Assessment of Reproducible Computing as an E-Learning Tool in Statistics Education

Patrick Wessa[1]

[1]K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium

patrick@wessa.net

Abstract: Statistics Education, Constructivism, Reproducible Computing, R language

This paper attempts to assess the learning-related satisfaction of students who have participated in an experimental, constructivist, undergraduate statistics course that is based on a newly developed Computational R Framework (http://www.wessa.net/) and the Compendium Platform for Reproducible Computing (http://www.freestatistics.org/). The analysis of the survey responses is based on aggregated measures that are easy to interpret and clearly demonstrate that the current implementation of this new Reproducible Computing technology is (very) successful (in terms of reported student satisfaction).

1. Introduction

Within the context of computer-assisted and mathematical education, the pedagogical community has shown great interest in the role and importance of social and individual constructivism ([Von Glasersfeld, 1987], [Smith, 1999], [Eggen and Kauchak, 2001]) and its implementation in statistics education in particular ([Mvududu, 2003]). While the relevance of a constructivist pedagogical paradigm is well documented there seems to be no direct or obvious relationship with the problem of irreproducible research. Nevertheless, the problem of our inability to reproduce statistical computations that are presented in papers has received quite a bit of attention within the statistical computing community. The most prominent citation about the problem of irreproducible research is Claerbout's principle (source: [de Leeuw, 2001]):

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures.*

The importance of the irreproducibility problem has been highlighted by many authors and is related to science, the dissemination of science, and academic education. Some of the leading arguments can be found in [Peng et al., 2006], [Schwab et al., 2000], [Green, 2003], [Gentleman, 2005], [Koenker and Zeileis, 2007], [Donoho and Huo, 2004]. In his comments [de Leeuw, 2001] defines the following additional requirements:

- any type of computational output should be reproducible
- reproducibility should be assured for all academic publications, course texts in particular
- the software environment should be freely available

Several approaches to solve the problem have been suggested and implemented. Some of the more promising attempts have been described in [Buckheit and Donoho, 1995], [Donoho and Huo, 2004], [Leisch, 2003]. These solutions however, have not been implemented in statistics education because of several reasons that make them impractical for students:

- students are required to download, install, and execute software on their local machines
- students have to understand the underlying technicalities (such as LaTeXand R)
- it is not easy for students to create reproducible documents (for example when they have to submit a term or assignment paper)

In addition, and most importantly, the existing solutions have not been designed with educational research in mind. Computational and learning-related activities are not measured and stored for the purpose of quality control or research - even though there is an active interest in measuring and exploring educational activities within e-learning environments [Romero C., 2008]. Allowing academic educators to do research about student's learning bevhavior might prove a strong incentive to improve the quality of course materials, software, pedagogical approaches, etc ...

The solution that is proposed within the context of this paper is new and differs from previously developed solutions in the sense that it can be used by anyone and without the need to understand the technicalities of scientific word processing (LaTex) or statistical programming (R code) [R Development Core Team, 2008]. In addition, all computations are performed through a distributed computing network of servers which implies that the user only needs a browser and a live internet connection. Finally, the e-Learning system that was created automatically stores all software-related activities of students - this includes the use of statistical software, the creation of documents that are reproducible, and communication streams between students that are related to peer review.

This paper attempts to assess the learning-related satisfaction of students who have participated in an undergraduate statistics course that is based on the Computational R Framework and the Compendium Platform for Reproducible Computing. For this purpose we use standardized surveys which measure student's satisfaction about their learning experiences, and computer system usability. The main purpose of this paper is to show that the technology that facilitates Reproducible Computing (as developed in our computer-assisted statistics course) enjoys a high degree of reported student satisfaction even if the use of such technology is associated with a heavy workload.

2. Compendium Platform

The R Framework (hosted at http://www.wessa.net) allows educators and scientists to develop new, tailor-made statistical software (based on the R language) within the context of an open-access business model that allows us to create, disseminate, and maintain software modules efficiently and with a very low cost in terms of computing resources and maintenance efforts [Wessa, 2008a]. The so-called R modules empower students to perform statistical analysis through a web-based interface that does not require them to download or install anything on the client machine. This permits students to focus primarily on the interpretation of the analysis - however, the R Framework also allows advanced students and scientists to inspect and change the R code that was coded by the original author. This results in the creation of so-called derived R modules that may be better suited for particular purposes.

If a derived R module contains generic improvements or if a computation needs to be communicated to other students/scientists then it is necessary to have a simple, transparent mechanism that allows one to permanently store the computation in a repository of computational objects that can be easily retrieved, recomputed, and reused. Such a repository was recently created within the OOF 2007/13 project of the K.U.Leuven Association and is called the Compendium Platform. The main reason for creating the R Framework and the Compendium Platform, is that it allows anyone to create and use Compendia of reproducible research. A Compendium is defined as [Wessa, 2008c]: any document with (open-access) references to (remotely) archived Computations (including Data, Meta-data, and Software) that allow us to reproduce, and reuse the underlying analysis. Such documents can be easily created (even by students) and permit any reader to (exactly) recompute the statistical results that are presented therein. A few simple clicks are sufficient to have the R Framework reproduce the results and to reuse them in derived work. The practical implications of this technology are explained in [Wessa, 2008c].

3. Course Design

The main sections of the statistics course are built around a series of research-based workshops that require students to reflect and communicate about a variety of statistical problems, at various levels of difficulty. The workshops have been carefully designed and cannot be solved without additional information that is provided within the Virtual Learning Environment or by the tutor.

Based on reported information from students and extrapolations based on web server log files, I estimate that each workshop involves about 9 hours of work per student, per week. In addition, students were required to perform detailed peer reviews of about 5-7 submissions from other students. Even though students had to assess the submitted workshops and give them a score, the peer review was not intended as an evaluation method (it did not count towards their final score). On the other hand, it enabled students to provide feedback, learn from mistakes made by others, communicate solutions about a variety of problems, and provide an incentive in the form of encouragement to fellow students.

This feedback-oriented process is similar to the peer review procedure of an article that is submitted to a scientific journal. The process of (anonymous) assessment by peers is an intrinsic part of scientific endeavour, and may help students in nurturing their scientific attitudes (through peer review experiences) and non-rote learning (through construction of knowledge).

A group of 240 undergraduate business students participated in the course and completed a total of 1907 workshops which were subjected to peer review. Every submission was assessed with respect to 3-6 criteria. For every graded criterion students had the ability to provide verbal feedback to the other student. As a consequence, a total of 41960 grades and 34438 verbal feedback communications were received by students. This implies that, on average, 22 grades and 18 verbal feedback messages were generated (per workshop, per student).

Fortunately, this did not require any intervention by me: the otherwise time-consuming administration of the Peer Assessment procedure was automatically performed by the use of the Virtual Learning Environment called Moodle [Moodle, 2008] which is freely available. One of the main reasons why Moodle features the administration of Peer Assessment is the fact that it has been designed with a constructivist, pedagogical philosophy in mind. The grades that were generated by the peer review process did not count towards the final score of students. Instead, I graded the quality of the verbal feedback messages that were submitted to other students based on semi-random sampling techniques.

The course contains a wide variety of statistical techniques and methods such as: probability, discrete and continuous distributions, descriptive statistics, explorative data analysis, hypothesis testing (about the mean, the variance, and proportions), multiple linear regression, and univariate time series analysis (Box-Jenkins analysis). A total of 73 different types of statistical techniques are covered by the course with a large variety of model parameters.

For each technique, students had one or several web-based software modules available. The modules are based on the R Framework and are available free of charge at http://www.wessa.net/. The R Framework allows educators and scientists to develop new, tailor-made statistical software and at the same time the end-user is able to change the underlying source code and improve the software [Wessa, 2008a].

There is strong empirical evidence that the use of Reproducible Computing is related to non-rote learning of statistical concepts which is measured by objective exam questions [Wessa, 2008b]. In addition, it can be shown that the Compendium Platform allows educators to improve the e-learning experiences because the underlying technology allows us to perform monitoring and control of activity-based learning processes based on actual, objective measurements that are otherwise not available [Wessa, 2008d].

4. Data

The survey data were obtained from three well-known questionnaires (ATTLES, COLLES, and CSUQ). The response rate for each survey was extremely high because it was easily accessible (within the learnig environment) and because the importance of the survey results for our research was explained in great detail.

The first survey (called ATTLES) is available in Moodle (as a standard questionnaire) [Moodle, 2008] and aims to measure student's attitudes towards thinking and learning [Galotti et al., 1999]. The first ten questions relate to "connected" (empathic) ways of learning whereas the ten last questions are associated with "separate" (critical, detached) ways of knowing: http://www.freestatistics.org/moodle/mod/survey/view.php?id=36. The non-response rate was 8%.

Students perception of their online learning experience during the semester was measured with the Constructivist On-Line Learning Environment Survey (COLLES) as implemented in Moodle [Moodle, 2008]. The survey focused on a spectrum of important aspects: relevance, reflection, interaction, educator, peers, and understanding (for a complete list see: http://www.freestatistics.org/moodle/mod/survey/view.php?id=37). For every aspect there are eight questions, four of which are related to the actually perceived experience. The remaining four questions have identical phrases but are related to the degree of what students prefer. The survey was submitted by the students before receiving the scores of the multiple choice test. The non-response rate was 15%.

The third survey is based on IBM's Computer System Usability Survey (called CSUQ) [Lewis, 1993] with additional questions that were specifically related to the relationship between software usability and statistics learning. The questions were made available within a "Quiz" module in Moodle and can be examined at: http://www.freestatistics.org/moodle/mod/quiz/view.php?id=410. The non-response rate was 17%.

5. Assessment Methodology

The analysis of the survey responses is performed in such a way that anyone is able to interpret the results. Each question was based on a 5-point Likert scale (5 is excellent, 3 is neutral, and 1 is poor). By subtracting a fixed constant (= 3) we obtained scores that are contained in the interval $[-2, 2]$ where the neutral score is zero valued. This score $S_{i,j}$ represents the transformed reply (for all questions $i = 1, ..., Q$ and for all students $j = 1, ..., N$) for which the following definitions can be formulated:

- $D^+_{i,j} = 1$ if $S_{i,j} > 0, D^+_{i,j} = 0$ and $S_{i,j} \leq 0$
- $D^-_{i,j} = 1$ if $S_{i,j} < 0, D^-_{i,j} = 0$ and $S_{i,j} \geq 0$
- P^s_i is the sum of all positive scores: $P^s_i = \sum_{j=1}^{N} D^+_{i,j} S_{i,j}$ for $i = 1, ..., Q$
- N^s_i is the sum of all absolute values of negative scores: $N^s_i = \sum_{j=1}^{N} D^-_{i,j} |S_{i,j}|$ for $i = 1, ..., Q$
- P^c_i is the number of positive scores $P^c_i = \sum_{j=1}^{N} D^+_{i,j}$ for $i = 1, ..., Q$
- N^c_i is the number of negative scores $N^c_i = \sum_{j=1}^{N} D^-_{i,j}$ for $i = 1, ..., Q$

It is now possible to define three aggregated measures (AM) for each question:

1. the arithmetic mean: $\frac{1}{N} \sum_{j=1}^{N} S_{i,j}$ for $i = 1, ..., Q$
2. the difference between positive and (absolute) negative scores, divided by the absolute sum of all scores $\frac{(P^s_i - N^s_i)}{(P^s_i + N^s_i)}$ for $i = 1, ..., Q$
3. the difference between the number of positive and negative scores, divided by the sum of all absolute scores $\frac{(P^c_i - N^c_i)}{(P^c_i + N^c_i)}$ voor alle $i = 1, ..., Q$

The first two measures can be used if a quasi-interval scale can be assumed. The third measure does not make the assumption of a quasi-interval scale because the scores are substituted by frequencies (counts). The drawback of the third measure is that is does not differentiate between extreme answers (±2) and moderate answers (±1). In other words, the third measure has the advantages that are associated with ordinal (rank-based) measures but at a cost of loss of information. The first measure is contained in the interval $[-2, 2]$ and last two measures lie in the interval $[-1, 1]$.

In each survey, and for all questions, a high AM is associated with a favorable situation. A negative AM indicates a weak point that should be considered for improvement.

6. Results

The ATTLES survey scores are shown in Table 1. The survey was used to measure student's attitudes at the start of the semester (before the first workshop was completed). Hence, the results from this computation indicate how good student's attitudes were at the start of the course.

The conclusion from table 1 is positive for most aspects of the ATTLES survey. Negative AMs are found for the following questions:

- The most important part of my education has been learning to understand people who are very different to me.
- I like playing devil's advocate - arguing the opposite of what someone is saying.
- I often find myself arguing with the authors of books that I read, trying to logically figure out why they're wrong.
- I spend time figuring out what's 'wrong' with things. For example, I'll look for something in a literary interpretation that isn't argued well enough.

The negative AM scores for these four questions indicate that our students lack the - arguably - most fundamental attitude of good scientists which allows them to be critical and question any assumption that underlies our thinking or analysis. Hence, the introduction of new learning technologies that allow students to reproduce (c.q. challenge) computations from peers is expected to be difficult and lead to negative learning experiences. If students

Question	mean	(Ps-Ns)/(Ps+Ns)	(Pc-Nc)/(Pc+Nc)
1	0.45	0.55	0.51
2	0.68	0.74	0.72
3	0.58	0.63	0.6
4	0.47	0.57	0.55
5	0.76	0.79	0.77
6	0.74	0.73	0.72
7	0.7	0.73	0.68
8	0.98	0.86	0.84
9	-0.11	-0.16	-0.11
10	0.83	0.82	0.78
11	-0.37	-0.38	-0.35
12	0.91	0.86	0.84
13	1.15	0.79	0.73
14	0.5	0.59	0.54
15	0.22	0.36	0.36
16	-0.59	-0.63	-0.59
17	0.41	0.54	0.5
18	0.53	0.61	0.61
19	0.52	0.66	0.62
20	-0.17	-0.2	-0.13

Table 1. ATTLES survey scores [Wessa, 2008e]

dislike to challenge the analysis of others, they are not likely to appreciate assignments that are related to Reproducible Computing and Peer Assessment.

Fortunately, table 2 shows overwhemling evidence that students perceive their learning experience (at the end of the semester) as positive. This comes as a surprise because of the fact that the course involves a heavy workload, and the observation that Reproducible Computing goes against student's attitudes towards learning and thinking (as measured in the initial ATTLES survey). All questions have a positive AM - some are even close to the maximum score.

Q	(Pc-Nc)/(Pc+Nc)	Q	(Pc-Nc)/(Pc+Nc)	Q	(Pc-Nc)/(Pc+Nc)	Q	(Pc-Nc)/(Pc+Nc)
1	0.65	13	0.81	25	0.91	37	0.61
2	0.9	14	0.85	26	0.95	38	0.81
3	0.69	15	0.61	27	0.84	39	0.1
4	0.93	16	0.81	28	0.93	40	0.33
5	0.54	17	0.51	29	0.84	41	0.69
6	0.92	18	0.69	30	0.96	42	0.88
7	0.54	19	0.57	31	0.86	43	0.6
8	0.91	20	0.81	32	0.92	44	0.86
9	0.76	21	0.24	33	0.5	45	0.87
10	0.91	22	0.51	34	0.76	46	0.97
11	0.84	23	0.43	35	0.4	47	0.86
12	0.91	24	0.81	36	0.7	48	0.94

Table 2. COLLES survey scores (count-based aggregated measures) [Wessa, 2008f]

Table 3 shows that the web-based software was highly rated by students. The only exception is related to question 9: "The website gives error messages that clearly tell me how to fix problems." This negative AM is due to the fact that error messages (produced by the R language) are of a "technical" or "purely statistical" nature. For this reason, students were instructed to archive computational results with error messages and send the link to me (by e-mail). The Compendium Platform allowed me to quickly reproduce errors, detect problems, and solve any computational or software-related issue and report back to the student [Wessa, 2008c]. This method of error handling is not only very efficient - it also provides me with a lot of insight into the nature of problems that are commonly encountered.

Q	(Pc-Nc)/(Pc+Nc)	Q	(Pc-Nc)/(Pc+Nc)	Q	(Pc-Nc)/(Pc+Nc)
1	0.89	12	0.57	23	0.45
2	0.82	13	0.15	24	0.87
3	0.84	14	0.7	25	0.49
4	0.62	15	0.77	26	0.93
5	0.67	16	0.51	27	0.54
6	0.73	17	0.52	28	0.94
7	0.62	18	0.82	29	0.71
8	0.67	19	0.88	30	0.9
9	-0.26	20	0.61	31	0.72
10	0.53	21	0.61	32	0.95
11	0.69	22	0.8	33	0.81

Table 3. CSUQ survey scores (count-based aggregated measures) [Wessa, 2008g]

Some results in Table 3 are of particular interest:

- Q20: Overall, the website was helpful in learning statistics
- Q21: Learning Statistics with this website is more effective than with a traditional handbook
- Q22: I intend to use this website when I need to apply statistics in the future
- Q27: To learn statistics, this website is better than the statistical courses I have had so far

The AM for each of these four questions is larger than 0.5 which implies that students appreciate the fact that the web-based Compendium Platform helps them to learn statistics. The appreciation is very strong and may compensate the fact that the learning process involves alot of work, and that Reproducible Computing goes against their initial attitudes towards thinking and learning.

Overall, the results from Tables 2 and 3 demonstrate that student satisfaction can be very high even though the introduction of the Reproducible Computing technology and associated Constructivist Pedagogy comes at the price of a heavy workload. In addition, students are able to adopt new learning technologies and engage in many types of technology-based learning activities (such as experimentation, communication and collaboration).

On the one hand, this paper clearly suggests that Reproducible Computing was well accepted by my students and beneficial for their learning experiences. On the other hand, the technology that was developed also allows us to obtain accurate measurements about computer-assisted learning that are otherwise not available. For instance, the study in [Wessa, 2008d] suggests that self reported measurements (based on questionnaires) may be strongly biased when compared to the actual activity-based measurements of computer usage, and streams of communication within the learning environment. Hence, the results in this paper are clearly limited in the sense that they only relate to reported measurements as obtained through the ATTLES, COLLES, and CSUQ surveys. Another limitation of this study is that perceived learning experiences may depend on various cofactors that are independent of Reproducible Computing:

- the structure of course materials
- the design and difficulty of workshops
- prior knowledge/education of students
- the role of the educator, etc...

In this sense, this paper does not provide a definitive answer to the question if Reproducible Computing is well-accepted by students or not. On the other hand, it is an illustration of a (very) succesful implementation of Reproducible Computing technology that is embedded in a constructivist pedagogical setting.

However, it cannot be disputed that we now have the technological tools available to thoroughly investigate statistics learning within a constructivist, controllable, and monitorable environment. This allows us to focus future research on new, and uncharted areas of computer-assisted, educational learning processes.

Acknowledgements

This research was funded by the OOF 2007/13 project of the K.U.Leuven Association. I would like thank Ed van Stee, and Stephan Poelmans (HUB) for their contributions.

References

Buckheit, J. and Donoho, D. L. (1995). *Wavelets and Statistics*, chapter Wavelab and reproducible research. Springer-Verlag.

de Leeuw, J. (2001). Reproducible research: the bottom line. In *Department of Statistics Papers, 2001031101*. Department of Statistics, UCLA.

Donoho, D. L. and Huo, X. (2004). Beamlab and reproducible research. *International Journal of Wavelets, Multiresolution and Information Processing*.

Eggen, P. and Kauchak, D. (2001). *Educational Psychology: Windows on Classrooms*. Upper Saddle River, NJ: Prentice Hall, 5th ed. edition.

Galotti, K. M., Clinchy, B. M., Ainsworth, K., Lavin, B., and Mansfield, A. F. (1999). A new way of assessing ways of knowing: the attitudes towards thinking and learning survey (attls). *Sex roles*, pages 745–766.

Gentleman, R. (2005). Applying reproducible research in scientific discovery. BioSilico.

Green, P. J. (2003). Diversities of gifts, but the same spirit. *The Statistician*, pages 423–438.

Koenker, R. and Zeileis, A. (2007). Reproducible econometric research (a critical review of the state of the art). In *Research Report Series*, number 60. Department of Statistics and Mathematics Wirtschaftsuniversitt Wien.

Leisch, F. (2003). Sweave and beyond: Computations on text documents. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria.

Lewis, J. R. (1993). Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. In *Technical Report 54.786*. IBM Corporation.

Moodle (2008). A free, open source course management system for online learning. In *http://www.moodle.org*.

Mvududu, N. . (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education*, 11(3).

Peng, R. D., Dominici, F., and Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Romero C., Ventura S., G. E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51:368–384.

Schwab, M., Karrenbach, N., and Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science & Engineering*.

Smith, E. (1999). Social constructivism, individual constructivism and the role of computers in mathematics education. *Journal of mathematical behavior*.

Von Glasersfeld, E. (1987). Learning as a constructive activity. In *Problems of Representation in the Teaching and Learning of Mathematics*, pages 3–17. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wessa, P. (2008a). A framework for statistical software development, maintenance, and publishing within an open-access business model. *Computational Statistics*.

Wessa, P. (2008b). How reproducible research leads to non-rote learning within a socially constructivist e-learning environment. In *Proceedings of the 7th European Conference on e-Learning*, Cyprus.

Wessa, P. (2008c). Learning statistics based on the compendium and reproducible computing. In *Proceedings of the International Conference on Education and Information Technology*, Berkeley, San Francisco, USA.

Wessa, P. (2008d). Measurement and control of statistics learning processes based on constructivist feedback and reproducible computing. In *Proceedings of the 3rd International Conference on Virtual Learning*.

Wessa, P. (2008e). Statistical computations. In *FreeStatistics.org*. Office for Research Development and Education, http://www.freestatistics.org/blog/date/2008/Sep/07/t1220787368wfytqddnvxjq53g.htm.

Wessa, P. (2008f). Statistical computations. In *FreeStatistics.org*. Office for Research Development and Education, http://www.freestatistics.org/blog/date/2008/Sep/07/t1220794944bzrkmfrlic23vxu.htm.

Wessa, P. (2008g). Statistical computations. In *FreeStatistics.org*. Office for Research Development and Education, http://www.freestatistics.org/blog/date/2008/Sep/07/t1220796393qktqk00wbmi19d7.htm.