MEASURING LEARNING WITHIN A LARGE DESIGN RESEARCH PROJECT

Sharleen Forbes[1], John Harraway[2], Megan Drysdale[2]
[1] StatEd Consultancy, New Zealand
[2] University of Otago, Dunedin, New Zealand
Sharleen.forbes@gmail.com

***C**onceptual learning of students from universities, schools and the workplace taking part in a research project to develop new material on bootstrapping and randomisation is investigated. The aim was to develop teaching strategies using dynamic visualisation software. Before and after instruction, students sat tests involving multi-choice and True/False questions on sampling and confidence intervals. Performance is analysed in terms of increases in correct answers and changes in responses. The percentage correct in both the pre-test and post- tests varied widely. Under two thirds of students answered the same in both tests but 5-18% changed correct to incorrect answers and 13-27% incorrect to correct answers. Relevance of questions, appropriateness of multi-choice and True/False questions in assessment and levels of learning (or unlearning) acceptable to teachers are discussed. Pre- and post-tests can measure student understanding and prior skills, but multi-choice and True/False questions may not be adequate for this purpose.*

PURPOSE

The purpose of this project was to investigate the conceptual statistics learning gained by students taking part in a large design research project. The overall purpose of the design research was to develop new learning trajectories, resource materials, and dynamic visualisation software, for new material on bootstrapping and randomisation being introduced into New Zealand's senior secondary school (http://seniorsecondary.tki.org.nz/Mathematics-and-statistics), followed by implementation with students, then retrospective analysis resulting in modification of teaching materials (Pfannkuch et al (2013). While a few of the students may have been exposed to simulation software in other subjects it is likely that this type of software would be a new experience for the majority of the students. The principles which drove the design of the visualisation software were to avoid cognitive overload, direct attention to salient features, build familiarity before introducing new concepts and combine pictorial, verbal and movement elements in key actions.

DESIGN

The design research project was conducted over 2 years and went through two developmental cycles. The first was a pilot study involving five year 13 secondary school and five university students. Pre- and post-test questions were developed and altered during the pilot. As stated in Liu (2014) some questions (question 6 and the three parts of question 8 below) were modified from the Comprehensive Assessment of Outcomes in Statistics (CAOS) test developed by delmas et al (2007). The second (main) study involved 2765 students from throughout New Zealand (14 year 13 secondary school classes, seven introductory first-year classes at two universities and one group from a statistics workplace). The learning format differed according to location with schools learning in short teaching sequences outside normal classes, university students in four 50-minute lectures and tutorials and workplace students in a full-day workshop.

Data collected were pre- and post-tests from all students, pre- and post- interviews with 38 students, task-interviews with 12 students, videos of three classes implementing the learning trajectories, and teacher and lecturer reflections. There were two versions of post-test in the main study, one focussing on bootstrapping (post-test A) and the other on randomisation (post-test B), to which students were randomly allocated. Differences in responses on the two post-tests were not explored. This paper reports on the quantitative results from the pre- and post-tests only.

Confidential unique identifiers were used to match an individual student's responses between common questions in the pre- and post-tests. The questions analysed involved either multi-choice or true/false (yes/no) responses. Students answers were categorised as correct, incorrect or non-response. Increase (or decrease) in the proportion of correct answers between the

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil. iase-web.org/Conference_Proceedings.php

two tests was analysed together with the proportions of students whose responses remained unchanged (status quo), changed from incorrect to correct or changed from correct to incorrect. This provides a simple method for measuring the learning gained by students. Analysis of the free-response questions in this study is reported on in detail in Pfannkuch et al (2013).

RESULTS

After data cleansing (removal of duplicates, etc.), there were 2757 students who had completed pre-tests; 2544 from Universities; 198 from various secondary schools and 15 from the statistics workplace. However, not all the students who sat pre-tests completed a post-test.

Table 1 gives the numbers of students completing both a pre-test and a post-test, by type of student and which post-test completed.

Table 1: Number of students completing both Pre- and Post-tests.

Type of Student	University	School	Workplace	Total
Post-test A	647	66	6	719
Post-test B	658	70	7	735
Total	1305	136	13	**1454**

Student characteristics:

A slightly higher percentage of females than males (just over 50%) completed the pre-test and one of the two post-tests. This varied by location with one University and the Schools having a statistically significant larger proportion of females than males.

A total of 2534 (92%) of students answered a question on previous statistics experience. Most responding students (83%) had some previous statistics learning at the school level, with just over half of the students having studied statistics in their final year (Year 13) of secondary schooling. As expected, there was a statistically significant difference in reported experience between the university and school groups. This was still the case after the Other statistics learning group was removed (Chi-squared = 95, p value <0.001). However, there was no statistically significant difference in level of previous statistics learning between the two sets of university students.

Table 2: Previous statistics learning

Institution	Secondary schooling					
	None	Year 11	Year 12	Year 13	Other statistics experience	Total
Universities	270	214	423	1436	163	2506
Schools	8	16	70	65	15	174
Workplace	2	0	0	4	9	15
Total	280	230	493	1505	187	2534
Percent	**10%**	**9%**	**18%**	**56%**	**7%**	**100%**

One question asked the number of lectures/hours of instruction received by students as part of the project. As expected with students being randomly assigned one of the two post-tests there was no significant difference in the number of teaching sessions attended between these groups. The majority (72%) of university and workplace students received at least 7 teaching sessions, but the school students were more evenly spread between 3-4 sessions (33%), 5-6 sessions (28%) and 7 sessions (29%).

LEARNING GAINED IN STATISTICS QUESTIONS

Seven questions asked in both the pre-test and at least one of the two post-tests are analysed to assess learning, lack of learning, or un-learning resulting from the teaching. There were fewer students doing question 6 as this was only given in post-test B.

Question 3 investigated sampling variation. Emma is interested in finding the typical right foot length (in cm) of Year 8 NZ girls. She takes a random sample of 30 Year 8 NZ girls. She can find out their right foot lengths without shoes on. The box plot of the right foot lengths is plotted with median 24 cm, lower and upper quartiles 22 and 25 cm respectively, and the whiskers extending to 18 and 27 cm. Emma looks at her graph and claims that the median right foot length of Year 8 NZ girls is 24 cm. This question had an open response in addition to a **yes/no** answer but only the latter is analysed here.

Question 4 explored the impact of sample size on the width of a confidence interval. Consider taking multiple random samples of the same size from the NZ population. From each sample the median height is calculated. The variation in the sample median heights for samples of size 100 is

(A) greater than / (B) about the same as / (C) less than (Circle the correct option)

the variation in the sample median heights for samples of size 25.

Question 5 also explored the impact of sample size on the width of a confidence interval. Based on random sampling from the NZ population an interval of plausible values for the population median height can be calculated. The width of the interval of plausible values for the population median height based on a sample of size 50 is

(A) greater than / (B) about the same as / (C) less than (Circle the correct option)

the width of the interval of plausible values for the population median height based on a sample of size 200.

Question 6 investigated the reason for random assignment in an experiment on diet and blood pressure. Prior to conducting a study the researchers conjectured that those on a fish oil diet would tend to experience greater reductions in blood pressure than those on a regular oil diet. Researchers randomly assigned 14 male volunteers with high blood pressure to one of two four-week diets: a fish oil diet and a regular oil diet. Therefore the treatment is the fish oil diet while the regular oil diet is the control. Each participant's blood pressure was measured at the beginning and end of the study, and the reduction was recorded. Why was the assignment of the 14 male volunteers to one of the two groups done randomly? Which **ONE** of the following statements gives the **best** response to this question?

A. To increase the accuracy of the research results.
B. To ensure that all male participants with high blood pressure had an equal chance of being selected for the study.
C. To reduce the amount of sampling error.
D. To produce treatment groups with similar characteristics.
C. To prevent skewness in the results.

Question 8 posed three questions on interpretation of a confidence interval requiring True/False answers based on the following study:
A statistics class wants to estimate the mean number of chocolate chips in a generic brand of chocolate chip cookies. They collect a random sample of cookies, count the chips in each cookie, and construct a confidence interval for the mean number of chips per cookie (18.6 to 21.3). Circle TRUE or FALSE for each of the following statements.

A. We believe that it is a fairly safe bet that each cookie for this brand has approximately 18.6 to 21.3 chocolate chips **True / False**

B. We expect almost all of the cookies to have between 18.6 and 21.3 chocolate chips. **True / False**

C. We believe that it is a fairly safe bet that the confidence interval of 18.6 to 21.3 includes the population mean number of chocolate chips per cookie. **True / False**

Table 3 summarises the proportions of students getting correct and incorrect answers in each of the common questions.

Table 3: Numbers (and percentages) of students getting correct and incorrect answers on the pre- and post-tests

Question	Responding Students				Non-responses	Total
	Pre-test		Post-test			
	Correct	Incorrect	Correct	Incorrect		
3	420 (32%)	887 (68%)	516 (39%)	791 (61%)	147 (10.1%)	1454
4	461 (33%)	946 (67%)	608 (43%)	799 (57%)	41 (2.8%)	1448
5	425 (31%)	938 (69%)	535 (39%)	828 (61%)	87 (6.0%)	1450
6	74 (10%)	641 (90%)	172 (24%)	543 (76%)	18 (2.5%)	733
8A	355 (31%)	807 (69%)	296 (25%)	866 (75%)	290 20.0%	1452
8B	628 (52%)	575 (48%)	761 (63%)	442 (37%)	249 (17.1%)	1452
8C	906 (72%)	356 (28%)	1094 (87%)	168 (13%)	190 (13.1%)	1452

Although 83% of the students had some previous school statistics learning, overall the percentages getting correct answers on the pre-test are low. Most of the questions are about sampling or confidence intervals but Question 6, on randomisation, has a markedly lower percentage correct on the pre-test suggesting that this is new material. As shown in Table 3, of the seven questions or parts of questions that were in both Pre- and Post-tests, only one showed a statistically significant decrease (of 6 percentage points) in the proportion of students giving correct answers after the learning. This question involved interpreting a confidence interval. For the five questions with gains these ranged from 4 to 15 percentage points. The most substantial gains were in question 6 and question 8C. The first is not surprising as this question was related to the randomisation process in experimental design - one of the two key elements of the teaching. Question 8C was the only one of the three confidence interval true/false questions where the answer was True rather than False.

The adequacy of multi-choice questions as an assessment tool has been questioned (summarised in Liu, 2014) and also its fairness across different gender and ethnic groups (as summarised in Forbes, 2000). However, here the students are all sitting identical questions in identical settings in the two tests so these biases should not affect measures of change in performance. Liu (2014), analysing an almost identical data set, used Chi-squared tests on these questions to determine that there was strong evidence that student answers were not simply the result of random guessing. It also should be noted that an incorrect answer on a multi-choice question does not necessarily mean that the student doesn't understand an underlying concept, and vice versa for a correct answer. Free response questions and also data from interviews with students can shed more insight into students' conceptual understanding.

Table 4 gives the proportions of students who gave answers on both tests by whether their answer remained the same (status quo) or changed from either incorrect to correct or vice versa. Students who did not respond to the question in one or both tests have not been included.

Table 4: Numbers (and percentages) of responding students by the type of change in answers.

	Status Quo		Change		Total responses
Question	Correct	Incorrect	From Correct to Incorrect	From Incorrect to Correct	
3	211 (16%)	582 (45%)	209 (16%)	305 (23%)	1307
4	302 (21%)	640 (46%)	159 (11%)	306 (22%)	1407
5	237 (17%)	640 (47%)	188 (14%)	298 (22%)	1363
6	38 (5%)	507 (71%)	36 (5%)	134 (19%)	715
8A	144 (13%)	655 (56%)	211 (18%)	152 (13%)	1162
8B	438 (36%)	252 (21%)	190 (16%)	323 (27%)	1203
8C	799 (63%)	61 (5%)	107 (9%)	295 (23%)	1262

For each question, over a half (between 57% - 76%) of the students gave the same sort of answer in the post-test as in the pre-test. However, those giving an incorrect answer may have given a different incorrect answer than in the previous test. These students can be described as *non-learners*, and on some questions this was well over half of the students. Across the questions between 5% -18% of students who had previously given correct answers changed to incorrect ones. These students were either guessing or *unlearning*. Another possibility given the tight teaching time (constrained by overall school and university course requirements) is that students were distracted by learning the visual tool and had not been given enough time to familiarise themselves with it before learning or revising the statistical concepts. It may also be that some students were not familiar with the language being used in the questions. From a teacher perspective, motivation is provided by the *learning* group of students. That is, the 13% - 27% who changed from incorrect to correct answers.

The above quantitative analysis is enriched by the qualitative analysis of student interviews. For example, Liu (2014) reported that in questions 3 and 4 there was a lack of understanding of sampling variation with some students thinking that a larger sample would have larger variation and that in parts 8A and 8B of the confidence interval questions not liking the question wording could lead to an incorrect response. Both Stephanie Budgett and Maxine Pannfkuch have separately co-authored papers using the interview results to explore different aspects of students' conceptual understanding (for example, Pfannkuch et al, 2013).

Implications for research and practice:

There were two aspects to the design research in this study; one was the use of dynamic visualisation software to aid students to understand underlying statistical concepts and the other was the introduction of new statistics techniques that would have been unfamiliar to almost all the students.

With respect to the first aspect, Pfannkuch et al (2013) found that *'the combination of dynamic visual imagery and verbalisations seems to have the potential to facilitate students' conceptual access to processes behind experiment-to-causation inference. However, the interpretation of the tail proportion and the argumentation still eludes many students …'* (p9).

Budgett and Wild (2014) used interview data to analyse the reasoning of two students and suggested that, for these students, experience with the dynamic visual simulation tools used in the teaching appears to have consolidated abstract inferential concepts. However, they also acknowledge the importance of preceding experience with the simulation tool by hands-on activities, as did Pfannkuch et al (2013).

The analysis presented here suggests that teachers face four types of students; those who already know the material, learners, non-learners and potential un-learners who get confused about material that they previously knew. What is not known, is what levels in each of these groups are acceptable to teachers. Without the sort of pre- and post-tests used here teachers may never really know the value added by their teaching. Further potential issues for discussion include the relevance of assessment questions to the taught material and the appropriateness of multi-choice and/or True/False question in assessing students' understanding of conceptual concepts.

IMPORTANCE

Overall, these results raise the possibility that the new learning may have been gained at the expense of some 'unlearning' of concepts that had been introduced earlier. Teachers need to ensure that new teaching methods being introduced into classrooms do not introduce new language or methods that conflict with past learning, and that sufficient time is given for students to familiarise themselves with new teaching tools before new concepts are introduced. Teacher motivation is related to the learning gained by their students. Pre- and post-tests can be used to measure student's understanding as well as to determine the skills that they come into the class with. While simple multi-choice and True/False questions may not be totally adequate for this purpose, they do provide a snapshot of the whole student group, particularly where change in performance is being measured. In-depth consideration of individual student's thinking from open-ended responses and interviews that give insight into their conceptual thinking is reported in Pfannkuch et al (2013).

REFERENCES

Budgett, S., & Wild, C. (2014) Students' visual reasoning and the Randomization test. In *Proceedings of the 9th International Conference on Teaching Statistics, ICOTS 9*. Online. 6pp. Available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A1_BUDGETT.pdf

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

Forbes, S. (2000) *Measuring students' education outcomes: Sex and ethnic difference in mathematics,* (Doctoral dissertation), Curtin University of Technology, Perth, Australia.

Liu, J. (2014) Multi-choice and true/false assessments in introductory statistics: What can they tell us about student understanding? Unpublished student BSc Honours Project, University of Auckland, New Zealand.

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. (2013). *"Bootstrapping" students' understanding of statistical inference."* Summary research report for the Teaching and Learning Research Initiative, Online.18pp. Available at http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report.pdf.