Is collecting anonymous but code-identified intervention assessment data worth the effort? Reflections on a recent study in electronics

Alexander P. Mazzolini
Swinburne University of Technology, Melbourne, Australia
amazzolini@swin.edu.au

Peter J. Cadusch
Swinburne University of Technology, Melbourne, Australia
pcadusch@swin.edu.au

Abstract: *Often the effectiveness of an educational intervention for a large lecture group is assessed by testing the cohort before and after the intervention and measuring any improvement in the aggregated data from pre- to post-testing. A limitation of this method is that not all students may attend the pre-test, post-test or the lectures where the intervention is administered, diluting the significance of the results. An alternative approach is for students to use a unique but anonymous research code that allows researchers to 'tag' each individual student and hence identify those students who participate in all intervention activities and tests ('complete responders'). This paper argues that tagged data can increase the statistical significance of an intervention hypothesis when compared to untagged data even when the statistical sample is small. In a recent study that tested the efficacy of interactive lecture demonstrations (ILDs) in improving students' conceptual understanding for an advanced topic in electronics (AC resonance), the 'complete responders' formed a relatively small subgroup (N=21) of the full group (N=86) that participated in all or only some of the activities or tests ('all responders'). The learning gains for the 'complete responders' were more significant than those of 'all responders'. The reasons for the increased significance are discussed in this paper.*

Introduction

The use of Interactive Lecture Demonstrations (ILDs) is one of many Active Learning (AL) strategies (e.g., Halloun and Hestenes, 1985, McDermott, 1991, Laws, 1997, Steinert and Snell, 1999, Crouch et al., 2004) that can be used in large class environments to engage students and improve their conceptual understanding. ILDs have led to significant student conceptual learning gains in many areas of introductory science and engineering (e.g., Sokoloff & Thornton, 1997, Thornton & Sokoloff, 1998, Meltzer & Manivannan, 2002, Zimrot & Ashkenazi, 2007, Mazzolini et al., 2011). Usually, measured learning gains are derived from the aggregated performance of the class in a diagnostic test delivered before (pre-test) and after (post-test) the ILD intervention. In large classes, results are often averaged over all students who attend the particular pre- or post-test. This paper discusses whether the learning gains due to a particular ILD teaching intervention as determined by pre- and post-testing of 'complete responders' (i.e. students who participated in the pre-test, post-test and all intervention activities) were more informative and meaningful than those determined from the aggregated data from 'all responders' (i.e. students who participated in all or only some of the intervention activities or assessment tests). 'Complete responders' in this study were identified uniquely and anonymously by individual research codes that each student used to 'tag' their own work (pre-test, post-test or ILD activity sheets).

In this paper, the authors reflect on the results obtained from a study into the effectiveness of an ILD intervention that aimed to improve students' conceptual understanding for an advanced topic (AC circuits and resonance) in an introductory electronics program at Swinburne University of Technology (Melbourne, Australia). In the study, students agreed to contribute their responses for research analysis by using their unique research code in their paper-based responses to the ILD activities and assessment tests. 86 students contributed to the pre-test, 67 contributed to the first set of ILD activities, 46 contributed to the second set of ILD activities and 38 contributed to the post-test. Of these students, 21 'complete responders' were identified as having participated in *all* activities and *all* tests, and 8 'post-test only responders' could be identified as having contributed only in the post-test and not in any other ILD activities or pre-test. The main reason for the reflections described in this paper centres around whether the lower number of participants that completed *all* activities (N=21) could yield more informative results than the aggregated data from the larger number of participants that contributed either to the pre-test (N=86) or to the post-test (N=38). In the latter case (i.e., N=86 or N=38), the authors chose not to identify whether students had contributed to all, some or none of the ILD activities (as would be the case with untagged data).

Methodology

The methodology used in this research study had been fully described in Mazzolini, Daniel and Edwards (2011) but a brief summary is given here. The ILDs were developed to improve students' understanding of key concepts around AC circuits and resonance in an introductory electronics unit, since academics teaching into this unit for several years had observed that many students had difficulty with this topic. The ILD intervention was delivered in a blended-learning mode. With this approach, the entire AC topic was taught via 8 hours of traditional lectures that were followed by an additional 2 hours of 'consolidation' ILD activities. The ILDs targeted concepts associated with a simple AC series RLC circuit:- (a) the amplitude and phase nature of AC voltages and current, (b) how voltages and current can vary with frequency in a resonant circuit, and (c) the phase relationship between the voltage and current associated with individual circuit elements (R,L or C). The ILD activities were facilitated using a Predict, Observe, Discuss, Synthesise (PODS) learning cycle as described in Lakhadar et al. (2006). The PODS cycle was designed to engage students in constructing their own understanding of concepts (Nachtigall, 1990) from careful observation of the RLC series circuit, which was constructed at the front of the class. Observations of the circuit and the outputs of various measuring instruments were displayed to the entire class in real time via a data projector. Students worked collaboratively in groups of 2 or 3 to discuss their predictions and observations, and any differences between the two.

A diagnostic test was developed to test the efficacy of the ILDs in improving students' conceptual understanding of various key ideas in AC circuits and resonance. The test consisted of 7 multiple-choice questions. Question 1 was designed to test a specific misconception in electronics (i.e., current is used up as it flows around a circuit); this misconception was discussed many times throughout the DC and AC sections of this electronics unit but it was not specifically covered in the AC consolidation ILDs. The other 6 questions tested students' understanding of specific AC concepts that were covered in the ILDs. The diagnostic test (pre-test) was administered to the class after the 8 hours of traditional lecture instruction but before the commencement of the ILD activities. The same diagnostic test (post-test) was then administered to the class about one week after the additional 2 hours of ILD activities.

Results and Discussions

Table 1 shows the results for the percentage of students in each subgroup who chose the correct answer for each question (Q1 to Q7) in the diagnostic test, which was administered either as the pre- or post-test. <All> shows the overall percentage averaged over all 7 questions. These results for the pre- and post-tests for the 'complete responders' (i.e., students who completed *all* ILD activities and tests) and for 'all responders' (i.e., all students who completed at least one test, but who did not necessarily complete all ILD activities or tests) are also shown in Figure 1 for ease of comparison.

To test a *null hypothesis* that the ILDs had no effect, the results for 'complete responders' (N=21) were statistically analysed. Experiment-wise *p*-values for the null hypothesis test were determined for each question and for the overall average for all questions. A *p*-value of 0.05 indicates a 5% probability that the difference observed occurred by chance alone, so *p*-values below this are considered significant. A statistical analysis of the data for 'complete responders' using the Simes/Benjamini-Hochberg test (Benjamini & Hochberg, 1995, Shaffer, 1995), as described in Mazzolini, Daniel and Edwards (2011), indicated the following:

- Questions 2, 3, 5 and 6 showed a significant improvement between pre- and post-tests, with experiment-wise *p*-values of 0.03, 0.02, 0.05 and 0.01 respectively.
- Questions 4 and 7 were indicative of an improvement, with experiment-wise *p*-values of 0.10 and 0.11 respectively.
- Question 1 did not show any significant improvement, with an experiment-wise *p*-value of 0.94. This was not surprising as this question tested understanding of a general concept that was not specifically addressed in the ILD intervention.
- Overall, for the average of all 7 questions, there was a highly significant improvement with an experiment-wise *p*-value of <0.01.

Table 1 Percentage of students in each subgroup who chose the correct answer for each question, and the average for all questions.

Subgroups of data set	Q1	Q2	Q3	Q4	Q5	Q6	Q7	<All>
Complete responders (Pre-test) N=21	90.5	14.3	47.6	9.5	23.8	28.6	19.0	33.3
Complete responders (Post-test) N=21	85.7	42.9	76.2	28.6	47.6	61.9	38.1	54.4
All responders (Pre-test) N=86	89.5	17.4	54.7	20.9	22.1	36.0	26.7	38.2
All responders (Post-test) N=38	89.5	39.5	76.3	28.9	42.1	52.6	31.6	51.5
Post-test Only Responders N=8	100	12.5	50.0	25.0	50.0	25.0	12.5	39.3

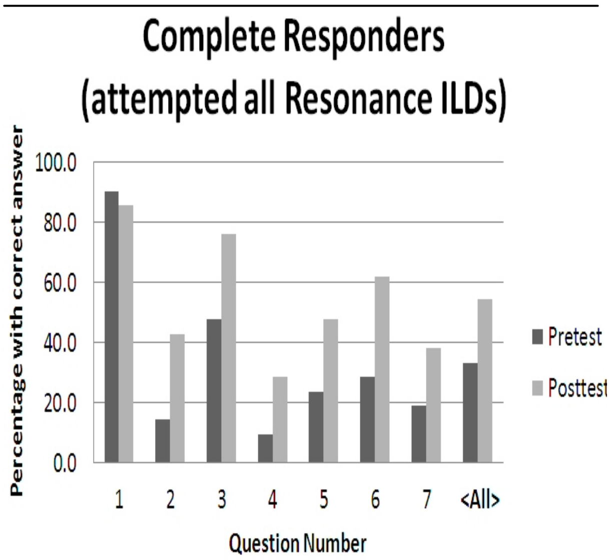

Figure 1a. Pre- and post-test results for 'complete responders' that completed all tests and all ILD activities

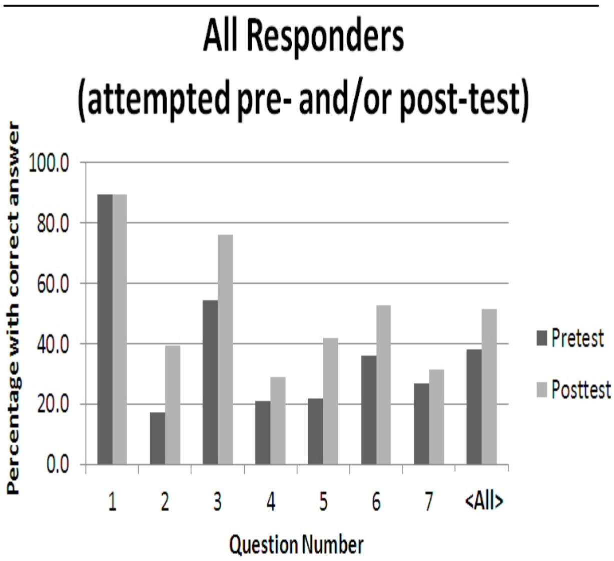

Figure 1b. Pre and post-test results for 'all responders' that attempted either the pre-test and/or the post-test, but did not necessarily complete all ILD activities

In general the pre- and post-test data from the 'complete responders' (see Table 1 and Figure 1a) indicated that the blended-learning approach, which included the consolidation ILD intervention, did result in significant student learning gains when compared to traditional lectures alone, even though the statistical sample (N=21) was quite small. The learning gains for 'all responders' (see Table 1 and Figure 1b) are less conclusive even though the sample sizes are relatively large compared to the 'complete responders'.

This analysis was possible because students were anonymously identified via their research codes, but often researchers conducting similar studies do not 'tag' students and only rely on aggregated data from students who attend the pre-test and those who attend the post-test. This sort of data does not differentiate whether the students attended both tests, or attended all, some or none of the intervention activities. As can be seen from Figure 1b, the aggregated untagged data indicated smaller improvements in each question (except Q1) and a smaller improvement in the overall average for all the questions combined. The research question investigated in this current study related to whether the smaller sample of the tagged 'complete responders' was a more informative and a more statistically significant indication of student learning gains than the aggregated data from the untagged full group ('all responders'), which was a much larger statistical sample (i.e., N=86 for the pre-test and N=38 for the post-test).

Dilution effect of students who have not completed all ILD activities

The first and probably most important observation from this study was that the inclusion of post-test results from students who had *not* participated in all the ILD intervention activities could significantly bias an estimate of the strength of the ILD intervention. In general, the effect of including the results for N_1 students who did not participate in any intervention activity in a post-test group of N_2 students, leads to an estimate of the size of the intervention effect, Δp_{biased}, which is smaller than the true effect, Δp_{true}, by a factor of $(N_2-N_1)/N_2$. In the current study, 8 students could be identified as completing the post-test and no ILD activities or pre-test. Hence, with $N_2=38$ and $N_1=8$, the factor is 0.79, leading to approximately a 20% underestimate of the size of the effect of the ILD intervention. In this particular study, the underestimate was probably higher as there were 9 additional students who only participated in some ILD intervention sessions. With tagged data, the students who had not participated in all ILD activities could be removed from the data set, thus removing this bias.

The dilution of the final data set also leads to an associated reduction in the statistical significance of the observed changes in test scores, which, depending on the size of the data set, could lead to an erroneous acceptance of the null hypothesis at the desired level of significance. This is even more important when attempting to identify the effects of the ILDs on a question by question basis.

The effect of self-selection

A second phenomenon that may be missed using untagged data in this type of study is that there appeared to be an element of self-selection in the makeup of the 'complete responders' group. Specifically, the performance of this group in the pre-test appeared to be weaker than the remainder of the overall pre-intervention group (p-value of 0.12 based on an unpaired permutation test). In addition, the 'complete responders' seemed to perform as well as, or better than, the remainder of the overall post-test group. This result suggested that students who appeared to actively engage with the ILDs were the weaker students who were struggling with concepts taught in the traditional lectures. This self-selection effect probably contributed to the nearly four-fold increase in the statistical significance of the improvement in the 'complete responder' group (p-value = 0.00013) compared with that of 'all responders' which formed the full set of available pre- and post-test data (p-value = 0.00049), notwithstanding the significant reduction in the size of the data set. Again, these p-value figures are based on an unpaired permutation test; if a paired test is used for the 'complete responder' group, the significance is even greater (p-value = 0.00002).

The identification of an inadvertent control group

Ethical restrictions prevented the use of a control group in this study, but the tagged data allowed the authors to identify the existence of a small, but not inconsiderable, subgroup of 'post-test only responders' (i.e., students who participated in the post-test but not the pre-test or any of the ILD activities). This subgroup (N=8) at least allowed the possibility of comparing their post-test results with those of the 'complete responders'. In this study, 'complete responders' usually performed considerably better in the post-test than the ‘post-test only responders’. In addition, the aggregated post-test results for the ‘post-test only responders’ were not statistically distinguishable from the pre-test results for ‘all responders’ (p-value = 0.49 using an unpaired permutation test), whereas the difference between the post-test results of the ‘post-test only responders’ and ‘all responders’ was significant (p-value = 0.06). Unfortunately the ‘post-test only responders’ group did not participate in the pre-test and so this result cannot help in untangling any effect due to familiarity with the test, although it is expected that this should be only a very small effect. Certainly, the methodology of using the *same* diagnostic test in pre- and post-assessment of a group that is exposed to a particular intervention is well established in education research (e.g., Hake, 1998). What can be said, however, is that the normal study in the intervening time between pre- and post-tests had no discernable effect in the group which did not participate in the ILD intervention, whereas the ILDs had significantly improved the post-test scores of those who participated in them. This weakens any counter assertion that the student learning gains could be attributed to normal study in the time between the two tests.

Conclusion

This paper has argued that asking students to use an anonymous but identifiable research code when recording data for educational intervention activities and assessment tests is well worth the effort. The use of student research codes allowed researchers in this study to identify ‘complete responders’ (i.e., students who had participated in all intervention activities and all tests). This study has shown that data from ‘complete responders’ provided more focused information and increased the significance of pre- and post-test results, when compared with the results from untagged aggregated data from ‘all responders’ (i.e., students who had participated in the pre-test and/or the post-test but not necessarily in all of the intervention activities). This outcome was not expected as the number of ‘complete responders’ was small (N=21) compared to the number of ‘all responders’ who participated in the pre-test (N=86) or post-test (N=38). The tagged results were more informative than the untagged results as they eliminated the dilution effect of students who had not fully participated in the intervention activities, and highlighted a self-selected group of weaker students who actively engaged with the intervention. The research codes also allowed researchers to identify an inadvertent control group (‘post-test only responders’), which could be used to help eliminate any non-intervention effects that might have occurred in the time between the pre- and post-test.

The small number of ‘complete responders’ was unexpected in this study. Analysis of the results indicated that some students did not fill in their research codes, or did not fill them in clearly enough for positive identification. The researchers also suspect that some students could not remember their original research codes for the entire study. It was also observed that the number of students in class reduced as the study progressed. This reduction could be due to students suffering from ‘survey fatigue’ as there were many paper-based worksheets and tests that students had to complete in class, and the distribution and collection of these papers required considerable class time. To try to address these issues and hence increase the number of ‘complete responders’ for future studies, the researchers in this study are trying to develop a new research code that is easier to remember but still preserves anonymity, and redesigning the tests and ILD worksheets to be compatible with audience polling devices (clickers). The use of clickers in the next iteration of the resonance ILD study will make data collection simpler, quicker and more reliable.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Ro. Stat. Soc., Series B (Methodological*), 57(1), 289-300.

Crouch, C. H., Fagen, A. P., Callan, J. P. & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment? *Am. J. Phys., 72(6)*, 835-838.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys., 66(1)*, 64-74.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students, *Am. J. Phys., 53(11)*, 1043-1055.

Lakhadar, Z. B., Culaba, I.B., Lakshminarayanan, V., Maquiling, J.T., Mazzolini, A., & Sokoloff, D.R. (2006). Introduction. In D.R. Sokoloff, (Ed.), *Active Learning in Optics and Photonics Training Manual* (pp. 1-14), Paris: UNESCO.

Laws, P. W. (1997). Millikan Lecture 1996: Promoting active learning based on physics education research in introductory physics courses, *Am. J. Phys., 65(1),* 14-21.

Mazzolini, A.P., Daniel, S. & Edwards, T. (2011). Using Interactive Lecture Demonstrations to Improve Conceptual Understanding of Resonance in an Electronics Course, *in preparation.*

Mazzolini, A., Edwards, T., Rachinger, W., Nopparatjamjomras, S. & Shepherd, O. (2011). The use of interactive lecture demonstrations to improve students' understanding of operational amplifiers in a tertiary introductory electronics course, *Lat. Am. J. Phys. Educ., 5(1)*, 147-153.

McDermott, L. C. (1991). What we teach and what is learned- Closing the gap, *Am. J. Phys*., 59(4), 301-315.

Meltzer, D. E. & Manivannan, K. (2002). Transforming the lecture-hall environment: the fully interactive physics lecture, *Am. J. Phys*., *70(6)*, 639-654.

Nachtigall, D.K. (1990). What is wrong with physics teachers' education? *Eur. J. Phys., 11(1),* 1-14.

Shaffer, P. J. (1995). Multiple Hypothesis Testing, *Annu. Review of Psych., 46*, 561-584.

Sokoloff, D. R. & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment, *The Phys. Teacher, 35(6)*, 340-347.

Steinert, Y. & Snell, L. S. (1999). Interactive lecturing: strategies for increasing participation in large group presentations, *Med. Teacher, 21(1),* 37-42

Thornton, R. K. & Sokoloff, D. R. (1998). Assessing Student Learning of Newton's Laws: the Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, *Am. J. Phys*., *66(4)*, 338-352.

Zimrot, R. & Ashkenazi, G. (2007). Interactive lecture demonstrations: a tool for exploring and enhancing conceptual change, *Chem. Educ. Res. Pract., 8(2),* 197-211.

Acknowledgements

We would like to acknowledge the support and valuable advice received from our colleagues in the Engineering and Science Education Research (ESER) group within the Faculty of Engineering and Industrial Sciences (FEIS) at Swinburne University of Technology. We would also like to acknowledge financial support for this project from Swinburne's LTPF and FEIS.

Copyright statement