ELECTRONIC JOURNAL FOR RESEARCH
IN SCIENCE & MATHEMATICS EDUCATION
VOL. 25, NO. 1, 94-105

ICRSME
International Consortium for Research in
Science & Mathematics Education

How Much is Lost? Measuring Long-Term Learning Using Multiple Choice Tests

Julio Benegas
Universidad Nacional de San Luis

Julio Sirur Flores
Universidad Nacional de San Luis

ABSTRACT

This work proposes a new approach for measuring long-term conceptual knowledge based on the after-instruction evolution of students´ answers to a research-based, multiple-choice, single-response test. The method allows for a quantitative determination of the fraction of students that, after instruction, attain long-lasting and temporary learnings, as well as those that did not learn. It also provides a plausible value of the experimental error. The method has been applied to analyze data obtained from a group comparison quasi-experimental design, in which two intact, equivalent high school classes have been subjected to two different instructional approaches. Conceptual knowledge of the subject, simple resistive electric circuits, was measured through the administration of the multiple-choice test DIRECT at three different times: before and immediately after instruction and one year later. Results indicate that the fraction of students achieving long-term learning is about four times larger in the group that followed active-learning activities, compared with the class that followed traditional instruction; drastically decreasing the no-learning group. The proposed method is relatively simple to implement and to interpret, providing more in-depth information, with higher accuracy and detail than the usual pre- and post-instruction data analysis. Some suggestions for complementary studies and to improve instruction are also given.

Keywords: MCSR tests, long-term learning, conceptual knowledge, electric circuits, tutorials

Introduction

While long-term learning is a central objective of instruction, it is well known to teachers and researchers that students lose some knowledge with time (Bernhard, 2001; Pollock, 2009). Therefore, regular post-instruction evaluation, which includes a certain (usually unknown) fraction of temporary learning, is not an accurate measurement of long-term, post-instruction knowledge. Although this fraction of labile knowledge is often useful to students for passing course examinations, especially in traditional instruction, after a certain time it disappears, becoming no longer available for future use, including to support further learnings.

Long-term learning studies are not abundant in literature, in part because, in most education systems, it is difficult to have the same student samples available for further examinations a long time after the experimental courses finish. Among the few available, Francis et al. (1998) and Bernhard (2001) show that college-level students achieve better long-term results if their instruction is based on research-based curricula, as compared with those students following traditional, lecture-based instruction. Similar results were achieved by Kohlmyer et al. (2009) and Pollock (2009) on regular college courses and by Benegas and Sirur Flores (2014), working with high school students of a very different education system. Persano Adorno et al. (2018) also report on long-term learning, but in a different type of experiments, based on supplementary, post-instruction active-learning activities. These experiences, taken as representative of studies run in different education systems and school levels, not only point out the beneficial effects of active-learning instruction, as compared with teacher-centered pedagogies, but also to the difficulties of running this type of longitudinal studies. Although the above experiments are based upon the application of the same test at two different times after instruction, they are based on the time changes of the average class performance and not on the evolution of individual student's answers to every test item, as proposed in the present work.

For several reasons, accurate determination of long-term learning is a relevant issue, in particular for assessing the effectiveness of instruction. For instance, in most Latin American countries, long-term scientific knowledge does not seem to be the usual outcome of high school instruction. According to the results of international PISA evaluations (OECD, 2019), the conceptual knowledge of regional middle school students is extremely low, with participating Latin American countries at the bottom of the world-wide performance scale. An Ibero-American study (Benegas et al., 2009; 2010) that complemented the PISA measurements, showed that just about 7% of more than 3,000 first-year science and engineering university students, attending seven universities in five different countries, have a sound conceptual knowledge of Coulomb´s law. With similar disappointing results obtained in all other tested topics, including free-fall motion, Newton´s laws, and simple dc electric circuits, all basic subjects included in the standard high school physics curricula of all participating countries. Since all these students had obtained passing grades in their high school general science and physics courses, but at the beginning of their university studies (In science and engineering!) their conceptual knowledge was so low, the immediate question regards how solid was the knowledge acquired in the corresponding high school instruction.

Towards this educational problem, this work proposes a new approach for measuring long-time conceptual learning based on the after-instruction evolution of students´ answers to a research-based, multiple-choice, single-response (RB-MCSR) test. The method follows the work of Lasry et al. (2014) who proposed the use of RB-MCSR tests to measure gains and losses of conceptual understanding by analyzing, for every test item, the options selected by each student before and after instruction.

Following a similar procedure, we propose that appropriate categorization of student´s answers to two after-instruction administrations of the same RB-MCSR test should provide an accurate measurement of long-term and temporary learnings. Therefore, this work has the following research objectives:

1. To present a method to measure long-term and temporary learnings based on the after-instruction evolution of students´ answers to individual items of RB-MCSR tests.

2. To apply this method to a group comparison classroom experiment to compare the long-term learning outputs of two different instructional approaches.

Conceptual Knowledge and Research-Based, Multiple-Choice Tests

This work is based upon the assumption that research-based, multiple-choice, single-response tests are not only a representative measure of conceptual knowledge but also a sound way to follow the evolution of the main learning difficulties and alternative models held by a given student group. The most representative RB-MCSR tests in physics and other STEM disciplines (https://www.physport.org/assessments/), have been constructed with questions that probe different aspects of a given subject. It is important to note that, for each question, the distractors (the wrong options) correspond to the most popular alternative models and learning difficulties on the tested subject. These distractors, which have been revealed by extensive qualitative and quantitative educational research on university and high school students of different school systems (see, for instance, Hestenes et al., 1992; Engelhardt & Beichner, 2004), are applied to close to everyday situations, appealing to students´ previous experiences even if they have not yet been exposed by instruction to the corresponding scientific concepts. In that regard, Bao & Redish (2006) in their model analysis, recalled that educational research has shown that alternative conceptions of a particular topic seemed to be limited to a few popular models and that different contexts -including students' mental state- could activate different, even contradictory conceptions (di Sessa, 1993; Vosniadou, 1994). Therefore, an individual with a solid scientific framework (Newtonian, for instance) should ideally answer all items in a consistently correct manner, but others -especially uninstructed participants- could choose different wrong answers, even shifting from one distractor to another without a solid reason or being particularly aware of the contradiction. In this framework, alternative models, which derive their resilience from their association with underlying presuppositions in students´ previous knowledge, should not be considered as deeply held specific theories. Consequently, students may change their local, situational models, moving from one distractor to another influenced by the context, without the need to be internally consistent. Considering furthermore, that RB-MCSR tests are relatively easy to apply, analyze and compare local results with those of other applications, it is clear that the use of RB-MCSR tests provides both practical applicability and sound pedagogical bases to the present approach.

Methods

The Classroom Experiments

To test the suitability of the proposed method and as an example of the type of data to be analyzed, we propose to study the after-instruction dynamics of high school students´ answers to an RB-MCSR test. To that end, a quasi-experimental group comparison study was designed, with pre- and post-instruction evaluation. The subject, simple resistive electric circuits, was taught to two 11th grade high school classes of a state-run mixed-gender school, attended by students coming from low to middle-class families. CTRL and EXP groups have N_{TRD}= 31 (15 females) and N_{EXP} = 30 (14 females) students, respectively, a rather common condition of local high schools. Students were assigned to each class following institutional rules, two years before the experiment. For this experiment one of the classes (called TRD heretofore) was randomly assigned to the traditional, teacher-centered instruction offered in previous years. The other class (EXP) followed an experimental instruction that used the instructional activities of the active-learning methodology Tutorials for Introductory Physics (Tutorials) (McDermott & Shaffer, 1998)._The evidence-based learning effectivity of Tutorials (Redish & Steinberg, 1999) determined its selection as the experimental teaching approach. Its learning cycle: *elicit* students´ previous ideas, *confront* them with the outcome of the Tutorials Worksheets and *resolve* the differences, is implemented through three complementary activities: Tutorial Pre-test, Tutorial Worksheet, and Tutorial Homework. Students in the EXP class, following this sequence, worked through two Tutorials didactic units: "A model for circuits Part 1: Current and resistance" and "A model for circuits Part 2: Potential difference". Pre-test and Homeworks are individual activities carried out outside the classroom, while the Tutorials Worksheets were worked out by small collaborative groups of 3-4 members in the regular classroom settings. To that end, students in each small group moved their desks so that they could face one another, building up in this way small working tables for circuit elements and paperwork. The traditional instruction consisted of demonstration-supported lectures and problem-solving sessions. The latter consisted of exemplary problem-solving demonstrated by the teacher, followed by students´ problem-solving individual practice. Homeworks consisted mainly of problem-solving activities. In both teaching approaches, Homeworks contributed to students´ grades. Both courses were taught by the same experienced teacher, who had previously participated in a Tutorials workshop.

Conceptual knowledge of the subject matter was measured through the application of the RB-MCSR test Determining and Interpreting Resistive Electric Circuits Concepts Test (hereafter DIRECT) (Engelhardt & Beichner, 2004). For this experiment this measuring instrument was applied after instruction at two different times: just at the end of instruction (Post I) and one year later (Post II). The time between Post I and Post II was determined by the availability of the students´ samples, with Post II given about one year after instruction, in the last month of these students´ high school studies. Therefore, "long-term learning" in this study case should be interpreted as the knowledge retained one year after instruction. Pre- and Post-instruction performances are used to calculate the normalized gain g, defined as g= (Post-Pre)/(100-Pre) (Hake, 1998). For the present case, we can define a "short-term" normalized gain g_I, using Post I to calculate g, and a long-term normalized gain, g_{II}, determined using Post II to calculate g.

Although in all test applications the full test (29 items) was given to students, for the present application only the 19 items (listed in Table 1) directly related to the taught subject were analyzed, excluding, for instance, those items related with energetic and microscopic aspects of electric circuits.

Equivalency of these institutionally formed groups was determined by their similar gender and socio-economics conditions, as well as their common previous experience in science and math courses. Equivalency in the subject matter was determined by the pre-instruction application (Pre for shorthand) of the test DIRECT. Average (and standard deviation) pre-instruction performances were 20(10)% for the CTRL group and 12(7)% for the EXP group, i.e., very close or lower than the random performance, pointing to the very low initial students´ knowledge about this subject. Even though an independent sample t- test found some statistical evidence of differences of pre-instruction knowledge between the two groups (t= 3.785, df= 59, p< 0.001), for the present experiments they are considered equivalent groups since their very low pre-instruction performances indicate a practically null initial knowledge of electric circuits in both courses.

Determining Long-Time Learning

The distinction between temporary, short-term, and stable, long-term learnings is a central issue in education. Soderstrom & Bjork (2015), for instance, discusses temporary and long-term learning in terms of *Performance* and L*earning*. In their framework, *Learning* refers to relatively permanent changes in knowledge or behavior, a primary goal of instruction. *Performance*, on the other hand, refers to temporary fluctuations in student´s knowledge as measured or observed during (or shortly after) instruction.

This work proposes that proper categorization and analysis of all possible (Post I, Post II) answer pairs, obtained from two post-instruction applications of the same RB-MCSR test, should provide a quantitative measurement of long-time and temporary learnings. The basic idea is to assign a plausible learning path to every possible correct/incorrect combination of (Post I, Post II) answer pairs. It is postulated that students acquiring stable, long-term learning, should systematically select, after instruction, the correct option, i.e., the appropriate scientific model. Temporary, short-term learning, on the other hand, corresponds with those students that, choosing the correct option immediately after instruction, return to an incorrect option (an alternative conception) a certain time afterward. To complete this picture, some students will, after instruction, systematically chose incorrect options. In the present model, it will be assumed that this fraction of students has failed to learn. Consequently, the following interpretation is proposed for the relative abundances of the five possible correct/incorrect (Post I, Post II) answer pairs:

CC: a *correct* answer immediately after the instruction (Post I), which is maintained a long time later (Post II), denotes a solid, stable scientific knowledge. This CC fraction is postulated to be the quantitative measure of long-time learning.

CI: a *correct* answer immediately after the instruction that turned *incorrect* later is attributed to labile, temporary learning.

$II_=$ and $II_{\neq}$: these *incorrect-incorrect* answer pairs denote the after-instruction presence and persistence of learning difficulties and alternative models. In particular, $II_=$, which measures the fraction of times the same wrong option is selected in both after-instruction test applications, indicates the presence of a very strong, prevalent alternative model, firmly held by students after instruction. Instead, the fraction of answers with different incorrect options, measured by $II_{\neq}$, indicates that students shifted between different distractors (alternative models) in Post I and Post II. In this framework, the total fraction of *incorrect-incorrect* answer pairs, $II_= + II_{\neq}$, is interpreted as a quantitative measure of the failure to learn.

IC: this answer pair corresponds to students that selected an *incorrect* option just after instruction and the *correct* answer in Post II. If the tested subject was not revisited by instruction in the time between Post I and Post II (and consequently no new learning is expected to have occurred in that period), it is assumed that this pair does not represent real knowledge at the time of Post II. Consequently, this answer pair is considered a measure of the experimental error inherent to the use of MCSR tests.

As an example of the type of analysis proposed in the present work, Figure 1 shows the evolution, from Post I to Post II, of students´ answers to Item 22 of DIRECT (Engelhardt & Beichner, 2004) in the CTRL and EXP classes.

Data represented in Figure 1 allow us to identify a few relevant features of the after-instruction evolution of students' answers to this particular item. For the CTRL class the main findings are:

1. The few after instruction (Post I) correct answers (6) changed to incorrect one year later (CI=6).

2. The two correct answers, given one year after instruction, corresponded to incorrect answers in Post I (IC=2).

3. Most incorrect answers given immediately after instruction (Post I) evolved to a different incorrect option one year later ($II_{\neq} = 18$), which is about three times the number of same incorrect options in both tests ($II_= = 5$).

For the EXP sample, the situation is quite different:

1. Correct-correct is the most abundant answer pair (CC=15).

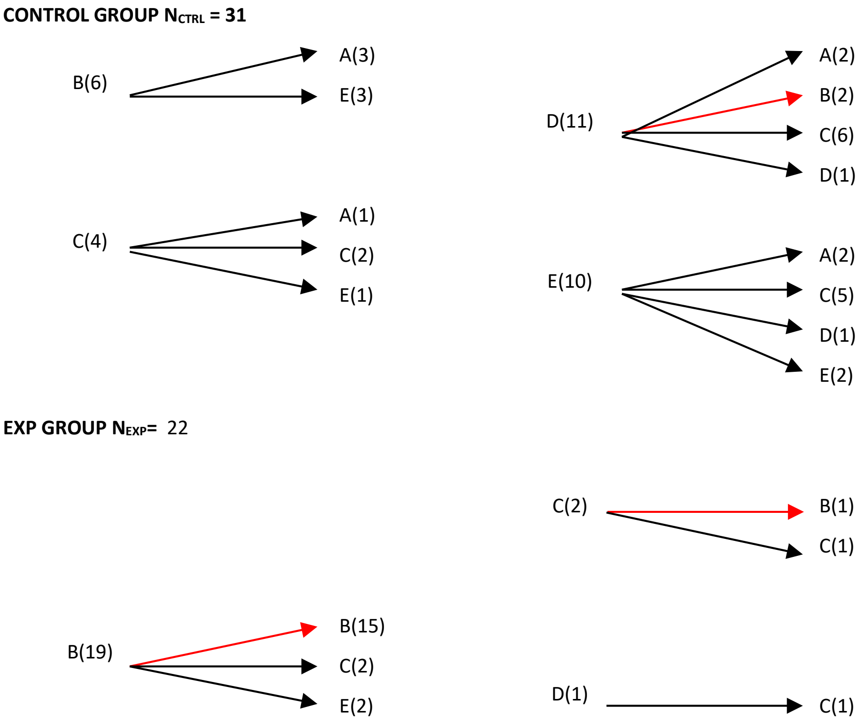
2. Only four initially correct answers turned incorrect (CI=4).

3. A very low number of incorrect-incorrect answer pairs ($II_{\neq} = 1$ and $II_= = 1$)

4. Non-significant number of incorrect to correct answer pair (IC= 1)

Figure 1

Evolution of Students´ Answers to Item 22 of the Test DIRECT from Post I to Post II



Note. For each answer choice (A to E) the numbers within parenthesis indicate the number of students selecting that choice. Arrows indicate how the answers in Post I evolve to Post II. Correct Answer: B.

A similar analysis of the other test items, and normalizing by the total number of answer pairs, allowed us to calculate the course average abundances of the five answer pairs shown in Table 1. Although it is beyond the scope of this work, this procedure also allows for more in-depth studies. Analysis by learning objective/dimension or by learning difficulty/alternative model can be readily carried out because the authors of the relevant RB-MCSR test usually identify or separate the test items in that manner. Similarly, the method could also be used to study different factors that might influence the learning processes, such as prior knowledge, reasoning ability, interest, academic achievement, self-concept, gender, and so on. The time series can also have more than two points, searching for the characteristics of the processes determining the loss of knowledge with time.

Results

The results of this experiment, separated for the CTRL and EXP groups are summarized in Table 1, which shows the statistical parameters of the traditional and new methods. The test items have been arranged according to the learning objectives proposed by the DIRECT test (Engelhardt & Beichner, 2004), relevant to the present experience: ¨Physical Aspects of DC Circuits" and "Current and Voltage." For each student´ group, the bottom row shows the corresponding whole class average results for the 19 items of DIRECT under analysis here. Columns 3 to 7 correspond to the statistical parameters calculated following a traditional MCSR test analysis: Pre, Post I, and Post II average course performances, and the corresponding normalized gains g_I and g_{II}. A simple inspection of Table 1 shows, in both groups, a rather similar performance behavior for both objectives, with small variations respect the total (bottom) row. A first general result is the important after-instruction performance difference between the two groups in both DIRECT objectives and for all tested items. An independent samples t-test shows that the Post I average performance of the EXP group is statistically higher than the CTRL sample performance ($t=5.573$, $df=59$, $p<0.001$). Similar results are found for the one-year after-instruction performances ($t=6.901$, $df=51$, $p<0.001$), which determine an effect size (Connolly, 2007) of 0.698 for the long-term performances. The difference in df happens, as noted in Methods, because 8 students of the EXP sample were absent at the time of the Post II evaluation, therefore the one-year after-instruction statistical parameters were calculated over the 22 students of the EXP sample that participated in all tests. The last row of each group also shows how time affects knowledge, with a mean performance drop of about 20% between Post I and Post II in both samples. This performance drop results in a drop in the normalized gains, $Dg= g_{II} - g_{I}$, of about -0.20, also very similar in both samples.

Table 1

Average Students´ Performances and Relative Abundances of the Five (Post I, Post II) Answer Pairs by Objective of the Test DIRECT.

DIRECT	Item #	Pre	POST	POST	g_I	g_{II}	CC	CI	$II_=$	$II_{\neq}$	IC
OBJECTIVE			I	II							
CTRL Group											
Physical aspects of DC Circuits	4,5,9,10,13,14,1819,22,23,27	21	40	20	0.24	-0.01	10	30	12	38	11
Current and Voltage	6,8,15,1617,26,28,29	18	38	21	0.24	0.04	11	27	16	36	10
TOTAL	All tested items	20 (10)	39 (20)	21 (13)	0.24	0.01	10 (12)	28 (10)	14 (9)	37 (16)	11 (6)
EXP Group											
Physical aspects of DC Circuits	4,5,9,10,13,14,1819,22,23,27	13	74	55	0.70	0.48	50	24	5	15	6
Current and Voltage	6,8,15,1617,26,28,29	10	62	40	0.58	0.33	38	35	12	14	2
TOTAL	All tested items	12 (7)	68 (22)	49 (16)	0.64	0.42	44 (20)	29 (13)	8 (7)	14 (12)	4 (5)

Note. Columns, from left to right, indicate DIRECT learning objective, DIRECT items that evaluate that objective, percent values of the average class performances in Pre, Post I, and Post II. The next two columns to the right indicate the normalized gains g_I (Post I relative to Pre) and g_{II} (Post II relative to Pre). The last five columns on the right show the percent values of the Correct-Correct, Correct-Incorrect, Incorrect-different Incorrect, Incorrect-same Incorrect and Incorrect-Correct, (Post I, Post II) answer pairs. TOTAL row represents the corresponding mean values (and standard deviations) over all tested items

The results of the present approach are represented by the (Post I, Post II) answer pairs shown in the last five columns on the right of Table 1. The first, striking result is the large difference in the CC pairs. Another feature is that these data show again, in both samples, similar behaviors by objectives and for the total of tested items. For the CTRL sample, Table 1 shows that about half (51%) of the answer pairs correspond to the incorrect-incorrect group, most of them of the different-incorrect options subgroup ($II_{\neq}$ = 37 %). Short-lived learning (CI) represents the second most relevant group (28%), while only 10% of the answers correspond to the CC pair, which measures long-time learning. The situation is very different for the EXP sample, where long-time learning is the most abundant category (CC=44%), i.e., after instruction almost half of the time these students systematically selected the correct answer. The CI answer pair, representative of short-lived learning, is again almost 30 %, while the fraction of incorrect-incorrect answer pairs is reduced by a factor of two, to 22%. It is also observed that, within our statistics, the EXP sample showed some preference for selecting, in both tests, the same wrong model ($II_{\neq}$ = 1.5 $II_{=}$), as compared to the CRTL sample ($II_{\neq}$ ~ 2.6 $II_{=}$). An independent sample t-test on the CC pair performance shows that there is a significant difference between these two samples concerning the selection of the CC pair (t=8.835, df=51, $p < 0.001$). If students are grouped according to their CC performance, it is found that 61% of the CTRL sample selected only between 0 and 10 % of CC pairs, with another 26% of this sample selecting between 10% and 30 % of the time a CC pair. The situation is almost reversed in the EXP sample where 50% of the sample selected CC pairs more than 50% of the time, with another 23% of this group selecting CC pairs between 30% and 50% of the time. These findings are reinforced by calculation of the loss parameter, L= CI/(CI+CC) (Lasry et al., 2014), which indicates the fraction of correct answers in Post I that turned incorrect in Post II. Data from Table 1 yields L_{CTRL}= 0.79 (0.19) and L_{EXP}= 0.39 (0.13), i.e., losses in the CTRL group double losses in the EXP group, pointing again to the labile nature of learning generated by traditional instruction.

Discussion

This work presents an alternative method of calculating long-term learning using data from a longitudinal study consisting of two post-instruction applications of the same RB-MCSR test. Traditional analysis, represented by the Pre, Post I, Post II, g_I, and g_{II} data of Table 1, indicates that some knowledge is lost with time and that this loss can be measured as the difference between Post I and Post II, or through the differences between the corresponding normalized gains g_I and g_{II}. Large differences between the two samples are observed in the Post I and Post II data. Surprisingly, the CTRL sample returned, one year after instruction, to the very low pre-instruction knowledge. This fact is reflected by the almost null value of long-term normalized gain g_{II}. Much to the contrary, the corresponding learning parameters of the EXP group show an important knowledge level, even one year after instruction.

The new approach presented in this work allows for more in-depth analysis. For instance, Table 1 shows that the important changes in long-time learning between the two groups are due to the large difference in the “no-learning” groups – about 50% in the CTRL sample -, which is reduced by a factor of two in the EXP sample. If we imagine these three learning categories as steps of a “learning ladder,” our data suggest that about 25-30% of the EXP sample has moved one-step up this ladder as compared to the CTRL sample. This change results in a notable (four times) increase of the fraction of answer pairs denoting durable learning, but with similar values of the temporary learning (the CI pair).

The relative abundance of the no learning categories is also worthy of analysis. While in the EXP group there is a clear predominance of same-incorrect distractors, in the CTRL group the number of students choosing the same-incorrect options is about 1/3 of those selecting different-incorrect answer pairs, which seems to indicate no preference for a particular distractor (in this test with four distractors/item). In terms of the Model Analysis of Bao and Redish (2006), the EXP group seems to be challenged by one prevalent learning difficulty (pure, but incorrect, model state in that framework), while answers in the CTRL group shifted between different-incorrect models, showing no preference for a particular alternative model (mixed model state). In that regard, Bao and Redish (2001) showed that the presence of two or more relevant distractors, implying that most students don't have a strong preference for any model on this topic, results in responses close to random guesses. This combination of *l*ow performance and *l*ow concentration of answers on a given option (the LL region in their model) characterizes uninstructed student samples. This position seems to confirm that, one year after instruction, there is little sign of the instruction received by the CTRL sample.

Finally, Table 1 shows the IC pair is more than twice larger for the CTRL sample compared to the active learning class. Since no instruction on the tested subject was given in the period between Post I and Post II, it has been assumed that this answer pair should not be considered as real understanding at the time of Post II. Consequently, it has been interpreted as the experimental error intrinsic to the use of multiple-choice tests. This position seems also supported by the adopted learning model (Bao & Redish, 2006; di Sessa, 1993; Vosniadou, 1994), which postulates that individuals that have not acquired the scientific model (the fraction of wrong answers in Post I) might change their answers without being particularly aware of it. In other words, we can assume that the evolution of their answers from Post I to Post II should be close to random. If this were the case, the measured IC pair should be the result of all incorrect answers in Post I that evolve randomly to Post II, yielding, for the present case, a value of the IC pair of 0.06 for the EXP group and 0.12 for the CTRL group, i.e., very close to the IC values shown in Table 1. According to this interpretation and values of the IC answer pair, the measuring error also seems to depend on the effectiveness of the teaching strategy.

Although the aim of this work is about measuring long-time, durable learning, it seems worthwhile to highlight a few points from the instructional point of view. First, and despite the large differences in the efficiency of the two teaching strategies, it is clear that even adopting a successful active-learning pedagogy, there is plenty of room for improving learning outcomes. As noted above, one out of three answer pairs selected by students of the EXP sample denotes short-lived learning. Considering labile learning as a transition state between the absence of learning and long-lived learning, it is clear that a relevant fraction of learners accomplished only precarious, unstable learning, and that further actions should be taken to consolidate the scientific model. In this regard, and since active learning teaching strategies are based on pedagogical principles that foster deep learning (Biggs, 2003; Meltzer and Thornton, 2012; Prosser and Trigwell, 1999), a reasonable recommendation is to strengthen this teaching position. One straightforward approach is to use complementary active learning strategies in the different activities of a given course (lectures, problem-solving, labs, etc.). This simple pedagogical approach, much in line with that proposed, for instance, by the Activity Based Physics Suite (Redish, 2003; The Physics Suite, 2015) explicitly avoids the drawbacks of the simultaneous use of conflicting learning approaches (Guidugli, Fernandez Gauna and Benegas, 2005). In the present case, for instance, the two Tutorials on DC circuits used by the EXP class could be complemented with the Interactive Lecture Demonstrations (Sokoloff and Thornton, 2004) "Introduction to DC circuits" and "Series and Parallel Circuits." This small change should provide further learning opportunities using only two extra hours of teaching time. Since these active learning strategies make use of coherent pedagogical principles to confront students with their learning difficulties, this approach should also be efficient for improving learning in the "no-learning" group.

Conclusions

The aim of this work has been to present a simple and more accurate approach to determine the fraction of students that, after instruction, achieve a solid long-term knowledge as compared to those getting only temporary, short-lived learnings. The method, based on the categorization of all possible answer pairs obtained from two after-instruction applications of a research-based MCSR test, readily provides not only a quantitative determination of long-term and temporary learnings but also the fraction of answer pairs associated with the absence of learning. Furthermore, the method allows separation of the "no-learning" group into two categories, i.e., those students that, after instruction, systematically selected the same incorrect option from those that shifted between different distractors. As noted in the previous sections, these features could furnish relevant insights regarding the characteristics of the learning obstacles faced by students.

Even though in the present classroom experiment both methods of analysis show that long-time conceptual learning is clearly higher in the experimental group, the new approach is more accurate than the standard determination of enduring learning. For instance, if one takes the results of Post II (Table 1) as a measurement of long-time learning, the achievement of the CTRL class would be overestimated by a factor of two (21% performance in Post II vs 10% of the CC pair). On the contrary, a similar comparison for the EXP sample results only in a 10% difference (49% vs. 44%, respectively). Since the IC pair, interpreted here as the experimental error, has been shown to depend on the type of instruction, the above results seem to confirm this dependence of the measuring error on the effectiveness of instruction. In terms of Soderstrom & Bjork (2015) model, the classical determination of *Learning* would be given by the results of Post II. The present model allows us to refine this measurement, correctly assigning the CC answer pair value to this long-time learning, leaving out the experimental error contribution to Post II.

The extremely low long-lasting learning achieved by the CTRL group could not be just idiosyncratic of the student groups analyzed in this study. Similarly, low conceptual knowledge (about 10%) has been reported for all relevant areas of basic physics (force and motion, free-fall motion, and Coulomb´s Laws) by the broad study cited above (Benegas et al., 2009; 2010). Although belonging to different school systems and countries, the common point of these student samples of first-year university students is that they had been subjected to traditional, lecture-based high school instruction. Therefore, the above results of the CTRL group provide a plausible explanation for the surprisingly low level of conceptual understanding, uniformly shown by these samples of incoming university students. In this regard it is noted that, since the proposed method is easily applicable to large-scale assessments, it should be of help to school officials that very frequently need an easy-to-use tool to measure the real, enduring impact of instruction on students´ conceptual knowledge.

Overall, this analysis makes clear that a substantial amount of basic and applied educational research is needed to improve our knowledge of the processes leading to solid, long-lasting conceptual learning, and to develop teaching approaches to achieve this goal. We think that these educational issues deserve further research and that the novel approach for measuring long-time learning presented here might be of help for designing and carrying out appropriate experiments.

This work was supported by the Universidad Nacional de San Luis, Argentina (grant number PROICO 03-0416).

Julio Benegas (jcbenegas@gmail.com) is Emeritus Professor of Physics at Universidad Nacional de San Luis. His current research focuses on applied research in physics education including the development and application of active-learning teaching strategies in physics and mathematics.

Julio Sirur Flores (juliosirur@gmail.com) is a high school physics teacher in San Luis, Argentina and an adjunct professor at Universidad Nacional de San Luis. His current research focuses on applied research in physics education including the development and application of active-learning teaching strategies in physics.

References

Bao, L. & Redish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *American Journal of Phys*ics, 69, S45. https://doi.org/10.1119/1.1371253

Bao, L. & Redish, E. (2006). Model analysis: representing and assessing the dynamics of student learning. *Physical Review Special Topics - Physics Education Research* 2, 010103. https://doi.org/10.1103/PhysRevSTPER.2.010103

Benegas, J., Villegas M., Pérez de Landazábal, M. del C. & Otero, J. (2009). Conocimiento conceptual de física básica en ingresantes a carreras de ciencias e ingeniería en cinco universidades de España, Argentina y Chile. *Rev. Iberoamericana de Física*, 35(1), 35.

Benegas, J., Pérez de Landazábal, M. del C. & Otero, J. (2010). Estudio de casos: Conocimientos físicos de los estudiantes cuando terminan la escuela secundaria: una advertencia y algunas alternativas. *Revista Mexicana de Física*, 56(1), 12–21. https://rmf.smf.mx/ojs/rmf-e/article/view/4621

Benegas, J. & Sirur Flores, J. (2014). Effectiveness of Tutorials for Introductory Physics in Argentinean high schools. *Physical Review Special Topics - Physics Education Research,* 10, 010110. https://doi.org/10.1103/PhysRevSTPER.10.010110

Bernhard, J. (2001) Does active engagement curricula give long-lived conceptual understanding? *Physics Teacher Education Beyond 2000*, edited by Pinto, R. & Surinach, S. Paris: Elsevier, pp. 752-759.

Biggs, J. (2003) *Teaching for Quality Learning at University* 2nd Ed. London: The Society for research into Higher Education & Open University Press.

Connolly, P. (2007). *Quantitative data analysis in Education. A critical introduction using SPSS*. London: Routledge. https://doi.org/10.4324/9780203946985

Engelhardt, P. V. & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics*, 72 (1), 98-115. https://doi.org/10.1119/1.1614813

diSessa, A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction,* 10(2/3), 105-225. https://doi.org/10.1080/07370008.1985.9649008

Francis, G., Adams, J. & Noonan, E. (1998). Do They Stay Fixed? *Physics Teach*er, 36, 488–490. https://doi.org/10.1119/1.879933

Guidugli, S., Fernandez Gauna, C. & Benegas, J. (2005). Graphical Representations of Kinematical Concepts. A comparison of Teaching Strategies. *Physics Teacher*, 43, 334-337. https://doi.org/10.1119/1.2033514

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, 64–74. https://doi.org/10.1119/1.18809

Hestenes D., Wells M. & *Swackhamer* G., 1992. "Force Concept Inventory," *The Physics Teacher* 30:3, 141-158. https://doi.org/10.1119/1.2343497

Kohlmyer, M.A., Caballero, M. D. Catrambone, R., Chabay, R. W., Ding, L., Haugan, M. P., Jackson Marr, M., Sherwood, B. A. & Schatz, M. F. (2009). A Tale of Two Curricula: The performance of 2000 students in introductory electromagnetism. *Physics Review Special Topics-Physics Education Research*, 5, 020105. https://doi.org/10.1103/PhysRevSTPER.5.020105

Lasry, N., Guillemette, J. & Mazur, E. (2014). Two steps forward, one step back. *Nature Physics,* 10, 402-403. https://doi.org/10.1038/nphys2988

McDermott, L. C., Shaffer, P. S. & The Physics Education Group (1998). Tutorials in Introductory Physics. New Jersey: Prentice Hall Inc. https://doi.org/10.1063/1.53118. Translated as *Tutoriales para Física Introductoria*. Buenos Aires: Pearson Education (2001).

Meltzer, D.A. & Thornton, R.K. (2012). Resource Letter ALIP–1: Active-Learning Instruction in Physics, *American Journal of Physics*, 80 (6), 478-496. https://doi.org/10.1119/1.3678299

OECD (2019). Organization for Economic Cooperation and Development: PISA 2018 Results, https://www.oecd.org/publications/pisa-2018-results-volume-i-5f07c754-en.htm (accessed December 10, 2019).

Persano Adorno, D., Pizzolato, N. & Fazio, C. (2018). Long term stability of learning outcomes in undergraduates after an open-inquiry instruction on thermal science. *Physics Review Special Topics-Physics Education Research*, 14, 010108. https://doi.org/10.1103/PhysRevPhysEducRes.14.010108

Pollock, S. J. (2009). A longitudinal study of student conceptual understanding in Electricity and Magnetism. *Physics Review Special Topics-Physics Education Research* 5, 020110. https://doi.org/10.1103/PhysRevSTPER.5.020110

Prosser, M. & Trigwell, K. (1999). Understanding Learning and Teaching, on Deep and Surface Learning. London: Society for Research into Higher Education & Open University Press, chapter 4.

Redish, E. (2003). *Teaching Physics with the Physics Suite.* Hoboken, NJ: Wiley.

Soderstrom, N. C. & Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, *10*, 176–199. https://doi.org/10.1177/1745691615569000

Sokoloff, D. R. & Thornton, R. K. (2006). *Interactive Lecture Demonstrations.* Hoboken, NJ: Wiley.

The Physics Suite (2015) http://www.wiley.com/college/sc/cummings/suite.html. (Accessed February 4, 2015).

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, 4, 45. https://doi.org/10.1016/0959-4752(94)90018-3