Development and assessment of a preliminary randomization-based introductory statistics curriculum

Nathan Tintle
Jill VanderStoep
Vicki-Lynn Holmes
Brooke Quisenberry
Todd Swanson
Hope College

Journal of Statistics Education Volume 19, Number 1 (2011),
www.amstat.org/publications/jse/v19n1/tintle.pdf

Key Words: CAOS; Inference; Permutation test.

Abstract

The algebra-based introductory statistics course is the most popular undergraduate course in statistics. While there is a general consensus for the content of the curriculum, the recent Guidelines for Assessment and Instruction in Statistics Education (GAISE) have challenged the pedagogy of this course. Additionally, some arguments have been made that the curriculum should focus on a randomization approach to statistical inference instead of using asymptotic tests. We developed a preliminary version of a randomization based curriculum which we then implemented with 240 students in eight sections of introductory statistics in fall 2009. The Comprehensive Assessment of Outcomes in Statistics (CAOS) assessment test was administered to these students and showed that students learned significantly more about statistical inference using the new curriculum, with comparable learning on most other questions. The assessment results demonstrate that refining content, improving pedagogy and rethinking the consensus curriculum can significantly improve student learning. We will continue to refine both content and pedagogy resulting in improved student learning gains on CAOS items and other assessment measures.

1. Introduction

The algebra-based introductory statistics course is the most widely offered and taken undergraduate statistics course (Moore and Legler, 2003). While the course has undergone some significant changes in the last twenty years (Aliaga, Cuff, Garfield, Lock, Utts, Witmer, 2005), the statistics education community has, to a large extent, agreed upon a consensus curriculum for the course. The consensus introductory statistics curriculum is typically presented in three major units: (1) Descriptive statistics and study design (first third of course), (2) Probability and sampling distributions (second third of course), and (3) Statistical inference (final third of course) (Malone, Gabrosek, Curtiss, Race, 2010). This curriculum is implemented in a number of the most popular textbooks (e.g. Moore 2007, Agresti and Franklin 2008, Utts and Heckard 2007).

Subsequent to the adoption of the consensus curriculum, the Comprehensive Assessment of Outcomes in Statistics (CAOS; delMas, Garfield, Ooms, Chance, 2007) was developed. The CAOS test represents the first standard, comprehensive assessment instrument for the consensus algebra-based introductory statistics course. In a nationally representative sample of undergraduate students, including both public and private four-year college and university students, as well as two-year college students who completed the standard introductory statistics course, the average posttest CAOS score was 54.0%, compared to 44.9% on the pretest, where 100% represents achieving all of the content learning goals for the course (delMas et al., 2007). While this difference represented a statistically significant increase in content knowledge ($p<0.001$), the results are striking. Specifically, students left the course having added only 9% to their final score, and the 54% average final score indicates that a significant portion of items are incorrectly answered by students at the end of an introductory statistics course.

In 2005, around the same time the CAOS test was being developed, the American Statistical Association endorsed the Guidelines for Assessment and Instruction in Statistics Education (GAISE, Aliaga et al., 2005). The guidelines gave six recommendations which, at the time, were not implemented by the majority of introductory statistics courses. Specifically, these recommendations are to: (1) Emphasize statistical literacy and develop statistical thinking; (2) Use real data; (3) Stress conceptual understanding rather than mere knowledge of procedures; (4) Foster active learning in the classroom; (5) Use technology for developing conceptual understanding and analyzing data; (6) Use assessments to improve and evaluate student learning. In short, while the GAISE guidelines suggest that part of the poor performance on the CAOS test may be the pedagogical approach to statistics education, the GAISE guidelines do not propose radical changes to the curriculum itself.

During this time, another change was occurring. Statistics, once taught almost exclusively at the college level, was being integrated throughout the K-12 curriculum (Aliaga et al., 2005). By the time today's high school students get to college many of them have already seen much of the material in the first third of the traditional statistics curriculum. This is validated by the results from the nationally representative CAOS data, which identifies eight questions that more than 60% of students correctly answered prior to entering the course. Of these eight questions, six are covered in the first third of the traditional course: most showed no change in the percent of students who could answer these questions correctly by the end of the course.

The familiarity with the material that many students experience near the beginning of the traditional course quickly dissipates as students begin learning about probability and sampling distributions. This portion of the course is conceptually difficult, technically complicated and often disconnected from real data analysis and inference (Cobb, 2007).

By the final third of the course, when statistical inference is introduced, many students have lost sight of the big picture of statistics (arguably, real data analysis and inference) and end up in survival mode. When these student attitudes are combined with a typical student's end of semester busyness and stress, what's left is a shallow level of understanding of inferential statistics, arguably the crux of the course. These statements are in part justified by CAOS data which shows poor student performance on questions about statistical inference (delMas et al. 2007). In essence, instead of a course that emphasizes the logic of statistical inference, students get a course that emphasizes a series of asymptotic tests with complicated conditions (Cobb 2007). Further, the content offered is becoming increasingly outdated since the tests covered in the introductory statistics curriculum are increasingly not being used in real research practice (Cobb, 2007; Switzer and Horton 2007).

In his article, Cobb (2007) argues that the GAISE guidelines, which give general pedagogical recommendations but do not suggest major revisions to the structure or content of introductory statistics, are not enough. Instead Cobb challenges statistics educators to purposefully reconsider both the pedagogy and the content of introductory statistics. Cobb argues that we can address two significant critiques of introductory statistics, namely complexity of conditions and lack of relevance to modern statistics, simultaneously, by motivating statistical inference through a randomization approach (e.g. permutation tests) instead of asymptotic sampling distributions. Using permutation tests to learn statistical inference provides students with both a conceptually easier introduction to statistical inference and a modern, computational data analysis technique currently lacking in the first course in statistics. In a recent NSF-sponsored project (NSF CCLI-DUE-0633349), Allan Rossman and Beth Chance, along with George Cobb, John Holcomb and others (Rossman and Chance, 2008), developed a series of learning modules motivating statistical inference using randomization methods. While these modules can, and are, being used in traditional introductory statistics courses as an alternative motivation to statistical inference, to date no curriculum has been implemented that totally embraces the randomization-based approach and, thus, revamps the consensus curriculum. In this manuscript we provide an overview of our preliminary efforts to redesign the consensus curriculum taking a randomization-based approach. We then compare student learning gains from the new course to those based on the traditional consensus curriculum.

2. Methods

2.1. Curriculum Development

We developed a preliminary version of a randomization-based curriculum for introductory statistics during summer 2009. This was loosely based on modules developed by Rossman and Chance (2008). The curriculum was compiled into a textbook titled *An Active Approach to Statistical Inference* (Tintle, Swanson and VanderStoep, 2009). An annotated table of contents for the textbook is available in Appendix A. We took an active-learning approach and implemented the GAISE pedagogy while completely re-ordering, re-emphasizing and adding and subtracting content from the consensus curriculum.

2.2 Assessment

The Comprehensive Assessment of Outcomes in Statistics (CAOS) tool is used to assess conceptual understanding of statistics students in introductory statistics courses (delMas, 2007). CAOS is a 40-question, online multiple choice test that assesses students' conceptual understanding of topics taught in a traditional introductory statistics course. We administered pretest and posttest versions of CAOS on two separate occasions. The first administration of the test was to introductory statistics students in fall 2007 at Hope College. We administered CAOS to eight sections of Math 210 students (each containing 25-30 students). Out of 216 students who completed the course, we have pretest and posttest data for n=195 of the students (90% response rate). These students participated in the consensus curriculum using the Agresti and Franklin *Art and Science of Learning from Data* (2008) textbook.

During the first semester (fall 2009) of implementing the new randomization-based curriculum using the *An Active Approach to Statistical Inference* textbook we also administered the CAOS test before and after the course. Out of 229 students, valid data is available on 202 of the students, for an overall response rate of 88.2%.

When administering the CAOS test in fall 2007, students took the test in a computer lab under the supervision of the instructor at the end of the first week of class, and again in a computer lab under supervision during the last week of the class. On the pretest, students received a 100% homework grade for taking CAOS, while on the posttest they received a performance based grade (e.g. 100% HW grade if scored 70% or higher on the CAOS test, etc). In fall 2009, students took the CAOS outside of the classroom and received 100% homework grades simply for completing the tests (both pretest and posttest). For the pretest in fall 2007, and both pretest and posttest in fall 2009, students were reminded that they would get a 100% homework grade for completing the CAOS test, and that, while we wanted students to do their best, their grade would not be based on their performance on the test. Two of the five instructors teaching introductory statistics in each of the two semesters being compared were the same. Administering and reporting CAOS test results was approved by the Hope College Institutional Review Board as part of our ongoing efforts to assess curricular changes in our statistics classes.

2.3 Statistical analyses

Statistical analyses of assessment data were conducted using SPSS Statistics 17.0 (2009). ANOVA is used to compare aggregate change in CAOS scores between cohorts while matched pairs t-tests are used to test for significant learning gains within each cohort. McNemar's test on each of the 40 CAOS questions is used to investigate learning gains within each of the three cohorts: the Hope student sample (2007) using the traditional curriculum (HT), the Hope student sample (2009) using the new, randomization-based curriculum (HR) and the nationally representative sample described in delMas et al. (2007) which used the traditional curriculum (NT). Differences in item-level posttest scores between cohorts are evaluated by using a logistic regression model predicting whether a student correctly answered the question on the posttest by pretest (correct or not) and cohort (HT, NT and HR). For item-level analyses a Bonferroni correction was used to set the significance level at 0.05/40=0.00125. All tests are two-sided.

3. Developing a Randomization-Based Curriculum

3.1 Content

Motivated by our own CAOS scores using the traditional curriculum (*Tables 1-3 in paper; A1-A2 in Appendix B*), by Cobb's paper (2007) and the initial work developing learning modules for teaching introductory statistics using a randomization approach (Rossman and Chance 2008), we set out to re-design the introductory statistics course from the ground-up using a randomization approach. Specifically, we proposed the following content learning goals for the new course: The curriculum:

1. Emphasizes the core logic of statistical inference by using randomization tests. Instructors begin the course talking about statistical inference, if possible. After clearly establishing the logic of statistical inference, students make the connections between asymptotic and randomization tests.

2. Presents sampling distributions in an intuitive manner based on randomization tests that feeds understanding of the core logic of statistical inference and can be directly utilized in real data analysis.

3. Reviews topics in descriptive statistics, but does not spend explicit time discussing descriptive statistics topics that most students in the class know already.

4. Presents confidence intervals as an outcome of tests of significance, instead of the other way around.

5. Promotes a deeper and more intuitive understanding of power and its use in study design.

6. Underscores crucial differences between experimental and observational studies as related to conclusions about cause-effect relationships.

7. Utilizes descriptive and inferential statistics techniques in a large-scale research project that shows students all aspects of the scientific process (from hypothesis formulation through presentation of results), and how each step is impacted by statistical thinking.

The aforementioned curriculum goals were met by modifying the traditional course. Specifically, we re-ordered course concepts, added significant treatment of randomization tests and power, de-emphasized descriptive statistics techniques, and did not explicitly cover probability and sampling distributions (though these topics are covered implicitly in teaching about randomization tests). The resulting course can be viewed in two main parts. The first part introduces the core logic of statistical inference using randomization tests, while the second part connects randomization tests to asymptotic tests and introduces confidence intervals and power.

In the first part of the course we focus on permutation testing and the core logic of statistical inference for tests of a single proportion, comparing two proportions, comparing two means and testing correlation/regression. Additionally, basic descriptive statistical techniques are reviewed in the context of inferential data analysis and important distinctions between experimental and observational studies are introduced. The second part of the course emphasizes the connection between randomization (permutation) and traditional (asymptotic) tests (independent samples t-test, ANOVA, chi-squared test, and asymptotic correlation/regression tests), confidence intervals and power. Confidence intervals are presented as part of the core logic of statistical inference and statistical power and sample size relationships are presented intuitively using web-applets developed in-house expressly for this purpose.

The textbook website (http://math.hope.edu/aasi) has copies of sample chapters of the textbook. Additional textbook chapters and course materials are available from the authors upon request.

3.2 Pedagogy

We sought to design a new course that, at its very core, addressed the GAISE guidelines in terms of course pedagogy. With this in mind we instituted a number of significant changes in our approach to teaching statistics. Most significantly, we transitioned from a more traditional mix of lecture and laboratory exercises, to a focus on tactile, self-discovery learning experiences supported by a mix of lecture and concept review. These learning experiences are implemented in the curriculum's companion textbook as "activities" which are designed to (1) utilize real and interesting datasets that are a mix of peer-reviewed data and datasets gathered by students; (2) engage students in a variety of learning strategies including reading/comprehension, simulation (computer-based and tactile), peer and instructor led discussions and written reflections; (3) help students discover new (minor) statistical ideas on their own and reinforce core concepts introduced by the instructor; and (4) make active use of computational resources (computers) wherever possible. A number of other important pedagogical changes have also been instituted:

1. All class periods now take place in a computer lab (no graphing calculator usage). In the second half of the course, the course software transitions from Fathom to SPSS as students transition from randomization tests to traditional tests. Most students are comfortable with using two software packages in the course.

2. During much of the semester students are actively participating in large-scale group research projects reinforcing course material.

3. All review exercises are based on real studies and real data.

4. Each chapter contains a case study (an in-depth statistical analysis of real research data) which integrates a variety of concepts from the chapter.

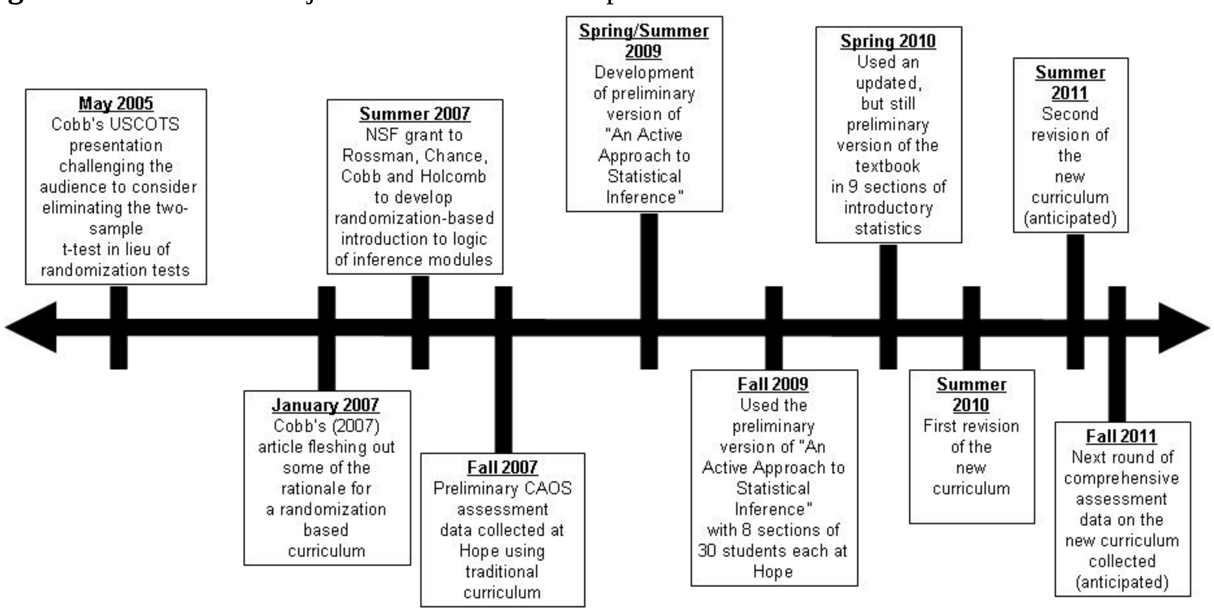
5. In each chapter there is at least one research article that students read and reflect on. In the beginning chapters these are from popular media (e.g. newspapers) but transition quickly to peer-reviewed primary literature as the course progresses. These research articles reinforce concepts from the chapter as well as ensure that students can translate their classroom knowledge to large, complex and relevant statistical issues.

6. We de-emphasize the use of symbolic notation and mathematical equations wherever possible. Instead, students write about statistical and mathematical ideas in prose which forces students to communicate about statistics in their own words, thus increasing their conceptual understanding of the topics taught.

3.3 Timeline

Using ideas presented by Cobb (2007), as well as by Rossman and Chance (2008) and teaching in a newly renovated "Statistical Teaching and Computing Laboratory" (funded by a Howard Hughes Medical Institute grant to Hope College), we pilot tested a preliminary randomization based curriculum in spring 2009 in two sections of introductory statistics. Following this pilot test we held a workshop with all Hope statistics instructors (May 2009) where discussion centered on outcomes from the pilot implementation, including discussion of the randomization approach to inference (new to some instructors). The workshop served as a time to develop a careful outline of the vision and curriculum for the course. Subsequent to the workshop, a subgroup of instructors, along with two student assistants, wrote a textbook (*An Active Approach to Statistical Inference*) and the accompanying materials (exercises, solutions, data sets, web applets) that were developed in fall 2009. Eight sections (each with 25-30 students) took the new course with the preliminary randomization-based curriculum in fall 2009, taught by five different instructors. Figure 1 gives an overview of the development of the course, and anticipated future work.

Figure 1. Timeline of major events in the development of a randomization based curriculum



4 Assessment results

4.1 Aggregate comparisons

During the first full rollout of the new curriculum in fall 2009, CAOS assessment data was collected both before and after the course. Table 1 shows the aggregate percent correct across the 40 question CAOS test from fall 2009 compared to fall 2007 at Hope and nationally.

Table 1. Aggregate comparisons of CAOS scores

	Sample size	Average percent correct		Difference in average percent correct
		Pre-test Mean (SD)	Post-test Mean (SD)	Mean (SD)
National sample (NT)	763	44.9 (-[1])	54.0 (-[1])	9.1 (12.0)
Hope sample Fall 2007 (HT)	195	48.4 (11.2)	57.2 (11.8)	8.9 (9.9)
Hope sample Fall 2009 (HR)	202	44.7 (9.3)	55.7 (11.8)	11.0 (11.3)

1. Standard deviations for the pre and posttest were not available, but the standard deviation of the difference was recorded in delMas et al. (2007)

In all cases, students average CAOS scores increased significantly ($p<0.001$; matched pairs t-tests). Pretest and posttest scores, as well as learning gains are similar between all curricula. Only weak evidence of a difference in aggregated learning gains between the three curricula exists (one-way ANOVA; $p=0.093$).

4.2 Question by question analyses

Despite similar aggregate CAOS results, there were a number of significant differences in student performance on particular questions. In the following sections we compare the three cohorts through a question by question analysis that looks at the differences between the pretest and posttest CAOS results. Comparing the two Hope cohorts (traditional curriculum (2007) vs. new curriculum (2009)) is important in identifying those areas in which the new curriculum may have made a difference in student learning. However, it is also important to compare both years of Hope's CAOS results to the nationally representative sample to ensure that differences between the new and old curriculum are not due to idiosyncrasies of Hope students, faculty or other Hope specific pedagogies.

In the following paragraphs we summarize question by question differences between the cohorts. In the first section we identify questions showing gains in student learning (posttest minus pretest scores) that are significantly higher in the Hope cohort that used the new curriculum as compared to the Hope cohort that used the traditional curriculum. In the second section we identify a single question that showed gains in student learning that is significantly lower with the Hope cohort using new curriculum compared to the Hope cohort using the traditional curriculum. The remaining sections briefly overview questions for which there were other significant differences among the cohorts or there were no significant differences among them.

Questions on which Hope students taking the new curriculum performed significantly better than Hope students taking the traditional curriculum

Table 2 shows six items that showed significant differences between cohorts such that the Hope students taking the new curriculum (HR) outperformed Hope students using the traditional curriculum (HT). In all but one of the cases, HR also significantly outperformed the national sample (NT).

Understanding the purpose of randomization in an experiment (*item 7*) was a question for which very few students knew the correct answer in any sample (NT: 8.5%, HT: 4.6%, HR: 3.5%) on the pretest. The HR cohort was the only cohort to show statistically significant learning gains on the question. The HR cohort also showed significantly greater posttest scores than the other two samples, though only slightly more than 20% of the students were able to correctly answer the question on the posttest.

In all three samples approximately half of the students entered the course able to correctly answer a question about low p-values (*item 19*). While all three samples showed statistically significant improvements in the percent of students correctly answering the question (McNemar's $p<0.001$), 96% of students using the new curriculum correctly answered the question, which was significantly more ($p<0.001$) than either the traditional curriculum nationally (68.5%) or at Hope (85.6%).

Students taking the new curriculum showed statistically significant gains (17.8%) on a question about lack of statistical significance and its interpretation (*item 23*), while students in both traditional curriculum cohorts showed no increase. On the posttest Hope students using the new curriculum outperformed the national sample and the Hope students using the traditional curriculum (85.1% vs. 64.4% and 72.7%, respectively).

When comparing Hope samples on a question about the correct interpretation of a p-value (*item 25*), students using the new curriculum had borderline significant learning gains ($p=0.002$) and performed significantly better on the posttest. Compared to the national sample the Hope sample performed better, though the difference was not statistically significant.

Another question on p-values showed even more striking results. The students taking the new curriculum performed significantly better than both of the other samples on their ability to recognize an incorrect interpretation of a p-value (*item 26*). Furthermore, the students taking the new curriculum were the only group to show statistically significant learning gains on this question.

Lastly, students taking the new curriculum performed significantly better than the other two samples on the posttest when asked to indicate how to simulate data to find the probability of an observed value (*item 37*). While there were no significant learning gains on this question using any curriculum, the learning gains using the new curriculum were borderline statistically significant ($p=0.009$).

Table 2. Items for which students in the HR cohort learned significantly more than the HT cohort

Item number on CAOS	Item Description (Topic)	Cohort[1]	% of Students Correct: Pretest	Posttest	Difference	McNemar's Test p-value	Cohort p-value[2]	aOR (95%CI)[2]
7	Understanding of the purpose of randomization in an experiment (Data Collection and Design)	NT	8.5	12.3	3.8	0.013	**0.001**	0.5 (0.3, 0.8)**
		HT	4.6	9.7	5.1	0.076		0.4 (0.2, 0.7)**
		HR	3.5	20.8	17.3	**<0.001**		1.0
19	Understanding that low p-values are desirable in research studies (Tests of Significance)	NT	49.9	68.5	18.6	**<0.001**	**<0.001**	0.1 (0.0, 0.2)***
		HT	56.9	85.6	28.7	**<0.001**		0.2 (0.1,0.5)**
		HR	56.9	96.0	39.1	**<0.001**		1.0
23	Understanding that no statistical significance does not guarantee that there is no effect (Tests of Significance)	NT	63.1	64.4	1.3	0.630	**<0.001**	0.3 (0.2, 0.5)***
		HT	66.2	72.7	6.5	0.130		0.4 (0.3, 0.8)**
		HR	65.2	85.1	19.9	**<0.001**		1.0
25	Ability to recognize a correct interpretation of a p-value (Tests of Significance)	NT	46.8	54.5	7.7	0.005	**<0.001**	0.8 (0.6, 1.1)
		HT	36.1	41.0	4.9	0.402		0.5 (0.3, 0.7)***
		HR	42.3	60.0	17.7	0.002		1.0
26	Ability to recognize an incorrect interpretation of a p-value. Specifically, probability that a treatment is not effective. (Tests of Significance)	NT	53.1	58.6	5.5	0.044	**<0.001**	0.4 (0.3, 0.5)***
		HT	59.8	68.6	8.8	0.085		0.6 (0.3, 0.9)*
		HR	58.9	79.7	20.8	**<0.001**		1.0
37	Understanding of how to simulate data to find the probability of an observed value (Probability)	NT	20.4	19.5	-0.9	0.713	**0.001**	0.5 (0.4, 0.7)***
		HT	20.0	20.0	0.0	1.000		0.5 (0.3, 0.8)**
		HR	20.0	32.2	12.2	0.009		1.0

1. NT= National sample with the Traditional curriculum, HT= Hope sample (2007) with the traditional curriculum, HR= Hope sample (2009) with the new curriculum
2. Results from a logistic regression model predicting post-test (right/wrong) by curriculum, controlling for pre-test right/wrong. Cohort p-value gives the overall p-value for the cohort term, and aOR gives the adjusted odds ratio (and corresponding 95% CI) comparing each curriculum to the new randomization based curriculum. *p<0.05, **p<0.01 and ***p<0.001.

Question for which Hope students taking the new curriculum performed worse than Hope students using the traditional curriculum

Table 3 shows the single question (*item 14*) that showed significantly poorer results with the sample of students taking the new curriculum, when compared to the other two groups. The question involved estimating standard deviations from boxplots. Both samples taking the traditional curriculum showed significant learning gains on this question, while gains for the new curriculum sample were only borderline significant. Posttest scores were significantly higher for both samples that used the traditional curriculum.

Other questions

In addition to the 7 questions described above, the remaining 33 questions can be roughly grouped into questions that showed significant differences on the posttest and those that didn't. Detailed tables, similar to Tables 2 and 3, for these two sets of questions are presented in Tables A1 and A2 in Appendix B. While there are many interesting patterns seen in these results, including seeing relative strengths and weaknesses of Hope's program versus the national sample and seeing areas that no or all curricula performed well, we do not discuss or interpret these results in detail here.

Summary of results by topic

Table 4 provides a summary of the places where differences occurred between the new and traditional curricula by topic using the topic groupings proposed by delMas et al. (2007). Four of the six questions for which the new curriculum showed more student improvement than the traditional curricula were related to tests of significance, one was related to study design and one was related to probability (specifically, simulation). The question showing poorer performance was related to descriptive statistics.

Table 3. Items for which students in the HR cohort learned significantly less than the HT cohort

Item number on CAOS	Item Description (Topic)	Cohort[1]	% of Students Correct			McNemar's Test p-value	Cohort p-value[2]	aOR (95%CI)[2]
			Pretest	Posttest	Difference			
14	Ability to correctly estimate and compare standard deviations for different histograms. (Descriptive Statistics)	NT	34.3	51.7	17.4	**<0.001**	**<0.001**	1.2 (0.9, 1.7)
		HT	44.8	70.8	26.0	**<0.001**		2.5 (1.6, 3.9)***
		HR	36.3	48.5	12.2	0.006		1.0

1. NT= National sample with the Traditional curriculum, HT= Hope sample (2007) with the traditional curriculum, HR= Hope sample (2009) with the new curriculum
2. Results from a logistic regression model predicting post-test (right/wrong) by curriculum, controlling for pre-test right/wrong. Cohort p-value gives the overall p-value for the cohort term, and aOR gives the adjusted odds ratio (and corresponding 95% CI) comparing each curriculum to the new randomization based curriculum. *p<0.05, **p<0.01 and ***p<0.001.

Table 4. Learning differences by topic[1]

Topic	Total number of items	New curriculum performed better than Hope traditional (Table 2)	New curriculum performed worse than Hope traditional (Table 3)	Other significant differences between samples (Table A1)	No significant differences between groups (Table A2)
Data Collection and Design	4	7			22, 24, 38
Descriptive Statistics	3		14		15, 18
Graphical Representations	9			6	1, 3, 4, 5, 11, 12, 13, 33
Boxplots	4			2	8,9,10
Bivariate Data	3			39	20, 21
Probability	2	37		36	
Sampling Variability	5			17	16,32,34,35
Confidence Intervals	4			28,29,30	31
Tests of Significance	6	19, 23, 25, 26			27,40

1. CAOS item numbers are in the table

5. Discussion and Conclusions

In this paper we have described an initial attempt to develop a randomization-based curriculum for the popular algebra-based introductory statistics course. Briefly, we have described how we have designed a completely new introductory statistics curriculum that focuses students' attention towards the core logic of statistical inference, treats probability and sampling distributions intuitively through the use of randomization tests, and minimizes time on descriptive statistics that students already know. Furthermore, as part of our complete redesign of the curriculum we significantly changed pedagogy to be in line with the GAISE guidelines. The development of such a curriculum and its successful implementation in eight sections of introductory statistics first and foremost provides evidence that such a curricular overhaul is possible. Based on assessment data from a preliminary version of the course, there is significant improvement in student’s knowledge of tests of significance, simulation and the purpose of randomization. While the new curriculum did show significantly less learning on a single question related to boxplots, the majority of questions did not show significant differences with the traditional curriculum.

A randomization-based curriculum addresses at least two major critiques of the traditional curriculum. First, it focuses students' attention towards the logic of inference instead of focusing their attention on asymptotic results which are disconnected from real data analysis and inference. Secondly, it gives students exposure to a modern, computationally intensive statistical technique which is rapidly growing in popularity. Furthermore, in this curriculum, we have addressed other issues in content suggested by CAOS (de-emphasizing descriptive statistics) as well as significant changes to pedagogy as suggested by GAISE (active learning approaches).

While the aggregate CAOS scores are similar between the curricula, there are significant differences in what students learned. Specifically, students better understood concepts about tests of significance, design, and simulation. These concepts are all emphasized in the new curriculum. Tests of significance are taught starting on day one of the course and emphasized throughout the curriculum, instead of only during the last 6-8 weeks of the semester, as in the traditional curriculum. The purpose of randomization in an experiment and understanding data simulation are emphasized by directly linking the data collection process to the null hypothesis simulation.

The CAOS test serves as one option for assessing student learning in an introductory statistics course. It is one of the only comprehensive content assessments currently available. However, it has a number of limitations in assessing our curriculum. First, it purports to assess the concepts in the traditional curriculum. Thus, concepts that we have added to the new curriculum (e.g. randomization tests, power) for which our students should perform much better than students in the traditional curriculum, are not directly assessed by CAOS. Additionally, since the CAOS is multiple choice (with between 2-4 response options), some questions for which students get correct 25, 33 or 50% of the time may represent nothing more than guessing and not true knowledge of the concept. Thus, in the future, we see the need for more comprehensive assessment tools for introductory statistics courses.

We did see poorer performance on a single question related to boxplots and standard deviation. In our implementation of the curriculum in fall 2009, standard deviation was not addressed until later in the course, and we presumed students understood the basics about boxplots. We have since modified the curriculum to introduce standard deviation earlier, and give a more explicit treatment to boxplots. Future assessment data is needed to assess the impact of these changes.

There are some limitations of our analysis. Briefly, there are a number of important differences between the test administration of the nationally representative sample, the Hope 2007 (traditional curriculum) sample and the Hope 2009 (new curriculum) sample. One key difference is the test administration setting. While two-thirds of the national sample took the exam in a supervised in-class setting, 100% of students in the Hope 2007 sample took the exam in-class, compared to 0% in 2009. Furthermore, a performance incentive was offered on the Hope 2007 posttest, but was not offered in 2009 (only a completion incentive); the national sample was a mix of many different incentives. These administration differences could be part of the reason why Hope students in 2007 performed better on some pretest and posttest questions compared to the new curriculum. Additionally, because students were in an uncontrolled environment in 2009, it is possible that they had more resources at their disposal when taking the exam. To address this limitation we explicitly instructed students to try their best, but that they should take the exam without using any other resources and that their course grade would be based only on completion of the test (100% for completion), not on their actual performance. Thus, there was little to no incentive for students to cheat or use additional resources to improve their performance.

Since most Hope students are from Michigan or other upper Midwest states, biases in the demographics represented and the K-12 mathematics and statistics curricula in those states limit how portable our findings about Hope students are to other populations. In our analysis we have chosen to use a Bonferonni corrected alpha value of 0.00125. While this choice limits false positive findings this conservative significance level may be hiding other questions that are impacted by the new curriculum. Further replication of the results shown here over different semesters and at other institutions is necessary.

It is very important to recognize that our curriculum changed in a number of different and important ways between the fall 2007 and fall 2009 semesters. Not only did we radically change our approach to content, but we radically changed our pedagogy. Additionally, there were different students and some different instructors the two semesters. While it appears that cohort performance differences may be a result of our new curriculum (reformed both in content and pedagogy), we cannot further attribute differences to content or pedagogy only. Two important factors are noteworthy. First, the six questions identified as significantly improved in HR (fall 2009; randomization) cohort all fall in topics that are foci of the randomization curriculum. Second, the teaching of the randomization content and active learning pedagogy are, in some ways, inextricably linked because a key advantage of the randomization approach is the ability for students to engage in hands-on and computer-based simulation. Thus, our study, and others assessing the impact of a randomization curriculum are faced with the difficulty in attributing significance to either content or pedagogy.

In conclusion, we have shown that it is indeed possible to revamp the introductory statistics curriculum to a randomization-based approach. Furthermore, assessment results show positive learning gains in a number of areas emphasized by this preliminary new curriculum. Overall, we are very encouraged by the assessment results and by the improvements in student learning from this new approach. Further curricular development will continue to refine both content and pedagogy to continue to improve student learning gains on CAOS items and other assessment measures.

Appendix A

Annotated Table of Contents for "An Active Approach to Statistical Inference"
(Tintle, Vanderstoep and Swanson 2009)

Chapter 1: Introduction to Statistical Inference: One Proportion. An introduction to statistics is given. The scientific method is discussed in how it relates to statistical inference. The basic process of conducting a test is introduced. Flipping coins and computer applets are used to model the null hypothesis in a one proportion test. The activities rely on a computer applet to simulate a model of a true null hypothesis and actual results are used to find the p-value.

Chapter 2: Comparing Two Proportions: Randomization Method. The randomization method is introduced to show how two quantities, in this case proportions, can be compared. Students are shown what explanatory and response variables are and how they are set up in a 2×2 table. Fathom is used to help determine the p-values.

Chapter 3: Comparing Two Means: Randomization Method. Tests to compare two means are done using the randomization method. Again cards are used to gain an understanding of how this method works and then Fathom is used to make this process more efficient. Type I and type II errors are introduced and the difference between an observational study and an experiment is reinforced.

Chapter 4: Correlation and Regression: Randomization Method. Scatterplots, correlation, and regression are reviewed. The randomization method is used to determine if there is a relationship between two quantitative variables. The meaning of r-squared is also introduced.

Chapter 5: Correlation and Regression: Revisited. Using inference on correlation, we transition to using traditional methods of tests of significance with the help of PASW and Fathom by showing a sampling distribution can be used to model the randomization distributions we saw in Chapter 4. Confidence intervals are introduced as a range of plausible values for a population parameter. Power is introduced and it is shown how power relates to sample size, significance level, and the population correlation.

Chapter 6: Comparing Means: Revisited. Standard deviation, normal distributions, and *t*-distributions are discussed. The independent samples *t*-test is introduced and it is shown how this traditional method is related to the randomization method. A confidence interval for the difference in means is discussed. Power of a test is discussed as it relates to this test in terms of sample size, significance level, difference in population means, and population standard deviation. The traditional analysis of variance test is shown. The meaning of the F test statistic is explored and the post-hoc Tukey test is used. Power again is looked at for this test in how it is related to sample size, significance level, maximum difference in means, and standard deviation.

Chapter 7: Comparing Proportions: Revisited. The traditional test for comparing two proportions is introduced. Power for this test is looked at as it relates to the difference in population proportions, sample size, significance level, and size of the two proportions. The chi-square test for association and a post-hoc test are discussed.

Appendix B - Supplemental Tables of Assessment Data

Table A1. Other items with significant differences between cohorts

Item number on CAOS	Item Description (Topic)	Cohort[1]	% of Students Correct: Pretest	Posttest	Difference	McNemar's Test p-value	Cohort p-value[2]	aOR (95%CI)[2]
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram) (Boxplots)	NT	45.5	56.3	10.8	**<0.001**	**<0.001**	0.5 (0.4, 0.7)***
		HT	49.0	66.7	17.7	**<0.001**		0.8 (0.5, 1.2)
		HR	53.0	72.3	19.3	**<0.001**		1.0
6	Understanding that to properly describe the distribution of a quantitative variable, a graph like a histogram is needed (Graphical Representations)	NT	15.1%	25.2%	10.10%	**<0.001**	**0.001**	2.2 (1.4, 3.6)**
		HT	9.7%	15.4%	5.70%	0.052		1.3 (0.7, 2.4)
		HR	5.9%	11.4%	5.50%	0.061		1.0
17	Understanding of expected patterns in sampling variability (Sampling Variability)	NT	42.8	50.3	7.5	**<0.001**	**0.001**	0.7 (0.5, 0.9)*
		HT	54.1	66.2	12.1	0.002		1.2 (0.8, 1.9)
		HR	40.1	57.9	17.8	**<0.001**		1.0
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits) (Confidence Intervals)	NT	48.4	43.2	-5.2	**<0.001**	**<0.001**	0.5 (0.4, 0.7)***
		HT	48.7	50.3	1.6	0.679		0.7 (0.5, 1.1)
		HR	52.0	58.7	6.7	0.762		1.0
29	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits) (Confidence Intervals)	NT	32.6	67.6	35.0	**<0.001**	**<0.001**	1.5 (1.1, 2.1)*
		HT	34.9	50.8	15.9	**0.001**		0.7 (0.5, 1.1)
		HR	35.3	58.4	23.1	**<0.001**		1.0
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits) (Confidence Intervals)	NT	31.4	44.2	12.8	**<0.001**	**<0.001**	1.6 (1.1, 2.3)**
		HT	35.2	26.3	-8.9	0.085		0.7 (0.5, 1.1)
		HR	35.1	32.8	-2.3	0.665		1.0

36	Understanding of how to calculate	NT	52.7	53.0	0.3	0.955	**<0.001**	0.4 (0.3, 0.6)***
	appropriate ratios to find conditional	HT	49.7	71.8	22.1	**<0.001**		1.0 (0.7, 1.6)
	probabilities using a table of data	HR				**<0.001**		1.0
	(Probability)		46.8	71.1	24.3			
39	Understanding of when it is not wise to	NT	17.9	24.5	6.6	**0.001**	**<0.001**	3.5 (2.0, 5.9)***
	extrapolate using a regression model	HT	9.7	11.8	2.1	0.618		1.5 (0.8, 2.9)
	(Bivariate Data)	HR	15	8.9	-6.1	0.066		1.0

1. NT= National sample with the Traditional curriculum, HT= Hope sample (2007) with the traditional curriculum, HR= Hope sample (2009) with the new curriculum
2. Results from a logistic regression model predicting post-test (right/wrong) by curriculum, controlling for pre-test right/wrong. Cohort p-value gives the overall p-value for the cohort term, and aOR gives the adjusted odds ratio (and corresponding 95% CI) comparing each curriculum to the new randomization based curriculum. *p<0.05, **p<0.01 and ***p<0.001.

Table A2. Items without significant differences between cohorts

Item number on CAOS	Item Description (Topic)	Cohort[1]	% of Students Correct: Pretest	Posttest	Difference	McNemar's test p-value	Cohort p-value[2]	aOR (95%CI)[2]
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram (Graphical Representations)	NT	71.1	73.6	2.5	0.291	0.088	0.7 (0.5, 1.0)
		HT	75.9	78.5	2.6	0.597		0.9 (0.5, 1.4)
		HR	68.3	80.2	11.9	0.012		1.0
3	Ability to visualize and match a histogram to a description (negative skewed distribution for scores on an easy quiz) (Graphical Representations)	NT	56.7	73.2	16.5	**<0.001**	0.019	0.9 (0.6, 1.3)
		HT	71.3	86.7	15.4	**<0.001**		1.7 (1.0, 3.0)
		HR	60.9	76.7	15.8	**<0.001**		1.0
4	Ability visualize and match a histogram to a description of a variable (bell-shaped distribution) (Graphical Representations)	NT	48.0	63.1	15.1	**<0.001**	0.931	1.0 (0.7, 1.4)
		HT	53.6	63.1	9.5	0.027		1.0 (0.6, 1.5)
		HR	41.3	60.9	19.6	**<0.001**		1.0
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution) (Graphical Representations)	NT	55.9	71.1	15.2	**<0.001**	0.066	1.2 (0.8, 1.7)
		HT	68.6	81.5	12.9	**<0.001**		1.8 (1.1, 3.0)*
		HR	55.9	68.3	12.4	0.004		1.0
8	Ability to determine which of two boxplots represents a larger standard deviation (Boxplots)	NT	54.7	59.2	4.5	0.068	0.004	1.6 (1.2, 2.2)**
		HT	52.8	62.6	9.8	0.025		1.9 (1.2, 2.8)**
		HR	56.7	48.0	-8.7	0.082		1.0
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles (Boxplots)	NT	23.3	26.6	3.3	0.114	0.742	1.0 (0.7, 1.5)
		HT	19.6	23.1	3.5	0.360		0.9 (0.5, 1.4)
		HR	10.0	23.4	13.4	**<0.001**		1.0
10	Understanding of the interpretation of a median in the context of boxplots (Boxplots)	NT	19.6	28.3	8.7	**<0.001**	0.197	0.8 (0.5, 1.1)
		HT	21.0	33.8	12.8	**<0.001**		1.0 (0.7, 1.6)
		HR	17.3	33.2	15.9	**<0.001**		1.0

11	Ability to compare groups by considering	NT	88.0	88.2	0.2	0.928	0.027	0.6 (0.3, 1.1)
	where most of the data are, and focusing on	HT	93.3	94.9	1.6	0.629		1.4 (0.6, 3.2)
	distributions as single entities (Graphical	HR	89.6	92.5	2.9	0.362		1.0
	Representations)							
12	Ability to compare groups by comparing	NT	85.3	85.8	0.5	0.804	0.716	1.0 (0.6, 1.6)
	differences in averages (Graphical	HT	89.2	88.7	-0.5	1.00		1.2 (0.7, 2.3)
	Representations)	HR	85.1	85.6	0.5	1.00		1.0
13	Understanding that comparing two groups	NT	61.8	73.5	11.7	**<0.001**	0.302	1.0 (0.7, 1.4)
	does not require equal sample sizes in each	HT	63.1	79	15.9	**<0.001**		1.3 (0.8, 2.2)
	group, especially if both sets of data are	HR	55.2	71.8	16.6	**<0.001**		1.0
	large (Graphical Representations)							
15	Ability to correctly estimate standard	NT	38.3	46.9	8.6	**<0.001**	0.688	1.0 (0.7, 1.3)
	deviations for different histograms.	HT	42.3	51.3	9.0	0.053		1.1 (0.7, 1.7)
	(Descriptive Statistics)	HR	41.8	48.5	6.7	0.203		1.0
16	Understanding that statistics from small	NT	22.8	31.9	9.1	**<0.001**	0.008	1.1 (0.7, 1.6)
	samples vary more than statistics from large	HT	23.1	42.1	19.0	**<0.001**		1.9 (1.2, 2.9)**
	samples (Sampling Variability)	HR	21.8	29.4	7.6	0.026		1.0
18	Understanding the meaning of variability in	NT	80.6	80.6	0.0	1.000	0.084	1.2 (0.8, 1.8)
	the context of repeated measurements, and	HT	86.7	87.7	1.0	0.856		1.9 (1.1, 3.3)*
	in a context where small variability is	HR	82.6	78.7	-3.9	0.280		1.0
	desired (Descriptive Statistics)							
20	Ability to match a scatterplot to a verbal	NT	90.5	92.5	2.0	0.159	0.644	0.7 (0.4, 1.4)
	description of a bivariate relationship	HT	95.4	92.8	-2.6	0.383		0.7 (0.3, 1.6)
	(Bivariate Data)	HR	92.1	94.5	2.4	0.541		1.0
21	Ability to correctly describe a bivariate	NT	73.6	83.7	10.1	**<0.001**	0.165	1.1 (0.7, 1.6)
	relationship shown in a scatterplot when	HT	80.5	89.7	9.2	0.010		1.7 (0.9, 3.1)
	there is an outlier (influential point)	HR				0.004		1.0
	(Bivariate Data)		71.1	82.7	11.6			
22	Understanding that correlation does not	NT	54.6	52.6	-2.0	0.404	0.011	0.6 (0.4, 0.8)**
	imply causation (Data Collection and	HT	52.1	54.4	2.3	0.640		0.7 (0.4, 1.0)
	Design)	HR	44.1	61.9	17.8	**<0.001**		1.0

24	Understanding that an experimental design	NT	58.5	59.5	1.0	0.731	0.441	0.9 (0.7, 1.3)
	with random assignment supports causal	HT	64.6	65.5	0.9	1.000		1.2 (0.8, 1.8)
	inference (Data Collection and Design)	HR	56.3	59.4	3.1	0.505		1.0
27	Ability to recognize an incorrect	NT	42.3	52.7	10.4	**<0.001**	0.128	1.3 (1.0, 1.8)
	interpretation of a p-value. Specifically, as	HT	37.1	47.7	10.6	0.027		1.1 (0.7, 1.6)
	the probability a treatment is effective.	HR				0.073		1.0
	(Tests of Significance)		35.8	44.6	8.8			
31	Ability to correctly interpret a confidence	NT	47.1	74.3	27.2	**<0.001**	0.017	1.4 (1.0, 1.9)
	interval (Confidence Intervals)	HT	46.2	80.5	34.3	**<0.001**		2.0 (1.2, 3.1)**
		HR	41.8	67.8	26.0	**<0.001**		1.0
32	Understanding of how sampling errors are	NT	16.9	17.1	0.2	0.941	0.009	1.3 (0.8, 2.1)
	used to make an informal inference about a	HT	14.4	8.2	-6.2	0.073		0.6 (0.3, 1.1)
	sample mean (Sampling Variability)	HR	17.4	13.4	-4.0	0.302		1.0
33	Understanding that a distribution with the	NT	41.5	39.7	-1.8	0.511	0.312	1.2 (0.8, 1.6)
	median larger than mean is most likely	HT	42.6	43.6	1.0	0.941		1.4 (0.9, 2.1)
	skewed to the left (Graphical	HR				0.755		1.0
	Representations)		37.6	35.8	-1.8			
34	Understanding the law of large numbers for	NT	55.3	65.2	9.9	**<0.001**	0.020	1.5 (1.1, 2.0)*
	a large sample by selecting an appropriate	HT	65.6	70.3	4.7	0.368		1.7 (1.1, 2.6)**
	sample from a population given the sample	HR	53.5	55.9	2.4	0.649		1.0
	size (Sampling Variability)							
35	Ability to select an appropriate sampling	NT	34.5	44.2	9.7	**<0.001**	0.341	1.0 (0.7, 1.4)
	distribution for a population and sample	HT	37.6	50.5	12.9	0.013		1.3 (0.9, 1.9)
	size (Sampling Variability)	HR	30.5	43.1	12.6	0.010		1.0
38	Understanding of the factors that allow a	NT	26.0	37.9	11.9	**<0.001**	0.143	1.4 (1.0, 2.0)
	sample of data to be generalized to the	HT	25.1	34.4	9.3	0.038		1.2 (0.8, 1.9)
	population (Data Collection and Design)	HR	20.8	29.5	8.7	0.033		1.0

40	Understanding of the logic of a significance test when the null hypothesis is rejected (Tests of Significance)	NT	41.9	52	10.1	**<0.001**	0.820	0.9 (0.7, 1.3)
		HT	36.4	53.3	16.9	**<0.001**		1.0 (0.7, 1.5)
		HR	40.6	53.5	12.9	0.010		1.0

1. NT= National sample with the Traditional curriculum, HT= Hope sample (2007) with the traditional curriculum, HR= Hope sample (2009) with the new curriculum
2. Results from a logistic regression model predicting post-test (right/wrong) by curriculum, controlling for pre-test right/wrong. Cohort p-value gives the overall p-value for the cohort term, and aOR gives the adjusted odds ratio (and corresponding 95% CI) comparing each curriculum to the new randomization based curriculum. *p<0.05, **p<0.01 and ***p<0.001.

Acknowledgements

We acknowledge the helpful feedback from the anonymous reviewers and editor that greatly improved this manuscript. Funding for this project has come from a Howard Hughes Medical Institute Undergraduate Science Education program grant to Hope College, the Great Lakes College Association Pathways to Learning Collegium, and the Teagle Foundation. We are extremely grateful to Joanne Stewart, current HHMI-USE director at Hope College, Dean Moses Lee and former interim HHMI-USE director Sheldon Wettack for their initial and continued support of the project. We also thank George Cobb, Allan Rossman, Beth Chance and John Holcomb for being willing to have us adapt materials from their project. Lastly we acknowledge the mathematics, biology and other departments at Hope College for their willingness to support and embrace radical changes to our statistics courses.

References

Agresti, A. and Franklin, C. (2008). Statistics: the Art and Science of Learning from Data, 1st Edition, Upper Saddle River, NJ: Pearson.

Aliaga, M., Cuff, C., Garfield, J., Lock, R., Utts, J. and Witmer, J. (2005). “Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report.” American Statistical Association. Available at: http://www.amstat.org/education/gaise/

Cobb, G. (2007). “The introductory statistics course: a Ptolemaic curriculum?” Technology Innovations in Statistics Education. 1(1).

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007). “Assessing students’ conceptual understanding after a first course in statistics” Statistics Education Research Journal 6(2):28-58.

Malone, C., Gabrosek, J., Curtiss, P., and Race M. (2010). “Resequencing topics in an introductory applied statistics course” The American Statistician. 64(1):52-58.

Moore, D. (2007). The Basic Practice of Statistics, 4th Edition, New York, NY: W.H. Freeman and Company.

Moore, T. and Legler, J. (2003). “Survey on Statistics within the Liberal Arts Colleges” Available at: www.math.grinnell.edu/~mooret/reports/LASurvey2003.pdf

PASW (SPSS) Statistics 17.0. (2009). SPSS: An IBM Company. Chicago, IL.

Rossman, A. and Chance, B. (2008). “Concepts of Statistical Inference: A randomization-based curriculum”. Available at: http://statweb.calpoly.edu/csi

Switzer S and Horton N. (2007). “What Your Doctor Should Know about Statistics (but Perhaps Doesn't).” Chance. 20(1): 17-21

Tintle, N., VanderStoep, J. and Swanson, T. (2009). An Active Approach to Statistical Inference, Preliminary Edition, Holland, MI: Hope College Publishing.

Utts, J. and Heckard R. (2007). Mind on Statistics, 3rd Edition, Belmont, CA: Duxbury.

Nathan Tintle
Department of Mathematics
27 Graves Place
Holland, MI 49423
Email: tintle@hope.edu
Phone: 616-395-7272

Jill VanderStoep
Department of Mathematics
27 Graves Place
Holland, MI 49423

Vicki-Lynn Holmes
Departments of Mathematics and Education
27 Graves Place
Holland, MI 49423

Brooke Quisenberry
Department of Psychology
35 E 12th St.
Holland, MI 49423

Todd Swanson
Department of Mathematics
27 Graves Place
Holland, MI 49423
