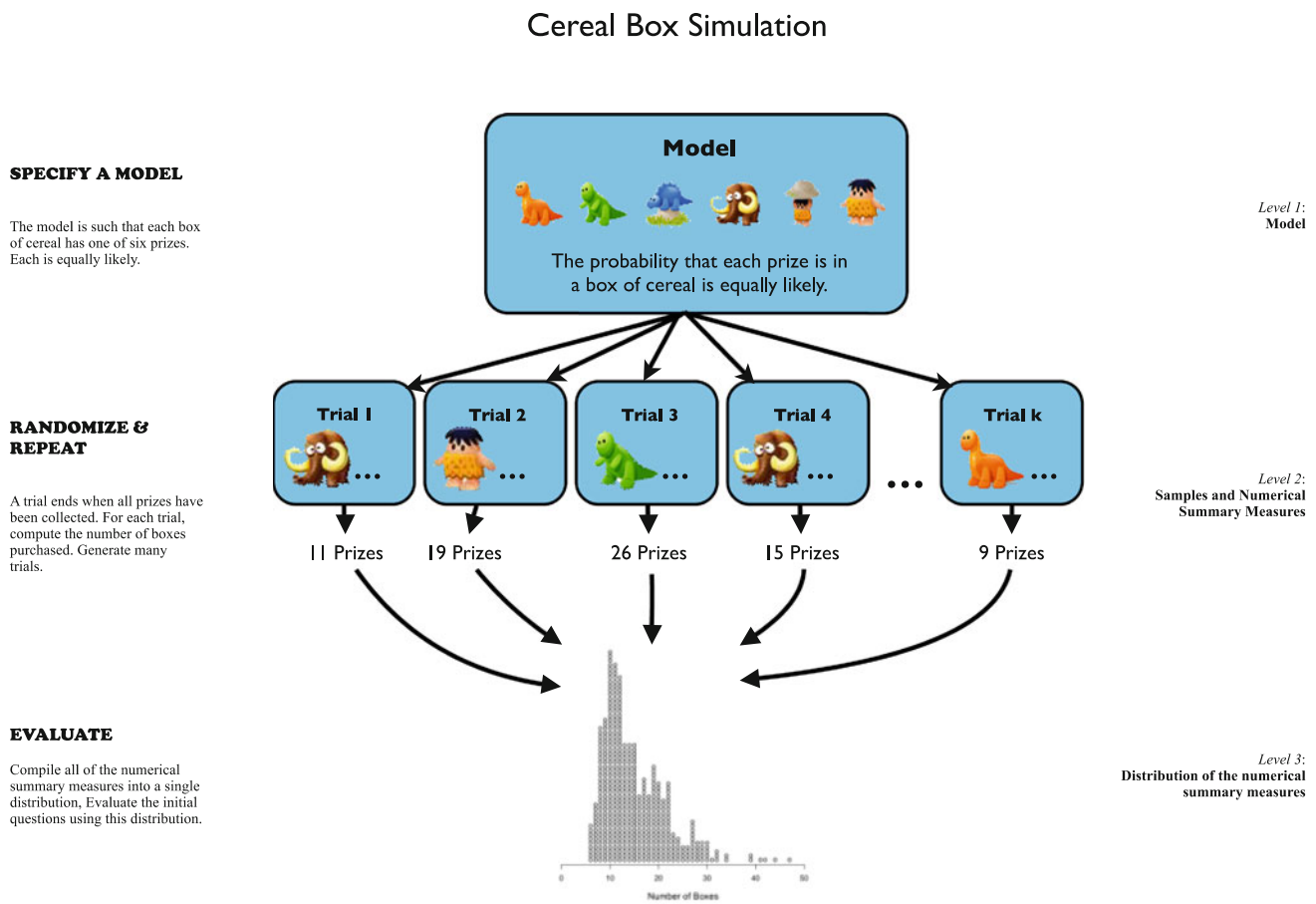

Fig. 3 An illustrated summary of the modeling and simulation process for an activity in Unit 1 of the CATALST curriculum

prize, and we can assume the manufacturer placed the prizes in the boxes at random. Imagine you are interested in collecting all six possible prizes, and you would like to know how many boxes of cereal you can expect to buy in order to collect all six prizes.

Each part of the simulation process depicted in Fig. 3 also corresponds to unique components in TinkerPlots™. For example, the *Specify the Model* section of Fig. 3 corresponds to setting up an appropriate sampling device in TinkerPlots™. The section labeled *Randomize and Repeat* illustrates the process of carrying out a trial in which "prizes" are sampled with replacement until all six are drawn. The total number of "prizes" drawn in the trial (the measure in TinkerPlots™) is recorded and collected, and this process is repeated many times. The *Evaluate* section of the figure shows the evaluation of the distribution of collected measures.

4 Methodology

In spring semester of 2011, the current version of the CATALST curriculum was taught in three sections of an introductory statistics class for liberal arts students at the University of Minnesota in the Department of Educational Psychology. Students who enroll in this course are typically not mathematics or statistics majors, and do not tend to be majoring in one of the sciences or related areas (e.g., physics, engineering). Each class session lasted 75 min, with 2 class sessions each week over a 14-week semester. Students could enroll in one of three different sections of the course, which met on different days and times. While a different instructor taught each course section, each was a doctoral student (pursuing a Ph.D. with an emphasis in statistics education) who had received training in teaching the CATALST curriculum. The curriculum was also implemented in a section of statistics for honors students at North Carolina State University taught by a second-year assistant professor, who was included as part of the instructional team. The instructional team, along with faculty researchers, met weekly to discuss the material and the implementation of the curriculum. Additional weekly meetings were used to plan and discuss day-to-day issues regarding the course.

Data were gathered from 78 students enrolled in the CATALST course at the University of Minnesota and 24

Researchers surveyed 1,000 randomly selected adults in the United States. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling the adults reported typically collecting in a week.

Please select the best interpretation of this result.

a **We cannot conclude whether earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.**

b This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.

c This result indicates that earning more money causes people to recycle more than people who earn less

Fig. 4 An item related to data collection and method of analysis in relation to types of conclusions that are allowed (correct answer is in *bold*)

students enrolled in the course at North Carolina State University (n = 102). The student characteristics were similar across all four courses (e.g., 65–85 % females, 55–75 % sophomores and juniors). The students at the two institutions did differ with respect to the students' declared majors. At North Carolina State University, all of the students had a declared major with about 70 % declaring a major with an emphasis in mathematics or the sciences (e.g., education major with a mathematics specialization; accounting; chemistry), whereas across the three sections at the University of Minnesota, about 20 % had not declared a major and about 65 % majored in an area of the arts or liberal arts (e.g., child psychology; urban studies; acting).

Data were gathered from classroom observations, student assignments, exams, midterm feedback forms, and end of semester assessments. This paper focuses solely on the data gathered from the end of semester assessments to answer the research questions on student outcomes stated previously. These assessments are described in the following section.

4.1 Assessment instruments

Three instruments were developed and used to gather summative data to provide information on students' understanding of content and on their attitudes toward the course. A comprehensive content-related exam consisting of two instruments—the Goals and Outcomes Associated with Learning Statistics (GOALS) and the Models of Statistical Thinking (MOST) assessments—was developed to assess students' achievement of the broader course learning outcomes. The third instrument—the Affect Survey—was designed to assess students' attitudes and perceptions about aspects of the course, what they had gained from the course, as well as about their perceptions of the value of statistics.

4.1.1 Goals and Outcomes Associates with Learning Statistics

The GOALS instrument included 20 forced-choice items and three open-ended items designed to measure students' statistical reasoning. Sixteen of the items were based on items initially created for the CAOS Test (delMas et al., 2007), with four of these items identical to the corresponding CAOS items and the other 12 items based on modifications. The remaining seven items had been created for evaluations of other curriculum projects that focused on the use of simulation methods for drawing inferences. GOALS items consisted of the following types:

- Items related to design and method of analysis in relation to types of conclusions that are allowed. An example item is provided in Fig. 4 (n = 5).
- Items on interpretation of graphical representations of data. One item involved interpreting a scatter plot and the other two questions were based on a comparison of two dot plots, representing two treatments in an experiment (n = 4).
- Items on reasoning about variability in samples of data or among samples of data. An example item is provided in Fig. 5 (n = 4).
- Items on interpretation of confidence intervals. There were three different interpretations offered for a particular confidence interval, and each was to be judged as valid or invalid (n = 3).
- Items on using and interpreting results of modeling and simulation to make an inference including interpretation of a *p*-value. The context for these items is provided in Fig. 6. Following this context, a simulation was described and data generated from the simulation were presented. The items included various questions such as finding the *p*-value and interpreting that result in the context of the experiment (n = 7).

A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag.

Which bag is more likely to have more than 70% brown candies?

a Sam, because there is more variability in the proportion of browns among larger samples.

b **Kerry, because there is more variability in the proportion of browns among smaller samples.**

c Both have the same chance because they are both random samples.

Fig. 5 An item related to reasoning about variability among samples of data (correct answer is in *bold*)

A research question of interest is whether financial incentives can improve performance. Alicia designed a study to test whether video game players are more likely to win on a certain video game when offered a $5 incentive compared to when simply told to "do your best." Forty subjects are randomly assigned to one of two groups, with one group being offered $5 for a win and the other group simply being told to "do your best." She collected the following data from her study:

	$5 incentive	**"Do your best"**	**Total**
Win	16	8	*24*
Lose	4	12	*16*
Total	*20*	*20*	*40*

Fig. 6 The context for the items on using and interpreting results of modeling and simulation to make an inference including interpretation of a *p*-value

The 16 items in the first four categories comprise the GOALS items that were based on CAOS Test items and represent assessment of statistical literacy and reasoning that is typically covered in a first course in statistics at the tertiary level.

4.1.2 Models of Statistical Thinking

The MOST assessment was designed to measure students' statistical thinking. The eight items on the assessment are based on four "real-world" contexts in which students are given a situation and observed data in the form of a summary statistic, and asked to make an inference or a judgment about the observed data. Two contexts involve an inference based on a single statistic, a third provides a situation where two randomly assigned groups were compared, and the last context involves an estimate of a population parameter. The description of each context is followed by a set of open-ended and forced choice format questions. Some of these items had originally been created to use in student interviews as part of a previous research study focused on the development of students' reasoning (see Zieffler, Garfield, delMas, Isaak, Ziegler, & Le, 2011). See Fig. 7 for an example of a MOST item.

The intention of the MOST assessment was to have students describe how they would solve the problems, not to actually use statistical software or perform computations. This way, whether students were taught using simulation based, or non-simulation based methods, they would be able to present their answers using the methods they had learned. Additional and more detailed aspects of statistical thinking were probed and examined in a set of student interviews, reported in delMas, Zieffler and Garfield (in review). For the research in this paper, we were interested in how well students enrolled in a CATALST course were able to describe a reasonable way to make an inference regarding a particular context.

A holistic scoring rubric was created to determine the extent to which students were exhibiting statistical thinking as evidenced in their written explanations. The students' explanations were classified as exhibiting complete thinking, partially complete thinking (at least three of the five

Some people who have a good ear for music can identify the notes they hear when music is played. One note identification test consists of a music teacher choosing one of the seven notes (A, B, C, D, E, F or G) at random and playing it on the piano. The student is standing in the room facing away from the piano so that he cannot see which note the teacher plays on the piano. The note identification test has the music student identify 10 such notes.

This note identification test was given to a young music student to determine whether or not the student has this ability. The student correctly identifies 7 notes out of the 10 that were played. Explain how you would use what you learned in this class to determine how surprising this result is and whether it is strong evidence that the student has the musical ability to accurately identify notes? *(Be sure to give enough detail that someone else could easily follow your explanation.)*

Fig. 7 An example of an item on the MOST assessment. Two additional questions followed this item based on the music test context

aspects included in the response), or incorrect thinking. The aspects of complete statistical thinking that we were looking for in the CATALST students' explanations included:

- Describe an appropriate model to use to simulate data
- Specify the need to generate multiple samples of simulated data and how many to generate
- Describe the outcome of interest to examine and collect across samples of data
- Describe how to find the *p*-value from the distribution of sample statistics
- States how the *p*-value would be evaluated to make a judgment.

The following student's response is an exemplar of the type of statistical thinking that we were looking for:

> "I would set up a sampler on TinkerPlots with two linked devices that randomly chose 1 out of 7 notes each time, and then calculate the number of times right. Repeat this 100 times and calculate a *p*-value by seeing how many times a result as or more extreme than the observed (7/10) is obtained. *p*-value below 0.05 is strong evidence against the null model".

4.1.3 Affect Survey

Students also completed an 11-item attitudinal assessment called the Affect Survey. Items were written to assess students' attitudes and perceptions about aspects of the course, what they had gained from the course, as well as about their perceptions of the value of statistics. Each item had four possible response options: strongly disagree, disagree, agree, strongly agree. These items were given to students to complete anonymously during the last week of the course.

4.2 Data analysis

The research design did not use random assignment in order to compare students who were and were not taught using the CATALST curriculum. Also, it cannot be argued that the sample is representative of all students who generally enroll in a liberal arts or general education undergraduate statistics course. The sample is, arguably, representative of students who typically enroll in such a course at the University of Minnesota. The purpose of the study was to use responses to carefully constructed assessment items to create a picture of the learning outcomes for the participating students with respect to their understanding of statistics and statistical inference. To this end, descriptive statistics are reported and interpreted to address each of the four research questions. Some comparisons are made to a national sample with respects to students' responses to the GOALS items, but statistical tests are not conducted because of the limitations stated above.

5 Results

Assessment data were analyzed and are summarized below as they help provide preliminary answers to the four research questions.

5.1 How do students respond to the demands of the course and to the exclusive use of TinkerPlots™ software?

Six of the items on the Affect Survey were used to provide an answer to this question, as shown in Table 1. In Table 1, UofMN 1 through 3 identify the three sections of introductory statistics taught at the University of Minnesota, and

Table 1 Percentage of students who chose Agree or Strongly Agree in response to each Affect Survey item

Course	UofMN 1 (n = 24) (%)	UofMN 2 (n = 29) (%)	UofMN 3 (n = 27) (%)	NCSU (n = 22) (%)	All courses (n = 102) (%)	Fisher exact test (p)
This course helped me understand statistical information I hear or read about from the news media	83.3	82.8	88.9	90.9	86.3	0.811
Learning to use TinkerPlots™ was an important part of learning statistics	91.7	72.4	77.8	86.4	81.4	0.271
I would be comfortable using TinkerPlots™ to test for a difference between groups after completing this class	95.8	93.1	96.3	100.0	96.1	0.758
I would be comfortable using TinkerPlots™ to compute an interval estimate for a population parameter after completing this class	95.8	89.7	92.6	100.0	94.1	0.403
Learning to create models with TinkerPlots™ helped me learn to think statistically	95.8	75.9	85.2	95.5	85.0	0.112
I think I am well-prepared for future classes that require an understanding of statistics	87.5	82.8	85.2	81.8	85.0	0.948

NCSU identifies the section taught at North Carolina State University (see Sect. 4 for more detail). Fisher exact tests were conducted for each Affect Survey item to test for independence of the proportion of students choosing "agree" or "strongly agree" among the four course in which the curriculum was taught. None of the Fisher exact tests produced statistically significant results. Results show positive responses across all four classes for these items, with 80 % or more of the students giving a positive response to all of the survey items.

5.2 How well are students able to think statistically and really "cook" after taking this course?

Responses to the MOST assessment were analyzed to answer this research question. Based on the preliminary analysis of student data, and despite some problems later found in item wording and testing constraints (too little time allowed) many students appeared to be stating the need for a model, the need to simulate data and collect a statistic from the samples, and to evaluate the observed result within the resulting distribution to find a *p*-value. Although the MOST items appeared to be quite challenging for the students, roughly two-thirds of the students gave partially complete responses or fully complete responses to five of the items. Students had the most difficulty on the items related to testing a claim about a percentage from a sample of data. However, problems with the wording of these items may have led to incorrect responses. Students also showed some difficulty in reasoning about the effect of increasing sample size on the resulting size of an effect, which has led the CATALST project team to create additional activities related to this content for inclusion in the third unit of the curriculum.

5.3 How well do students understand and reason about basic statistical ideas?

Students' responses to the GOALS items were used to evaluate this research question. Figure 8 shows the percentage of correct and incorrect responses to each of the 23 GOALS items. The items have also been categorized by content.

Students performed extremely well on the GOALS items involving graphical representation of data. Students also performed very well on most of the seven items involving the use of a randomization test to simulate a null-distribution to compare observed results (Modeling/Simulation). The weakest performance in this section of GOALS was observed on items related to the interpretation of the *p*-value as the probability of an effect given the data. However, a large percentage of students did correctly identify a valid interpretation of the *p*-value using the same problem context in a separate GOALS item. On average, students answered correctly 66 % (SD = 12.3 %) of the 16 items that assessed basic statistical literacy and reasoning, with 53 % of the students correctly answering two-thirds or more of these items (i.e., 11 or more items). On average, students answered correctly 81 % (SD = 18.7 %) of the seven items that assessed their understanding of modeling and simulation, with 79 % of the students correctly answering two-thirds or more of these items (i.e., 5 or more items).

In order to further answer this research question, students' responses on the GOALS items were compared to data on a national sample of 5,362 students who completed comparable items on the CAOS test in 2009–2011 (see Fig. 9). Using the data from four items that were exactly the same on both tests (GOALS items 9, 10, 14 and 15), students in the CATALST course had higher percentages of correct answers

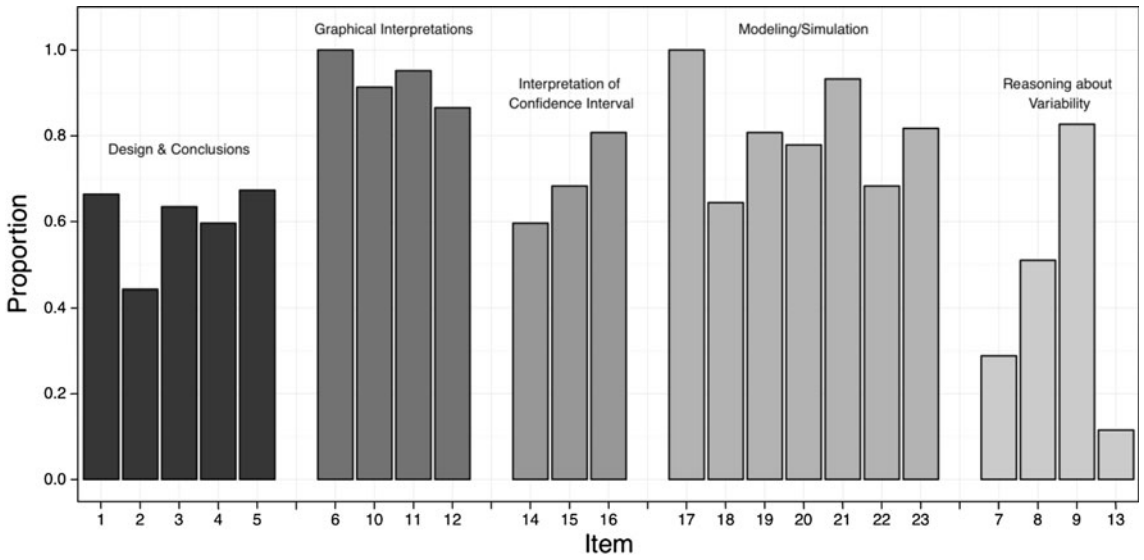

Fig. 8 Percentage of correct responses for each of the 23 GOALS items by content (n = 102)

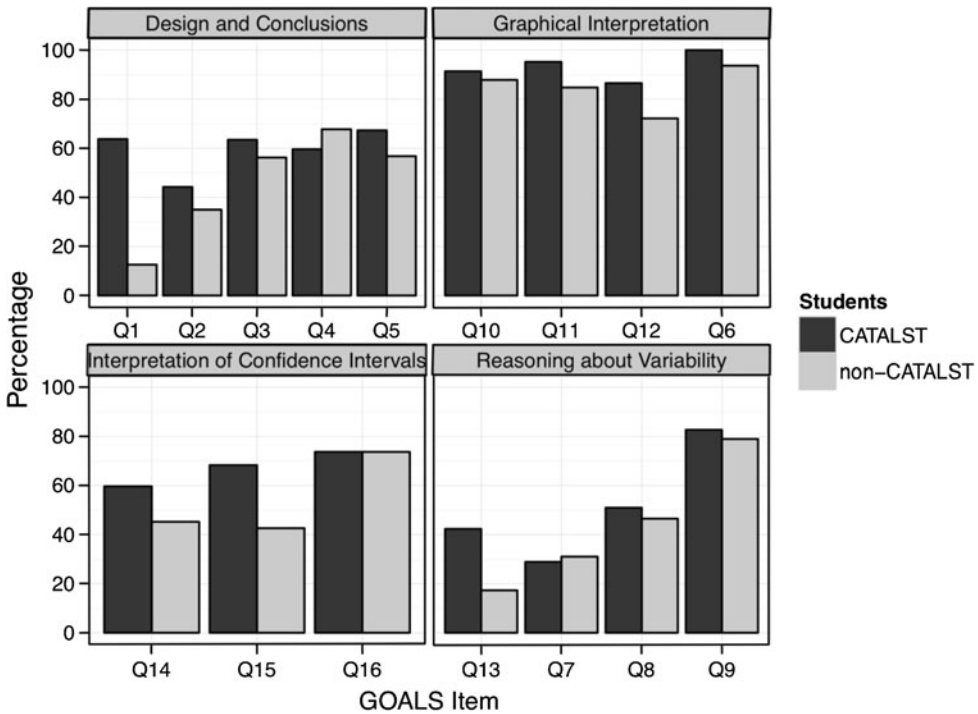

Fig. 9 Comparison of the percentage of CATALST students (n = 102) and non-CATALST students (n = 5,326) correctly answering each GOALS item (CATALST) or a similar item on the CAOS test (non-CATALST)

than on a national sample of students taking non-CATALST courses. For the remaining 12 items that had been modified from the CAOS test, students who had completed the CATALST curriculum performed better on 9 of the items. The CATALST students performed slightly worse than the national sample on two of the items. In general, there seems to be little difference between the CATALST and non-CATALST students on most of the items.

5.4 How do students view the discipline of statistics and the nature of statical problem solving?

For this final research question, students' responses to a relevant item on the Affect Survey were examined (see Table 2). The results were overwhelmingly positive with over 90 percent of students agreeing or strongly agreeing with a statement on the value of statistics.

Table 2 Percentage of students in CATALST courses who responded agree or strongly agree to a question about the value of statistics

Item	Course 1 (n = 24) (%)	Course 2 (n = 29) (%)	Course 3 (n = 27) (%)	Course 4 (n = 22) (%)	All courses (n = 102) (%)	Fisher exact test (p)
I feel that statistics offers valuable methods to analyze data to answer important research questions	91.7	96.6	96.3	100.0	95.0	0.674

6 Discussion

The data gathered from three instruments (Affect Survey, GOALS and MOST) revealed interesting aspects of student learning. Positive results were found regarding students' attitudes about the use of TinkerPlots™ software, the value of the course, and the value of the discipline of statistics. While positive attitudes may not be the most important student outcome, they are worth noting given the prevalence of negative attitudes towards statistics and often less than desirable perception regarding the discipline of statistics. It was also valuable to see that students did not dislike the use of the software and felt it helped them learn statistics. In fact, in a separate study that consisted of individual student interviews, the researcher found that TinkerPlots™ seemed to provide a structure that helped students think (see delMas et al., in review).

The data on the MOST assessment suggested that many students were beginning to think statistically. Their responses suggested that many students recognized the need to specify a model and use this to simulate data when making an inference. This was an important finding in light of the focus of the course on modeling and the importance of modeling in statistical thinking. Because of constraints on test time, students will be given more time to complete the assessment in future uses of this assessment. In light of the data collected, some of the items are being revised in order to provide students additional prompts to elicit more complete explanations and also for clarification.

The data from the GOALS test revealed that students seemed to have a good understanding of basic statistical ideas and could reason as well or better than students in other introductory courses, despite not having had explicit instruction on many of these topics. It may be that students have learned these topics in their secondary school classes or in other more applied research courses. Or, they may have developed these ideas informally as part of their work in the CATALST course.

7 Summary

A radically different approach to teaching introductory statistics was created based on the use of modeling and simulation, MEAs, and instructional design principles—all detailed in the first part of this paper. While it is possible in future research to look at these components in isolation, the curriculum was based on several foundations, all of which appeared to be important contributors to the strength of the CATALST approach. The MEAs appeared to engage students in thinking about real problems and inventing a variety of solutions for these problems. This process also seemed to motivate the students to learn the material in the unit by providing a type of prior knowledge that set the stage for the course content. The TinkerPlots™ software tool seemed to promote the development of students' statistical thinking and give them a sense of what it takes to really "cook" rather than "follow recipes". The design principles used to create the activities appeared to work well in promoting students reasoning by having them make and test conjectures and also by stimulating classroom discourse.

The data examined and reported in this paper were gathered from the initial pilot testing of the course, which took place in Spring 2011. The preliminary data gathered demonstrate that such a course can be taught, that students respond well to it, and that the outcomes students achieve are desirable. The analysis of the preliminary data collected in the teaching experiments, as well as our experience observing and studying the CATALST courses in the teaching experiment are that:

- Students can develop good statistical thinking about statistical inference, even in a first, introductory course.
- The content/pedagogy of the introductory tertiary-level course can be changed and students are not learning less or reacting in a negative way to this different course and approach.
- Software rooted in how students learn can be used at the tertiary-level, rather than software that is purely analytical.
- Students can be taught to "really cook" by using a modeling and simulation approach to statistical inference along with TinkerPlots™ software.

Since that time the curriculum has been modified and 15 more CATALST courses have been taught at 10 different institutions. Data collected using the same three assessments will again be collected and analyzed to learn more

about the impact of this content and approach on students. Furthermore, comparison data will be gathered from students enrolled in non-CATALST courses at these same institutions. Such data will help us better study the impact of this radically different course on tertiary students.

Acknowledgments The authors gratefully acknowledge the support of the National Science Foundation for the CATALST project. (Collaborative Research: The CATALST Project, Change Agents for Teaching and Learning Statistics, DUE-0814433). They also appreciate the contributions of their CATALST collaborators Beth Chance, George Cobb, John Holcomb and Allan Rossman. The advice given by Cliff Konold, Richard Lesh, Tamara Moore, and Rob Gould was extremely valuable. The work and dedication of graduate students Rebekah Isaak, Laura Le and Laura Ziegler, and of Dr. Herle McGowan at North Carolina State University, was a major contribution to this project. Lastly, the authors thank Anelise Sabbag for her copy-editing on this paper.

References

Antonioli, C., & Reveley, M. A. (2005). Randomised controlled trial of animal facilitated therapy with dolphins in the treatment of depression. *British Medical Journal, 331*(7527), 1231–1234.

Biehler, R., & Prömmel, A. (2010). *developing students' computer-supported simulation and modelling competencies by means of carefully designed working environments*. Paper presented at the ICOTS 8, Ljubljana, Slovenia.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Cobb, G. W. (2005). *The introductory statistics course: A saber tooth curriculum?* After dinner talk given at the United States Conference on Teaching Statistics.

Cobb, G. W. (2007). The introductory statistics course: a ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved September 28, 2010, http://escholarship.org/uc/item/6hb3k0nz#page-1.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 375–395). Dordrecht, The Netherlands: Kluwer Academic Publishers.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58. http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf.

delMas, R., Zieffler, A. & Garfield, J. (2012). Tertiary students' reasoning about samples and sampling variation in the context of a modeling and simulation approach to inference. *Educational Studies in Mathematics* (in review).

Diefes-Dux, H.A., Imbrie, P.K., & Moore, T.J. (2005). *First-year engineering themed seminar—A mechanism for conveying the interdisciplinary nature of engineering*. Paper presented at the 2005 American Society for Engineering Education National Conference, Portland, OR.

Doerr, H., & English, L. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education, 34*(2), 110–136.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics, 9*, 139–172.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Finzer, W. (2012). *Fathom® Dynamic Data™* (v. 2L) [computer software]. Emeryville, CA: Key Curriculum Press.

Garfield, J. & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Berlin: Springer.

Harper, S.R. & Edwards, M. T. (2011) A new recipe: No more cookbook lessons. *Mathematics Teacher 105*(3), 180–188.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning, 12*, 217–230.

Konold, C. & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education, 2*(1), Article 1.

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., et al. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning, 13*, 68–86.

Konold, C., & Miller, C. (2011). *TinkerPlots™ Version 2 [computer software]*. Emeryville, CA: Key Curriculum Press.

Lesh, R., & Doerr, H. M. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics teaching, learning, and problem solving* (pp. 3–33). Mahwah, NJ: Lawrence Erlbaum.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought—revealing activities for students and teachers. In A. E. Kelley & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 591–646). Mahwah, NJ: Lawrence Erlbaum Associates.

Maxara, C., & Biehler, R. (2006). Students' probabilistic simulation and modeling competence after a computer-intensive elementary course in statistics and probability (Electronic Version). In: *Proceedings of the 7th International Conference on Teaching Statistics (ICoTS 7)*. http://www.stat.auckland.ac.nz/~iase/publications/17/7C1_MAXA.pdf.

Maxara, C., & Biehler, R. (2007). Constructing stochastic simulations with a computer tool—students' competencies and difficulties (Electronic Version). In: *Proceedings of CERME 5*. http://www.erme.unito.it/CERME5b/WG5.pdf#page=79.

Moore, T. J., Diefes-Dux, H. A., & Imbrie, P. K. (2006). *The quality of solutions to open-ended problem solving activities and its relation to first-year student team effectiveness*. Chicago, IL: Paper presented at the American Society for Engineering Education Annual Conference.

Moore, T.J., Diefes-Dux, H.A., & Imbrie, P.K. (2007). How team effectiveness impacts the quality of solutions to open-ended problems. In: Distributed journal proceedings from the International Conference on Research in Engineering Education 2007 special issue of the *Journal of Engineering Education*.

Saldanha, L. A., & Thompson, P. W. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257–270.

Schoenfeld, A. H. (1998). Making mathematics and making pasta: From cookbook procedures to really cooking. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 299–319). Mahwah, NJ: Lawrence Erlbaum.

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 319–344). Mahwah, NJ: Erlbaum.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. Lesh (Eds.), *Handbook of research design in*

mathematics and science education (pp. 267–306). Dordrecht, The Netherlands: Kluwer.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.

Zawojewski, J.,K., Bowman, K., & Diefes-Dux, H.A. (Eds.). (2011). *Mathematical modeling* in *engineering education*: *Designing experiences* for *all students.* Rotterdam, The Netherlands: Sense Publishers.

Zieffler, A., Garfield, J., delMas, R., Isaak, R., Ziegler, L., & Le, L. (2011). *How do tertiary students reason about samples and sampling in the context of a modeling and simulation approach to informal inference?* Paper presented at the Seventh International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-7). Texel Island, The Netherlands.

Zieffler, A., Harring, J., & Long, J. (2011b). *Comparing groups: Randomization and bootstrap methods using R*. New York: Wiley.