RETENTION OF STATISTICAL CONCEPTS IN A PRELIMINARY RANDOMIZATION-BASED INTRODUCTORY STATISTICS CURRICULUM

NATHAN TINTLE
Dordt College
ntintle@dordt.edu

KYLIE TOPLIFF
Hope College
kylie.topliff@hope.edu

JILL VANDERSTOEP
Hope College
vanderstoepj@hope.edu

VICKI-LYNN HOLMES
Hope College
holmesv@hope.edu

TODD SWANSON
Hope College
swansont@hope.edu

ABSTRACT

Previous research suggests that a randomization-based introductory statistics course may improve student learning compared to the consensus curriculum. However, it is unclear whether these gains are retained by students post-course. We compared the conceptual understanding of a cohort of students who took a randomization-based curriculum (n = 76) to a cohort of students who used the consensus curriculum (n = 79). Overall, students taking the randomization-based curriculum showed higher conceptual retention in areas emphasized in the curriculum, with no significant decrease in conceptual retention in other areas. This study provides additional support for the use of randomization-methods in teaching introductory statistics courses.

Keywords: *Statistics education research; Simulation; Permutation tests; Active learning*

1. BACKGROUND

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) have set a new standard for how to teach the first course in statistics at the college level (Aliaga, Cuff, Garfield, Lock, Utts, & Witmer, 2005). GAISE gives recommendations for teaching statistics based on the latest research in statistics education (see Hulsizer & Woolf, 2009, for a review) and a reformed view of the learning objectives of an introductory statistics course (Aliaga, et al.). Recommendations of GAISE involve the use of what they term 'active learning,' which consists of less lecturing, more projects, lab

exercises, group problem solving and discussion; use of real data; use of technology; and an approach that emphasizes conceptual understanding, statistical literacy, and statistical thinking. These active learning reforms center on improved pedagogy more than on content reform.

For more than a decade, the algebra-based introductory statistics course, Stat 101, has had a generally accepted consensus curriculum (Malone, Gabrosek, Curtiss, & Race, 2010; Scheaffer, 1997). This curriculum focuses on the normal distribution to conduct statistical inference. Although the statistics education reform movement, culminating in GAISE, has greatly improved the pedagogy of the introductory course, there has been relatively little re-thinking about the core content of the curriculum (Cobb, 2007). Cobb argues that Stat 101 should focus on the core logic of inference by presenting concepts of inference through randomization-based methods. For example, Cobb makes the argument that a permutation test to compare two independent group means is conceptually much simpler for students to understand than the two-sample *t*-test.

In order to implement Cobb's recommendation and test his hypothesis, recent projects (e.g., Rossman, Chance, Cobb, & Holcomb, 2008) have developed modules to introduce and develop deeper understanding of statistical inference through randomization tests. Pilot testing of these modules and the resulting assessment data indicate that statistical inference can be successfully learned by students using a randomization-based approach (Chance, Holcomb, Rossman, & Cobb, 2010; Holcomb, Chance, Rossman, Tietjen, & Cobb, 2010). However, as suggested by Holcomb et al., the full learning benefits of a randomization-based approach may not be attainable until a fully integrated curriculum exists.

One such curriculum is the work of Tintle, Chance, Cobb, Rossman, Roy, Swanson and VanderStoep (2011) who are implementing a full-length, randomization-based curriculum utilizing simulation and randomization tests to motivate the logic of statistical inference. Evaluation of the learning gains of the full-length randomization-based curriculum showed significant gains in student learning compared to the consensus curriculum, likely attributable to a combination of improved pedagogy and content (Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). A key remaining question is whether the learning gains observed are merely temporary, or whether the randomization approach encourages conceptual knowledge to remain with students after the course ends.

For decades, active and experiential learning has been argued to improve student retention of both concepts and procedures (e.g., Dale, 1969). Indeed, some recent efforts at course design in introductory statistics use improved retention as a motivating argument for initiating their efforts (Lockwood, Ng, & Pinto, 2007; Parr & Smith, 1998). However, there is a noticeable dearth of research on retention among introductory statistics students, especially retention by students after they complete the course.

Retention is generally regarded as important in introductory statistics, and some guidance has been offered to assess retention both during the course and after (e.g., Berenson, Utts, Kinard, Rumsey, Jones, & Gaines, 2008). A handful of studies compared alternative modes of teaching introductory statistics and assessing student retention (Brandsma, 2000; Bude, 2007; Clark, Karuat, Mathews, & Wimbish, 2007; Kvam, 2000; Lovett, Meyer, & Thille, 2008; Richardson, 2008; Stangl, Banks, House, & Reiter, 2006). A few recurring themes emerge when looking at these studies. First, although learning gains may take place in certain content areas during the course (pretest to posttest), post-course retention is generally low (e.g., Bude; Clark et al.), a finding that is in line with recent findings across college courses (Arum & Roksa, 2011). Secondly, alternative pedagogical modes of teaching the consensus curriculum and/or changes in class settings had no impact on long term student retention (learning that lasts beyond the end of the

course) (Brandsma; Bude; Lovett et al.; Stangl et al.). In general, these studies had small sample sizes so that robust conclusions are difficult.

One of the most relevant articles involving student retention in introductory statistics involved an investigation of the impact of active learning pedagogy on introductory engineering statistics (Kvam, 2000). In the article, the author compared two classes of students: one taught with active learning methods (group projects and cooperative learning-based methods), the other without, and found that active learning tended to improve student retention eight months post-course. Although larger than many studies in this area, the sample sizes were still very small (23 and 15 in the two classes, respectively). This finding also is in contrast with the findings of Brandsma (2000) who report only a temporary gain in student learning from activity-based teaching with less retention in the long-term. It is possible that the differences in these findings are related to the differences in the items being assessed. For example, in an article comparing reform calculus to traditional calculus, students from the reform calculus course had better retention of concepts, whereas students in traditional calculus retained better procedural knowledge (Garner & Garner, 2001).

In this paper we will investigate the long-term retention of statistical concepts in students taking a randomization-based course in statistics compared to those using the consensus curriculum (details on implementation are described in Section 2.1). Our focus on conceptual knowledge, in contrast to procedural knowledge, is in line with general trends in statistics education, summarized well by GAISE (Aliaga et al., 2005).

To measure conceptual knowledge of students, we used the Comprehensive Assessment of Outcomes in Statistics (CAOS), a valid and reliable tool designed to assess students' understanding of important conceptual learning objectives for a first course in statistics. The CAOS test has been determined to validly measure outcomes that expert raters agree are necessary for successful mastery of the first statistics course (delMas, Garfield, Ooms, & Chance, 2007).

Specifically, the study was designed to address the following research question: Do students who complete a randomization-based course in introductory statistics at the tertiary level have better retention of statistical literacy and reasoning when compared to students who complete a consensus curriculum?

2. METHODS

2.1 DEVELOPMENT OF THE CURRICULUM

Here we provide a brief overview of the randomization-based curriculum used in this study. The first half of the curriculum covered inferential approaches for a single proportion, comparing two proportions, comparing two means, and regression/correlation using both randomization and simulation approaches. The core logic and scope of inference were emphasized throughout the first half of the course. In the second half of the curriculum, asymptotic tests, confidence intervals, and statistical power were introduced in multiple data contexts (including comparing three or more groups). Appendix A provides a chapter by chapter description of the topics covered. In short, the full-length randomization-based curriculum was based on adding sufficient content and expository content around the modules of Rossman et al. (2008).

The consensus curriculum was implemented using Agresti and Franklin (2007) with particular content and timing shown in Appendix A. In summary, the course follows the consensus approach of starting with descriptive statistics, talking briefly about design, then discussing probability and sampling distributions, and lastly, covering confidence intervals and tests of significance for multiple types of data.

In addition to significant content changes from the consensus curriculum, the pedagogy of the curriculum was substantially changed to an active-learning approach that emphasized exploratory, self-discovery-like activities, discussion, and tactile and computer-driven simulation; all activities were motivated entirely by real and compelling research data. In contrast, the consensus curriculum was taught in a more traditional lecture style, with computer-based lab exercises one hour per week, and lectures three hours per week. Detailed comments on the pedagogy of both curricula and more detailed descriptions of the curriculum development process for the randomization-based curriculum can be found in Tintle, VanderStoep, et al. (2011). Lastly, we note that the curriculum continues to be developed (Tintle, Chance, et al., 2011); the latest materials and assessment data available at: http://math.hope.edu/isi.

2.2 PARTICIPANTS

The CAOS test was administered electronically three separate times to each of two cohorts of introductory statistics students at Hope College. One cohort consisted of fall 2007 students who took the introductory statistics course using the consensus curriculum (consensus cohort), and the second cohort consisted of fall 2009 students who took the introductory statistics course using the randomization-based curriculum (randomization cohort). All students took a full-semester (15 week) version of the course that met 4 hours per week.

In fall 2007, 216 students completed Math 210 (introductory statistics) at Hope College. These students were taught across eight sections (each with 25-30 students) taught by five different instructors. Of these 216 students, 195 students completed the CAOS test during the first week of class (September 2007), as well as during the last week of class (second week of December 2007; response rate 90%). All 195 students were recruited by email during the second to last week of April 2008 to participate in a follow-up study. Students were offered a $5 gift card to a local coffee shop to take the CAOS test a third time. Seventy nine students chose to participate (40.5% response rate).

Similarly, in fall 2009, 229 students in eight sections (taught by five different instructors, two of whom were the same as in fall 2007) completed Math 210 (introductory statistics) at Hope College, 202 of whom completed the CAOS test during the first week of class (September 2009), as well as during the last week of class (second week of December 2009; response rate 88%). All 202 students were recruited by email during the second to last week of April 2010 to participate in a follow-up study. Students were offered a $5 gift card to a local coffee shop to take the CAOS test a third time. Seventy-seven students chose to participate (38.1% response rate). One of these students was enrolled in a statistics class during the time period from January-April 2010 and so was eliminated from the analysis, leaving a final sample of seventy-six students.

2.3 ASSESSMENT

As described in detail in delMas et al. (2007), the Comprehensive Assessment of Outcomes in Statistics (CAOS) test is designed to assess students' statistical reasoning after any first course in statistics with a focus on statistical literacy, conceptual understanding, and reasoning about variability. Detailed validity and reliability information on CAOS is available in delMas et al. with additional details available on the CAOS website (https://app.gen.umn.edu/artist/caos.html).

As shown in delMas et al. (2007), the 40 items on the CAOS test can be broken into nine topic-based sets of questions: Data Collection and Design, Descriptive Statistics,

Graphical Representations, Boxplots, Bivariate Data, Probability, Sampling Variability, Confidence Intervals, and Tests of Significance.

The overall reliability (combined sample $n = 76 + 79 = 155$) of the 40-item CAOS test as measured by Cronbach's alpha was 0.58 (pretest), 0.70 (posttest) and 0.72 (4-month retention), which, although lower than found by delMas et al. (2007) (0.82), is still an acceptable level of reliability (Pedhazur & Schmelkin, 1991).

2.4 STATISTICAL ANALYSIS

Statistical analysis was performed using a variety of approaches. Aggregate comparisons of the two cohorts used paired *t*-tests to test for changes in scores over time, as well as a multiple regression model predicting each student's four month retention score by cohort (consensus/randomization), while controlling for the students pretest and posttest scores. CAOS items were also analyzed by topic. In order to assess topics showing similar patterns of student retention, each of the 40-items was placed into one of the nine topics groups as defined by delMas et al. (2007). Each student was assigned a score (ranging from 0-100%) for each topic group, based on the percent of correct responses obtained for the topic. The effect of the cohort was modeled in multiple regression models predicting the topic group score at 4-month retention by cohort, while controlling for posttest and pretest topic group scores. Additionally, each of the 40 items was considered separately in an item-by-item analysis that was conducted using logistic regression models predicting whether the answer to each question was correct using cohort, posttest, and pretest scores for that item yielding adjusted odds ratios (aORs). Lastly, a demographic analysis of the non-respondents in the 4-month retention sample was conducted to investigate potential bias in the participating four-month retention sample. Chi-squared and independent samples *t*-tests were used to compare gender proportions and pretest and posttest scores among respondents and non-respondents in each cohort. Additionally, a multiple regression model predicting posttest scores by cohort (consensus/randomization), non-respondent status (yes/no), and the interaction of cohort and non-respondent status was also fit to test for potential differences in the non-response bias between the two samples. In general, assumptions for models (equal variance, cell counts, etc.) were met. In the few cases where assumptions were questionable, an alternative model/result is provided. All analyses were run using SPSS Statistics 18.0. In all cases two-tailed tests were used. A significance level of $\alpha=0.05$ is used when considering the overall cohort differences on the 40-question CAOS test (Section 3.2) and in the non-respondents analysis (Section 3.1). For individual item and topic scale analyses (Section 3.3) *p*-values are reported. However, given the exploratory nature of these analyses and the limited sample sizes available, the *p*-values in Section 3.3 are meant to be interpreted as an objective measure of the strength of evidence, not as statistically significant when below a certain threshold (α, adjusted for multiple testing) as would be the case in a confirmatory style analysis.

3. RESULTS

3.1. ANALYSIS COHORTS

As noted earlier, approximately 40% of both cohorts participated in the four-month retention analysis. In order to understand potential preferential participation in the retention study, a non-response analysis was conducted. No major differences were found. Females participated at a higher rate than males in both cohorts (consensus: 44.2% of females participated vs. 35.4% of males; randomization: 40.7% of females participated

vs. 29.9% of males). These differences were not statistically significant (chi-squared p = 0.21 and 0.19, respectively).

In both cohorts four month retention participants tended to have higher scores on both the pretest and posttest. Specifically, in the consensus cohort, four month retention participants had an average pretest score of 50.9 compared to 46.7 among non-participants (p = 0.012), and an average posttest score of 62.0 compared to 54.0 among non-participants (p < 0.001). In the randomization cohort the average pretest score of participants was 48.0 compared to 42.8 among non-participants (p < 0.001), with posttest scores of 57.9 among participants compared to 54.4 among non-participants (p = 0.04).

Finally, a multiple regression model predicting posttest CAOS scores using cohort, four-month retention participation status (*y*/*n*), pretest score, and the interaction between cohort and four-month retention status was also run. This model yielded a p-value of 0.012 on the interaction term, indicating a significant difference in the selection bias between the two cohorts. Specifically, students who participated in the retention analysis in the consensus cohort scored, on average, 5.1 percentage points higher on the posttest than students participating in the retention analysis in the randomization cohort (95% CI: [1.1, 9.0]). We also ran a logistic regression model predicting gender by participation status and cohort, including an interaction between participation status and cohort. The interaction term was not significant (p = 0.91) indicating no evidence of different gender participation rates between the two cohorts.

These analyses further underscore our control of pretest and posttest scores in the models presented in Sections 3.2 and 3.3. Although there is no evidence of selection bias related to gender, we re-ran the models testing for cohort effects in the aggregate CAOS 4-month retention scores and each of the nine topic groups. Estimated cohort effects remained similar to those presented in the following sections in all cases.

3.2. AGGREGATE COMPARISONS OF CAOS SCORES

Table 1 compares aggregate scores on the CAOS test between students in each of the two cohorts (consensus (2007), randomization (2009)) at three different times: pretest (first week of class), posttest (last week of class), and four-month retention (last two weeks of subsequent semester). At the time of both the pretest and posttest administration, the consensus cohort had higher average aggregate scores (two sample t-test p-values of 0.10 and 0.05, respectively), whereas at four-month retention, the randomization cohort had a higher average aggregate score (p = 0.66). In both cohorts, average four-month retention scores were lower than on the posttest; however, for the consensus cohort the decline in scores from the posttest was larger (5.28 vs. 0.61). Furthermore, the change was significant for the consensus cohort (61.92 vs. 56.64, paired t-test; p < 0.001), whereas for the randomization cohort the change was not statistically significant (58.16 vs. 57.55, paired t-test; p = 0.53).

Table 1. Aggregate comparison of CAOS score retention

	n	Average % correct (SD)			Average change in % correct (SD)	
		Pretest	Posttest	4-month	Post-pre	4-month vs. post
Consensus	78	51.00	61.92	56.64	10.92	-5.28
		(12.0)	(12.3)	(14.0)	(9.5)	(10.1)
Randomization	76	48.12	58.16	57.55	10.04	-0.61
		(9.2)	(11.4)	(11.5)	(12.3)	(8.3)

In order to test whether there was a significant difference in aggregate retention between the two samples, a multiple regression model predicting retention scores for the combined sample was used. The model predicted retention scores using an indicator variable for cohort (consensus vs. randomization) and controlled for the aggregate pretest and posttest score for each individual. The cohort variable had an estimated effect of 4.26 (95% CI: [1.59, 6.93]) after controlling for pretest and posttest, indicating a significant difference (p = 0.002) in aggregate retention between the two samples, with the randomization cohort having higher retention.

3.3. TOPIC AND INDIVIDUAL ITEM COMPARISONS OF CAOS SCORES

Having determined that, overall, students demonstrated improved retention in the randomization cohort, the 40-item CAOS test was analyzed separately for each of the nine statistical topics covered (Appendix B shows which items contribute to each topic scale). Table 2 summarizes the results for each of the nine topics, whereas Appendix B gives detailed results for each of the 40 items. We briefly summarize the results for the three topics showing the largest differences in retention. Lastly, we describe individual items showing the strongest change in retention.

Data Collection and Design Table 2 illustrates that, after controlling for pretest and posttest scores, the randomization cohort averaged 10.3 percentage points (95% CI: [2.7, 17.8]) higher scores as compared to the consensus cohort on the four items related to data collection and design. The consensus cohort showed a substantial loss in knowledge about data collection and design (12.5 percentage point decline), whereas the randomization cohort exhibited a minor loss (1.97 percentage point decline). As shown in Appendix B, three of the four items (items 7, 22 and 24) showed better retention in the randomization cohort (as indicated by aORs of 1.5, 1.7 and 2.4, respectively), with the most improvement in retention coming on an item related to the impact of randomness on causal inference (item 24). The fourth item in the group (38) showed virtually no change (aOR = 0.9).

Tests of Significance There are six items in the CAOS test related to tests of significance. The average improvement in retention for items related to tests of significance was 6.1 percentage points higher in the randomization cohort compared to the consensus cohort (95% CI: [0.8%, 11.4%]), with five of the six items (items 19, 23, 26, 27 and 40) showing more retention in the randomization cohort (aORs of 8.3, 1.9, 1.8, 1.1 and 1.9, respectively). The two items showing the most improvement in retention were related to the ability to recognize that low p-values are desirable in research studies (item 19) and recognizing an incorrect interpretation of a p-value (item 26). The sixth item in the scale (item 25) showed slightly lower retention (aOR = 0.9). See Appendix B for details.

Descriptive Statistics The topic showing the strongest evidence for decreased retention was descriptive statistics, a topic consisting of three items (14, 15 and 18). As seen in Appendix 2, these three items had aORs of 0.5, 0.7 and 1.5, respectively. Items 14 and 15 both involved the standard deviation.

Item by item analyses Four of the forty items on the CAOS test yielded individual item aORs larger than 3 (items 5, 11, 19 and 20), whereas none of the forty items yielded individual item aORs less than 0.33. Items 5 and 11 both relate to distributions, item 19 relates to p-values and item 20 relates to scatterplots. Two items had aORs between 2 and

3 (items 1 and 24), covering topics in distributions and data collection and design, respectively. Four items had aORs between 0.33 and 0.5 (items 6, 14, 30 and 39). These four items covered topics in distributions, standard deviations, confidence intervals, and bivariate data analysis.

4. DISCUSSION

In this paper we have presented a comprehensive comparison of retention among introductory statistics students four months after the completion of an introductory statistics course. We have shown that students in the randomization cohort showed higher levels of retention than students in the consensus cohort, with the strongest evidence of higher levels of retention in the areas of data collection/design and tests of significance. This evidence is in line with previously reported evidence of increased student learning from the randomization-based curriculum, as compared to the consensus curriculum, in areas related to tests of significance and data collection and design (Tintle, VanderStoep, et al., 2011).

We note that topic areas showing strongest evidence of increased retention in the randomization cohort are representative of two main emphases of the curriculum: namely, the logic and scope of inference. A key advantage of the randomization-based curriculum may be that students are able to conduct formal and informal inference on data early in the curriculum. In the Tintle et al. (2009) randomization-based curriculum, students are using simulation to conduct tests of significance on day one of the course, with the formal notion of *p*-value and null and alternative hypotheses occurring within the first week. Thus, students are generating and interpreting *p*-values for the entire course, instead of only during the last few weeks of the course, as is the case in the consensus curriculum. Similarly, the randomization curriculum emphasizes the impact of study design on scope of inference: both random samples and random assignment. Scope of inference issues are also introduced early and are revisited throughout the semester. Thus, the areas showing the greatest improvements in retention are precisely the areas emphasized by the curriculum. It is precisely because the Tintle et al. approach uses simulation and randomization methods, that logic and scope of inference topics can be introduced early in the semester and revisited often.

Because changes in the randomization curriculum are found in both content and pedagogy, this study does not allow us to attribute retention differences to content changes alone. In particular, the randomization-based curriculum was implemented using active learning as defined by GAISE as less lecture, more projects, lab exercises, group problem solving and discussion, while the consensus curriculum was generally implemented using a more traditional style (3 hours of lecture and 1 hour of computer laboratory exercises per week). We note, however, that active learning and randomization-based approaches go hand-in-hand. Coin-flipping to simulate null 50/50 chance models and card-shuffling and dealing to simulate re-randomization in a permutation test are two tactile simulations that are not typically considered in the consensus curriculum. Pedagogy and content changes are inextricably linked in the randomization-curriculum, confounding the ability to draw conclusions about the impact of specific changes in the randomization-curriculum on conceptual retention.

However, that the areas showing the strongest evidence of improvement in retention are exactly the areas emphasized for the entire semester highlights the success of this content/pedagogical reform. The consensus curriculum chooses not to emphasize the logic of inference, relegating it to the last few weeks of the semester, or the impact of study design, which is often relegated to only a few class periods. The randomization-

Table 2. Comparison of CAOS topic retention

Item Description (Topic)	Cohort	Average score on Topics					Cohort *p*-value[b]	Marginal Effect[c] (95%CI)
		Pretest	Posttest	4-month retention	Change[a]	Paired *t*-test [a,b]		
Data collection and design	Randomization	31.58	43.09	41.12	-1.97	0.564	0.008	0.103 (0.027, 0.178)
	Consensus	41.02	47.44	34.94	-12.50	<0.001		1.00
Descriptive statistics	Randomization	58.67	62.28	58.77	-3.51	0.335	0.105	-0.067 (-0.149, 0.014)
	Consensus	64.50	74.36	70.09	-4.27	0.206		1.00
Graphical Representations	Randomization	60.30	69.01	69.30	0.29	0.880	0.107	0.046 (-0.010, 0.103)
	Consensus	64.96	74.79	68.52	-6.27	0.005		1.00
Boxplots	Randomization	37.67	44.41	44.74	0.33	0.919	0.625	0.020 (-0.059, 0.098)
	Consensus	37.01	53.85	47.15	-6.70	0.036		1.00
Bivariate Data	Randomization	64.04	61.84	64.04	2.20	0.321	0.672	0.013 (-0.046, 0.072)
	Consensus	64.10	67.95	64.96	-2.99	0.225		1.00
Probability	Randomization	40.79	57.24	48.68	-8.56	0.074	0.423	0.044 (-0.065, 0.153)
	Consensus	38.46	48.08	41.67	-6.41	0.191		1.00
Sampling Variability	Randomization	35.79	41.32	42.63	1.31	0.687	0.667	0.015 (-0.055, 0.086)
	Consensus	41.04	51.54	44.36	-7.18	0.024		1.00
Confidence Intervals	Randomization	42.43	55.92	53.62	-2.30	0.502	0.842	-0.007 (-0.081, 0.066)
	Consensus	40.06	53.53	53.53	0.00	1.000		1.00
Tests of Significance	Randomization	51.54	71.27	72.37	1.10	0.622	0.022	0.061 (0.008, 0.114)
	Consensus	51.51	67.31	64.31	-2.95	0.175		1.00

[a] Posttest to 4-month retention

[b] Given the explanatory nature of these analyses and the limited sample sizes available, *p*-values are meant to be interpreted as an objective measure of the strength of evidence, not as statistically significant/not statistically significant as would be the case in a confirmatory style analysis.

[c] The partial regression weight for the dichotomous "cohort" variable after controlling for pretest and posttest.

based curriculum, however, provides a vehicle for emphasizing the logic and scope of inference in such a way that retention is enhanced.

There were differences in test administration between cohorts. However, as has been stated previously (Tintle, VanderStoep, et al., 2011), these differences are likely in favor of potentially improved scores for students in the consensus cohort. We also note that there are differences in the students in the two cohorts, which we have attempted to address by way of controlling for pretest and posttest scores. Additionally, teacher and demographic differences are potential limitations of our analysis. In short, because this is an observational study there are numerous potential alternative explanations for the differences observed between the two cohorts.

Caution should also be exercised in the generalization of results, as the students represented in the cohorts investigated here represent students from a single Midwestern college and so may not represent more diverse pools of college students. However, Tintle, VanderStoep, et al. (2011) show similar results between this sample and a national sample when both groups took the consensus curriculum, suggesting that generalizing conclusions beyond Hope College may be reasonable. Furthermore, our results should not be used to estimate overall retention of introductory statistics students, as our sample is biased towards better students. However, as shown in our non-response analysis, the sample from the consensus cohort was more biased towards better students than the sample from the randomization cohort.

Lastly, we note that we have used the CAOS test to assess student retention. Our results indicate two particular areas (tests of significance and data collection/design), showing the strongest evidence of higher retention. Notably, however, these topic groups should not be interpreted as complete measures of all aspects of tests of significance or data collection and design. Although the developers of the CAOS test argue that there was no single topic agreed upon by all reviewers that is not included on the CAOS test, additional testing and additional tools are needed to more precisely conclude the extent to which students master tests of significance, data collection and design, or any other topic in introductory statistics. Furthermore, we note that the CAOS test focuses on conceptual understanding by students, not procedural knowledge. Thus, although there was little evidence of areas of lower conceptual retention by students in the randomization-based cohort, it is unknown whether this is true for procedural knowledge.

There is a dearth of research on retention of conceptual and procedural knowledge in introductory statistics courses. In particular, little is known about what levels of conceptual and procedural knowledge retention are normative, the impact of pedagogy and class size on retention, or student level factors that influence retention like attitudes and motivation. Further research is needed to better understand retention in introductory statistics students and should consider multivariate analysis with a larger sample. However, at least two major hurdles face future studies of retention. First, participation of students after they finish the class is a difficult problem. Our participation rates of 35-40% are in line with or better than previous studies (e.g., Kvam, 2000; Lovett et al., 2008). Second, once students leave the class some students may go on to take other statistics courses, or other courses that reinforce statistical concepts (e.g., Research Methods). In results not reported here in detail, we dropped 43 students (22 and 21 from the consensus and randomization cohorts, respectively) who took either a research methods course or a general education math course (with a descriptive statistics unit) from the analysis and re-analyzed the data. In general, the results were similar, but showed less statistical significance. Specifically, both topics (data collection/design and tests of significance) with the strongest evidence in Table 2 yielded low p-values (p=0.041 for data collection; p=0.088 for tests of significance), and were the most and third most significant topics, respectively, in the analysis on the reduced sample (the

second most significant topic was graphical representations). Detailed results are available from the authors.

In this analysis we have focused on student retention as the difference between posttest and four-month post-course scores on the CAOS test. We note that when comparing aggregate CAOS scores (Table 1), there is little difference between the two cohorts, though topic comparisons and individual item comparisons show some differences (detailed results not provided). However, as noted in Section 3.1, the selection bias towards students with higher CAOS scores was larger in the consensus cohort than the randomization cohort; thus, direct comparisons of 4-month posttest scores should be done with caution.

Retention among introductory statistics students is an underexplored area in the research literature. Our analysis provides one of the largest studies of retention to date and shows higher levels of four-month retention among a cohort of students completing a randomization-based introductory statistics course compared to students taking a course using the consensus curriculum. Although these results give further quantitative support for the use of randomization methods in teaching introductory statistics, additional studies including randomized experiments are needed to pinpoint teaching and content changes that directly impact concept retention.

REFERENCES

Agresti, A., & Franklin, C. A. (2007). *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., & Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education (GAISE): College report*. Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/education/gaise/]

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.

Berenson, M. L., Utts, J., Kinard, K. A., Rumsey, D. J., Jones, A., & Gaines, L. M. (2008). Assessing student retention of essential statistical ideas: Perspectives, priorities and possibilities. *The American Statistician, 62*(1), 54–61.

Brandsma, J. A. (2000). *Data collection and analysis: Examining community college students' understanding of elementary statistics through laboratory activities* (Unpublished doctoral dissertation). North Carolina State University, Raleigh, NC. Synopsis in *Statistics Education Research Journal Newsletter*, *2*(3), September, 2001.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/Newssep01.pdf]

Bude, L. M. (2007). *On the improvement of students' conceptual understanding in statistics education* (Unpublished doctoral dissertation). Universiteit Maastricht, The Netherlands.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Bude.Dissertation.pdf]

Chance, B., Holcomb, J., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_5F1_CHANCE.pdf]

Clark, J., Karuat, G., Mathews, D., & Wimbish, J. (2007). *The fundamental theorem of statistics: Classifying student understanding of basic statistical concepts*. Unpublished manuscript.
[Online: http://www1.hollins.edu/faculty/clarkjm/stat2c.pdf]

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1).
[Online: http://escholarship.org/uc/item/6hb3k0nz]

Dale, E. (1946, 1954, 1969). *Audio-visual methods in teaching*. New York: Dryden.

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.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ6%282%29_delMas.pdf]

Garner, B. E., & Garner, L. E. (2001). Retention of concepts and skills in traditional and reformed applied calculus. *Mathematics Education Research Journal, 13*(3), 165–184.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8D1_HOLCOMB.pdf]

Hulsizer, M. R., & Woolf, L. M. (2009). *A guide to teaching statistics: Innovations and best practices*. Chichester, UK: Wiley-Blackwell.

Kvam, P. H. (2000). The effect of active learning methods on student retention in engineering statistics. *The American Statistician, 54*(2), 136–140.

Lockwood, C. A., Ng, P., & Pinto, J. (2007). An interpretive business statistics course encompassing diverse teaching and learning styles. *Academy of Educational Leadership Journal,11*(1), 11–23.

Lovett, M., Meyer, O., & Thille, C. (2008). The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*.
[Online: https://oli.web.cmu.edu/openlearning/publications/71]

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1), 52–58.

Parr, W. C., & Smith, M. A. (1998). Developing case-based business statistics courses. *The American Statistician, 52*(4), 330–337.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Richardson, A. (2008, December). Retention of knowledge between statistics courses: Results of a pilot study. Paper presented at the *Third Annual Applied Statistics Education and Research Collaboration Conference*. Newcastle, Australia.
[Online: www.uow.edu.au/content/groups/public/@web/@inf/@math/documents/doc/uow074264.pdf]

Rossman, A., Chance, B., Cobb, G., & Holcomb, J. (2008). NSF/CCLI/DUE-0633349. *Concepts of statistical inference: A randomization-based curriculum.*
[Online: http://statweb.calpoly.edu/csi]

Scheaffer, R. (1997). Discussion to New pedagogy and new content: The case of statistics. *International Statistics Review, 65*(2), 156–158.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/isr/97.Moore.pdf]

Stangl, D., Banks, D., House, L., & Reiter, J. (2006). Progressive mastery teaching: Does it increase learning and retention? Yes and No. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/C315.pdf]

Tintle, N. L., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2011). *Introduction to statistical investigations*. Unpublished manuscript. [Online: http://math.hope.edu/isi]

Tintle, N. L., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization based introductory statistics curriculum. *Journal of Statistics Education, 19*(1). [Online: www.amstat.org/publications/jse/v19n1/tintle.pdf]

Tintle, N. L., VanderStoep, J., & Swanson, T. (2009). *An active introduction to statistical inference* (Preliminary edition).

NATHAN TINTLE
Mathematics, Statistics and Computer Science Department
498 4th Ave. NE
Dordt College
Sioux Center, IA 51250

APPENDIX A: SIDE-BY-SIDE COMPARISON OF THE TWO CURRICULA

Month	Consensus curriculum (Fall 2007)[a]	Randomization curriculum (Fall 2009)[b]
August/ September (4-5 weeks)	*Chapter 1: Statistics: The Art and Science of Learning from Data*. The process of learning how to investigate using data. Using samples to learn about populations. Considering the role of computers in statistics. *Chapter 2: Exploring Data with Graphs and Numerical Summaries*. Learning about the types of data and using graphical summaries to describe data. Measures of center and spread for quantitative data. Misuse of descriptive summaries. *Chapter 3: Association: Contingency, Correlation, and Regression*. Exploring associations between two categorical or two quantitative variables. Cautions in analyzing associations. *Chapter 4: Gathering Data: Experiments vs. Observational studies*. Learning about what makes a study an experiment or an observational studies, which should we choose and good and poor ways to do each. *Chapter 5. Probability in our Daily Lives*. Understanding how probability can quantify randomness. Finding probabilities and conditional probabilities, and applying the probability rules.	*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.

Month	Consensus curriculum (Fall 2007)[a]	Randomization curriculum (Fall 2009)[b]
November/ December (5-6 weeks)	*Chapter 8. Statistical Inference: Significance Tests about Hypotheses*. Steps for performing significance tests. Tests on a single proportion and mean. Errors and limitations of significance tests. *Chapter 9. Comparing Two Groups*. Testing methods for comparing two proportions and means. Analyzing dependent samples. *Chapter 10. Analyzing the Association Between Categorical Variables*. Independence and association. Testing for association with the chi-squared test. *Chapter 11. Analyzing Association between Quantitative Variables: Regression Analysis*. Modeling the relationship between two quantitative variables with regression. Strength of association with correlation. Inferences on association between two quantitative models. Residuals and model fit. *Chapter 13. Comparing Groups: Analysis of Variance Methods*. One-way ANOVA and post-hoc considerations.	*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. Learning about how power for this test 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.

[a]Chapter numbers and contents for the consensus curriculum are from Agresti & Franklin (2007). [b]Chapter numbers and contents for the randomization curriculum are from Tintle et al. (2009).

APPENDIX B: ITEM-BY-ITEM RETENTION ANALYSIS

CAOS item	Item Description (Topic)	Cohort	% of Students Correct					Cohort *p*-value[b]	aOR (95%CI)[b]
			Pretest	Posttest	4-month retention	Change[a]	Paired *t*-test[a]		
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram (Graphical representations)	Randomization Consensus	69.7 70.5	78.9 85.9	81.6 71.8	2.7 -14.1	0.567 0.015	0.053	2.4 (1.0, 5.6) 1.0
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram) (Boxplots)	Randomization Consensus	64.5 48.1	72.4 70.5	71.1 69.2	-1.3 -1.3	0.849 0.859	0.879	0.9(0.5, 1.9) 1.0
3	Ability to visualize and match a histogram to a description (negative skewed distribution for scores on an easy quiz) (Graphical representations)	Randomization Consensus	72.4 78.2	84.2 93.6	82.9 80.8	-1.3 -12.8	0.765 0.007	0.216	1.8(0.7, 4.8) 1.0
4	Ability visualize and match a histogram to a description of a variable (bell-shaped distribution) (Graphical representations)	Randomization Consensus	49.3 60.3	69.7 69.2	75.0 69.2	5.3 0.0	0.288 1.0	0.291	1.6(0.7, 3.7) 1.0
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution) (Graphical representations)	Randomization Consensus	69.7 83.3	72.4 88.5	82.9 78.2	10.5 -10.3	0.045 0.020	0.024	3.4(1.2, 9.8) 1.0
6	Understanding that to properly describe the distribution of a quantitative variable, a graph like a histogram is needed (Graphical representations)	Randomization Consensus	11.8 5.1	14.5 20.5	10.5 23.1	-4.0 2.6	0.442 0.640	0.051	0.4(0.2, 1.0) 1.0
7	Understanding of the purpose of randomization in an experiment (Data collection and design)	Randomization Consensus	1.3 7.7	18.4 14.1	17.1 11.5	-1.3 -2.6	0.810 0.483	0.428	1.5(0.6, 4.1) 1.0

8	Ability to determine which of two boxplots represents a larger standard deviation (Boxplots)	Randomization Consensus	55.3 53.8	48.7 65.4	51.3 48.7	2.6 -16.7	0.718 0.006	0.347	1.4(0.7, 2.7) 1.0
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles (Boxplots)	Randomization Consensus	16.0 17.9	23.7 33.3	23.7 32.1	0.0 -1.2	1.0 0.798	0.568	0.8(0.4, 1.8) 1.0
10	Understanding of the interpretation of a median in the context of boxplots (Boxplots)	Randomization Consensus	15.8 26.9	32.9 46.2	32.9 38.5	0.0 -6.1	1.0 0.083	0.457	1.4(0.6, 3.5) 1.0
11	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities (Graphical reps)	Randomization Consensus	90.8 94.9	93.4 98.7	92.1 87.2	-1.3 -11.5	0.658 0.002	0.050	4.4(1.0, 19.8) 1.0
12	Ability to compare groups by comparing differences in averages (Graphical representations)	Randomization Consensus	85.5 84.6	88.2 85.9	86.8 83.3	-1.4 -2.6	0.784 0.567	0.653	1.2(0.5, 3.3) 1.0
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large (Graphical representations)	Randomization Consensus	55.3 60.2	80.3 83.3	71.1 79.5	-9.2 -13.8	0.109 0.369	0.306	0.6(0.3, 1.5) 1.0
14	Ability to correctly estimate and compare standard deviations for different histograms (Descriptive statistics)	Randomization Consensus	43.3 53.2	56.6 80.8	51.3 75.6	-5.3 -5.2	0.321 0.374	0.074	0.5(0.2, 1.1) 1.0
15	Ability to correctly estimate standard deviations for different histograms (Descriptive statistics)	Randomization Consensus	44.0 50.0	47.4 51.3	43.4 55.1	-4.0 3.8	0.605 0.567	0.228	0.7(0.3, 1.3) 1.0
16	Understanding that statistics from small samples vary more than statistics from large samples (Sampling variability)	Randomization Consensus	26.3 30.8	32.9 43.6	36.8 37.2	3.9 -6.4	0.409 0.167	0.261	1.7(0.7, 4.3) 1.0

17	Understanding of expected patterns in sampling variability (Sampling variability)	Randomization Consensus	50.0 50.0	60.5 73.1	57.9 55.1	-2.6 -18.0	0.658 0.001	0.231	1.6(0.7, 3.4) 1.0
18	Understanding the meaning of variability in the context of repeated measurements, and in a context where small variability is desired (Descriptive statistics)	Randomization Consensus	89.5 89.7	82.9 91.0	81.6 79.5	-1.3 -11.5	0.784 0.019	0.350	1.5(0.6, 3.8) 1.0
19	Understanding that low *p*-values are desirable in research studies (Tests of significance)	Randomization Consensus	67.1 62.8	97.4 89.7	98.7 89.7	1.3 0.0	0.567 1.0	0.049	8.3(1.0, 68.7)[c] 1.0
20	Ability to match a scatterplot to a verbal description of a bivariate relationship (Bivariate data)	Randomization Consensus	94.7 94.9	94.7 96.2	98.7 89.7	4.0 -6.5	0.083 0.058	0.032	17.7(1.3, 244.8)[c] 1.0
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point) (Bivariate data)	Randomization Consensus	81.6 83.3	84.2 94.9	84.2 84.6	0.0 -10.3	1.00 0.020	0.494	1.4(0.5, 3.7) 1.0
22	Understanding that correlation does not imply causation (Data collection and design)	Randomization Consensus	47.4 57.7	59.2 62.8	63.2 53.8	4.0 -9.0	0.516 0.163	0.135	1.7(0.9, 3.5) 1.0
23	Understanding that no statistical significance does not guarantee that there is no effect (Tests of significance)	Randomization Consensus	75.0 69.2	84.2 76.9	86.8 74.4	2.6 -2.5	0.596 0.640	0.150	1.9(0.8, 4.7) 1.0
24	Understanding that an experimental design with random assignment supports causal inference (Data collection and design)	Randomization Consensus	55.3 67.9	60.5 65.4	56.6 41.0	-3.9 -24.4	0.605 <0.001	0.016	2.4(1.2, 4.7) 1.0
25	Ability to recognize a correct interpretation of a *p*-value (Tests of significance)	Randomization Consensus	32.9 37.6	60.5 50.0	63.2 61.5	2.7 11.5	0.686 0.028	0.719	0.9(0.4, 1.8) 1.0

26	Ability to recognize an incorrect interpretation of a *p*-value. Specifically, probability that a treatment is not effective (Tests of significance)	Randomization Consensus	63.2 75.3	81.6 79.5	78.9 66.7	-2.7 12.8	0.673 0.040	0.127	1.8(0.8, 3.8) 1.0
27	Ability to recognize an incorrect interpretation of a *p*-value. Specifically, as the probability a treatment is effective (Tests of significance)	Randomization Consensus	30.3 32.5	48.7 50.0	47.4 47.4	-1.3 -2.6	0.820 0.686	0.886	1.1(0.5, 2.1) 1.0
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits) (Confidence intervals)	Randomization Consensus	57.9 52.6	56.6 59.0	59.2 52.6	2.6 -6.4	0.708 0.373	0.371	1.3(0.7, 2.6) 1.0
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits) (Confidence intervals)	Randomization Consensus	34.2 32.1	35.5 24.4	28.9 39.7	-6.6 15.3	0.339 0.013	0.057	0.5(0.2, 1.0) 1.0
31	Ability to correctly interpret a confidence interval (CIs)	Randomization Consensus	36.8 41.0	67.1 75.6	64.5 65.4	-2.6 -10.2	0.698 0.131	0.948	1.0(0.5, 2.0) 1.0
32	Understanding of how sampling errors are used to make an informal inference about a sample mean (Sampling variability)	Randomization Consensus	18.4 16.7	11.8 6.4	11.8 11.5	0.0 5.1	1.00 0.251	0.915	0.9(0.3, 2.6) 1.0
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left (Graphical representations)	Randomization Consensus	35.5 47.4	39.5 47.4	40.8 43.6	1.3 -3.8	0.859 0.605	0.805	0.9(0.5, 1.8) 1.0
34	Understanding the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size (Sampling variability)	Randomization Consensus	51.3 66.7	56.6 75.6	60.5 64.1	3.9 -11.5	0.625 0.140	0.548	0.8(0.4, 1.6) 1.0

35	Ability to select an appropriate sampling distribution for a population and sample size (Sampling variability)	Randomization Consensus	32.9 42.9	44.7 59.0	46.1 53.8	1.4 -5.2	0.877 0.496	0.427	0.8(0.4, 1.5) 1.0
36	Understanding how to calculate appropriate ratios to find conditional probabilities using a table of data (Probability)	Randomization Consensus	52.6 51.2	72.4 67.9	64.5 50.0	-7.9 -17.9	0.276 0.012	0.081	1.8(0.9, 3.5) 1.0
37	Understanding how to simulate data to find the probability of an observed value (Probability)	Randomization Consensus	28.9 25.6	42.1 28.2	32.9 33.3	-9.2 5.1	0.196 0.374	0.558	0.8(0.4, 1.7) 1.0
38	Understanding the factors that allow a sample of data to be generalized to the population (Data collection and design)	Randomization Consensus	22.4 30.7	34.2 47.4	27.6 33.3	-6.6 -14.1	0.321 0.048	0.791	0.9(0.4, 1.9) 1.0
39	Understanding when it is not wise to extrapolate using a regression model (Bivariate data)	Randomization Consensus	15.8 14.1	6.6 12.8	9.2 20.5	2.6 7.7	0.531 0.083	0.109	0.4(0.2, 1.2) 1.0
40	Understanding the logic of a significance test when the null hypothesis is rejected (Tests of significance)	Randomization Consensus	40.8 30.7	55.3 57.7	59.2 46.2	3.9 -11.5	0.567 0.049	0.083	1.9(0.9, 3.8) 1.0

[a]Change is from posttest to retention. [b]Results from a logistic regression model predicting posttest (right/wrong) by curriculum, controlling for pretest 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. [c]Because of the high percentage of students answering these questions correctly, expected cell count requirements are not met in these models suggesting that the *p*-values may be poorly estimated. In both cases, however, Fisher's exact test comparing the 4-month retention percentages between cohorts yielded similar *p*-values (0.034 in both cases)