Reviewing Peer Reviews – A Rule-Based Approach

Patrick Wessa[1] and Antoon De Rycker[2]
[1]Catholic University of Leuven, Belgium
[2]University of Malaya, Malaysia
patrick@wessa.net
teundr@um.edu.my

Abstract: Peer reviews may be used as an effective tool in non-rote learning, especially if one wishes to provide learners with computer-assisted learning environments within the general theory of pedagogical constructivism. Our past research and development efforts have resulted in the creation of an innovative approach to statistics education in which peer review activities, based on reproducible computing technology, play an important role in the constructivist learning of statistical concepts. The associated research has also shown that many types of objective measurements are available and that these are of the utmost importance when explaining students' learning outcomes. The first problem that is addressed in the current paper is how these peer review measurements can be used to predict learning outcomes. In order to engage students in taking peer reviews seriously they should be motivated. This can be achieved by reviewing and grading the peer reviews that they submit during the course. Any such attempt is likely to raise the problem, however, of how the educator should review the reviews. This is the second problem that will be discussed in our paper. Using a theoretical framework as our starting point, we are able to derive empirical rules that allow us to perform Peer Review-Reviews under the principles of implementability, predictability, comparability and purposefulness. It may be argued that the review of peer reviews becomes obsolete if students have to write a term paper which is going to be evaluated by the educator. However, the underlying rationale is that a term paper involves all concepts that are deemed important while the peer review of an assignment only focuses on a small number of topics. The third problem that is investigated in this paper focuses on the relative importance of grading term papers versus reviewing peer reviews. In other words, is it more efficient to have students submit a term paper which is graded by the educator, or is it better to review the Peer Reviews of weekly assignments? Our findings will be based on a detailed quantitative analysis of the peer review and student assessment data that were collected over a period of three years, involving 285 university-level students who took an introductory statistics course.

Keywords: peer review, assessment, pedagogical constructivism, reproducible computing, technology

1. Introduction

Even though peer reviewing is perceived as one of the important, formative assessment tools, there seem to be only few studies in which the effects on learning outcomes are actually tested (Strijbos et al. 2009a, 2009b; van Gennip et al. 2009). A plausible explanation may be the lack of easily testable foundations for relating peer review activities, rather than reported learning effects, to objectively measured learning outcomes (Strijbos et al. 2009b). Still, the availability of innovative eLearning tools provides us with ample opportunities to study the role and effects of peer review in computer-assisted learning.

A related issue is that many educators but also researchers perceive peer review as an assessment and grading tool rather than a collaborative learning activity (e.g. Dabbagh 2005) rooted in the traditions of pedagogical constructivism, experiential learning, learner autonomy and similar concepts (van Gennip et al. 2009). Pedagogical constructivism loosely refers to a view of meaningful learning associated with such scholars as Piaget, Vygotsky, Ausubel and Novak. Following Henze (2008), it can be more strictly defined as a learning theory that claims that knowledge is derived from both individual and social experiences within a broader historical and societal context, that learning is not a “spectator sport” but an ”process of communicating, discovering, organizing and conceptualizing”. It explicitly does not include any metaphysical or epistemological presuppositions. It is clear that peer evaluation and peer review activities stimulate this kind of constructivist learning. To quote (Strijbos et al. 2009b), peer assessment is “an interactive and communicative process in the service of learning” and “a cyclical and interactive process”.

However, if viewed solely as an assessment tool, peer review practices may restrict a learner’s freedom to experiment, to be creative and to collaborate in the joint construction of knowledge and the negotiation of alternatives through debate and argumentation (Dabbagh 2005). If grades are all important and the only visible incentive, then students are likely to engage in copying and other free-riding behaviour rather than take the time to develop their non-rote learning skills.

Almost all empirical peer review studies focus on the effect on the receiver of the feedback, i.e. the reviewee (Strijbos et al. 2009a). One notable exception is a study that investigates the benefits to both the receiver and the reviewer (Lundstrom et al. 2009). Their empirical findings clearly show that the reviewer benefits more than the receiver. Though perhaps surprising at first sight, this observation makes perfect sense if we realise that writing a peer review involves cognitive processes that encourage deep learning. In contrast, receiving review messages may or may not involve actions that might impact learning or thinking. ELearning tools cannot measure what happens with feedback messages that are received, e.g., opening a web page does not necessarily imply intensive reading and comprehension. Even so, the benefits of peer review to either the reviewer or the reviewee are not generally accepted and still cause a lot of debate:

> *Literature reviews [...] indicate that although various studies seem to have found positive effects of peer assessment on learning, the results are still inconclusive. Moreover, it is unclear under what conditions peer assessment is effective. (van Gennip et al. 2009)*

As a final point, the literature on peer review is primarily focused on language education and the teaching of writing skills. The concept of peer review-based learning in university-level statistics education is largely uncharted. This is odd as our ability to critically review statistical papers has never been disputed. In passing, the problem of irreproducible research and the proposition of accessible solutions has received a great deal of attention within the statistical community (Wessa 2009c). If statisticians find it difficult (if not impossible) to reproduce the empirical findings reported in scientific papers, then it is extremely unfair to expect students to be able to reproduce, and make sense of, empirical results that are presented in course materials and research papers. It is for this reason that we have been engaged in the development of a novel Reproducible Computing technology that allows anyone to produce an empirical paper (the so-called “Compendium”) that can be reproduced without the need to install software or the need to understand the underlying technicalities (Wessa 2009c). A more detailed discussion would lead us too far; for present purposes it is sufficient to observe that Reproducible Computing supports peer review and collaborative work.

To return to the lack of research into peer reviewing activities in statistics education, one notable exception is Wessa (2009a). This study found that the submission of peer reviews is strongly related to learning outcomes insofar as they are measured *objectively*, i.e. by means of independent summative exams that set out to assess the true understanding of statistical concepts rather than rote memorization.

The main contribution of the present paper lies in our attempt to take this objective, measurement-based approach one step further. More particularly, our objective is to specify a theoretical framework for Peer Review-Review (henceforth PRR) that is based on unambiguously defined and well-measured concepts that can be easily implemented and which predicts learning outcomes sufficiently accurately. In Section 2 of this article we will first describe the main design features of our PRR model. Next, Section 3 will present the implementation of the model in an introductory statistics course and discuss its predictive capabilities. In Section 4 we will test these predictive capabilities empirically. The significance of the PRR model to eLearning in general will be further discussed and illustrated in Section 5. Section 6 concludes the paper.

2. Design features of a Peer Review-Review (PRR) framework

In the development of a theoretical framework it is important to first identify a list of underlying principles that should be used to evaluate its merits and shortcomings. The design principles that we adopt have been inspired by our pragmatic beliefs about course management and student empowerment but we hope that they will appeal to other academic researchers and practitioners alike.

2.1 Implementability

This principle refers to the requirement that any methodology for PRR should be within human and technological capabilities. The implementability principle implicitly implies simplicity or a reliance on well-known concepts. It ensures that the PRR is not dependent on a complicated technological design that may be expensive to maintain; also, the workload for the instructor should remain acceptable, even in educational contexts where the student population is very large, and even when there are many assignments to be reviewed.

2.2 Predictability

What are good peer reviews and how do we grade them? There have been several attempts to define the characteristics of good peer review messages – see, for example, Benos et al. (2003). In essence, the overall rating is derived from a so-called rubric which generally contains a number of criteria that are scored on a pre-defined scale. One of the most frequently cited ones is the Review Quality Instrument (RQI) which contains 8 criteria on a 5-point Likert scale (Van Rooyen et al. 1999).

Some of the RQI items may not be well-suited, however, to evaluate review quality within the context of statistics education. The main reason is that RQI has not been designed with specific learning goals in mind. Therefore, we should only apply rating criteria that can be related to the desired learning outcomes, i.e.: any measurable property of a peer review message that helps predict a student's failure or success in an end-of-course examination.

It is important to emphasize that the predictability principle is not equivalent with any form of real or implied causality. For instance, if the count of submitted feedback messages is found to be "predictive" then this does not imply that submitting more messages "causes" higher exam grades. The true cause of favourable exam scores is most likely related to the learning process in the brain – this however, is a process which cannot be measured directly. Therefore, we employ indirect measures that are believed to be strongly related to the true underlying causes for the purpose of prediction.

Each predictive measure that is used to rank or categorize students falls into one of the following categories:

- Observations without prejudice or bias from the instructor (preferably based on computer log files)
- Intermediate assessments from the instructor (preferably based on a rubric)
- Scores that are reported by students (e.g. self-assessment)

An example of an intermediate assessment is an end of term-paper which is due before the final examination and which is graded by the instructor.

The principle of predictability has the huge advantage that it allows us to focus on those aspects of peer review messages that are known to be favourable in terms of the probability of success. Then again, the predictability principle raises the issue of comparability, which brings us to the next paragraph.

2.3 Comparability

Measured peer review properties cannot be interpreted on a standardized scale and therefore we need to consider an appropriate calibration/benchmarking mechanism which may be achieved through two complementary methods. The first method is to compare the feedback properties of one student to the statistical quantiles of the entire student population of the same course. This is helpful in obtaining grades that are "fair" because quantiles preserve the rank order of students. The second method is to use the estimated threshold parameters that are contained in the predictive model that relates feedback properties to learning outcomes. These parameters provide us with an objective benchmark (see Section 4).

2.4 Purpose

PRR is not a summative grading technique but should serve a clear and formative purpose. This principle means that the peer reviews should be viewed as learning activities in which the grades are unimportant and do not count towards a student's final score. The underlying rationale can be found in the observation that peer grades do not necessarily reflect performance accurately. Clearly, they do not automatically correlate with instructor grades. Secondly, peer grading runs the risk of preventing students from experimenting and being creative.

The "purpose" principle can be satisfied in at least the three following ways. First, since constructivist learning stems from peer review, any assignments that are to be submitted should be easily reviewable. Next, we should ask students to submit peer reviews based on pre-specified rubrics. Even though a rubric is always associated with some sort of scale, the actual grade is not relevant. The rubric is intended to guide students and help them to produce feedback messages which are focused

on those aspects that are known to be relevant. These rubrics may be inspired on RQI and may also depend on the assignment questions. Finally, we need to make sure that students know how they are graded. For example, there is no score for weekly assignment papers because students should be allowed to make mistakes.

In other words, it is the actual composition and subsequent submission of the peer reviews themselves that should be the focus of formative assessment. Moreover, the measurements and grades based on the reviews should be relatable to the relevant cognitive processes of the reviewer – of course, as long as it is the instructor who evaluates the quality of the review messages. On the other hand, the messages that are received may, or may not, have an impact on learning at all. After all, there is no way to ascertain that the receiver, i.e. the student reviewee, actually assimilates the feedback, even though this is implicitly assumed in most studies about formative feedback (Strijbos et al. 2009a).

3. Implementation of the Peer Review-Review (PRR) framework in a statistics course

In this section we will show how the PPR framework has been successfully implemented in an introductory university-level statistics course. After briefly describing the educational setting, we will discuss the various predictive relationships between key components in the course. On the basis of the data that have been collected, we will also identify the relevant variables that play a role in students' course performance and that can thus be used to develop the empirical prediction models to be tested in Section 4.

3.1 Setting

The learning environment of our entry-level statistics course was described in previous research (Wessa 2009a) and primarily consists of a series of weekly workshops (WS1, WS2, etc.) about a variety of topics (Figure 1). Each lecture (L1, L2, etc.) starts with an overview of the frequently made mistakes typical of the workshop held the preceding week while also providing answers and solutions. Students use this information to perform double-blind peer reviews (Rev1, Rev2, etc.) of five workshop submissions from the previous week. The second part of the lecture provides an introduction to the next workshop. Each week, before the start of the lecture, students submit their workshop papers (Compendia) and peer reviews.

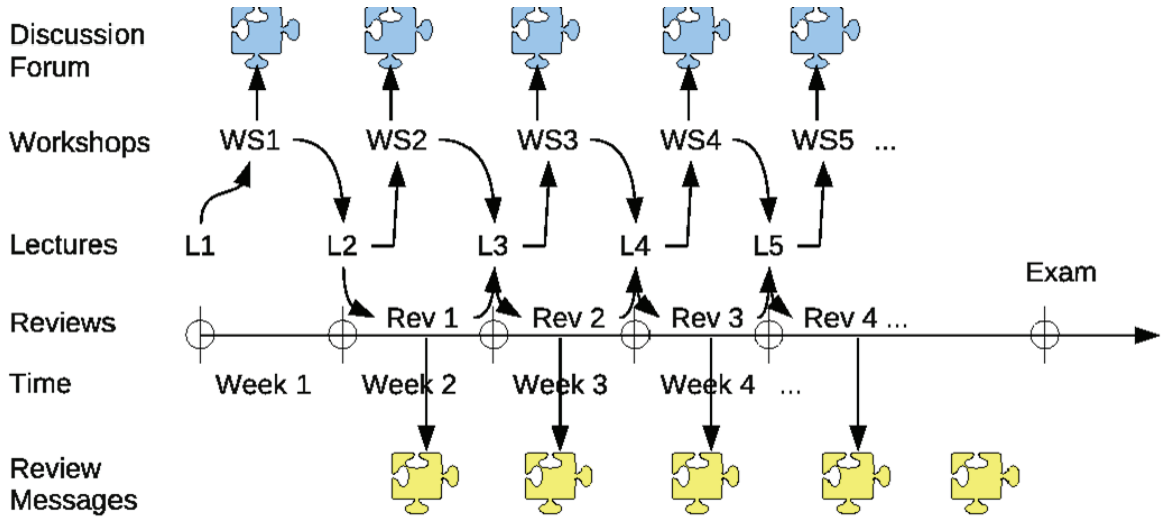

Figure 1: Workshops, lectures and peer reviews

As a result, the maximum number of peer review messages is given by the following equation:

$$\#\text{ Messages} = \#\text{Students} * \#\text{Weeks} * \#\text{Reviews/Week} * \#\text{Criteria/Review}$$

It is obvious that the main problem in this approach is to find a good balance between keeping the educator's workload to an acceptable level and maintaining a strong incentive for students to take the peer reviews seriously. It is also critical that the workshop assignments to be reviewed are doable. As

observed in Section 2.4, this relates to the formative purpose of any PRR model. In our statistics course, this requirement was met by ensuring that all statistical computations are fully reproducible and reusable through the use of the so-called Compendium Platform. Reproducible Computing is an indispensable component in the PRR of any statistics course but more on this in the next section.

3.2 The Peer Review model

Within the context of our statistics course, we have identified four key components that can be connected through predictive relationships. Figure 2 shows that the availability of reproducible computing is a technological prerequisite for the other three components: the peer reviews submitted by the individual students; the term paper based on a collaborative writing effort (Noël & Robert 2004); and the final, summative exam to be taken by the students individually. Note that the exam is based on a series of multiple-choice questions that relate to the computer output that is made available in the form of a Compendium.

As can be seen from Figure 2, the peer reviews should have predictive power for both the term paper scores and the final examination scores. The underlying assumption is that constructivist-learning activities of the type described here lead to more effective forms of non-rote learning (Wessa 2009a) – a claim that should be evidenced by favourable term paper and exam scores.

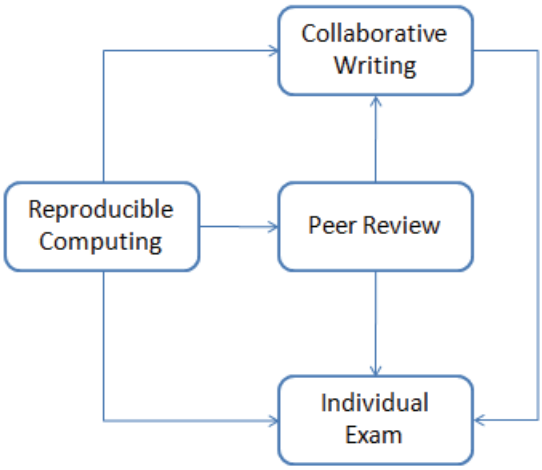

Figure 2: Predictive relationships

The term paper should be predictive of the final exam because it consists of a written report that summarizes the knowledge interactively constructed by students over the entire course. The paper is completed at the end of the term and handed in before the final exam. Obviously we expect that the prediction of final exam scores is more difficult than of the term paper scores. It should also be noted that the estimated prediction models will involve a great number of potentially useful variables. The selection of such variables is an empirical matter and relies entirely on the statistical methodology that is used to build the prediction models.

3.3 Data

Over the past three years we have collected a substantial amount of data from the statistics courses that made use of the Compendium Platform and peer review facilities. In this section we will briefly describe the variables that we have used in the empirical models to be examined in Section 4.

Table 1 shows the number students, grouped by year, gender, and type of prior education. The “Ba” subgroup corresponds to full-time students of the second Bachelor year. They have a one-year “academic” background with a sufficiently adequate knowledge of mathematics and basic statistics. The “Prep” students consist of those who have obtained a three-year professional Bachelor degree. These students can only enter the Master in Business Studies programme after completing a preparatory year. Prep-students are generally more mature and better motivated – on the other hand, they usually have a weaker mathematical background than students from the “Ba” group.

Table 1: Population

	Ba		Prep.		Total
Year	Female	Male	Female	Male	
0	58	51	53	78	240
1	49	66	51	85	251
2	52	79	69	85	285

R command:source

("http://www.freestatistics.org/blog/index.php?v=date/2010/May/25/t1274792028qzkzutois62ofld.htm&rcode=T")

Table 2 gives an overview of the variables that have been found to be relevant in predictive terms within the context of the empirical analysis. The statistical methodology that we employed, allowed us to discriminate between important and unimportant variables. The complete list of all candidate variables is not discussed and beyond the scope of this paper.

Table 2: Variables

Name	Description
Gender	0: Females 1: Males
Pop	0: Bachelor 1: Prep. Programme
Year	0: Fall 2007 1: Fall 2008 2: Fall 2009
NNZFG	# Submitted Feedback Messages
AFL	Average Feedback Length (average number of characters for all (non empty) feedback messages)
LPM	Levenshtein Distance (average *difference* between the feedback messages that is submitted over all distinct papers that were subject to review by the student)
BC	# Blogged Computations (umber of reproducible computations that the student produced)
WORDSPA	# Words per Author (word count of the submitted Term Paper divided by the number of authors)
PSCORE	Score of Term Paper

4. Empirical evidence for the predictive capabilities of the Peer Review model

The statistical model that is used to generate predictions is called “Pruned Classification and Regression Tree” (henceforth PCRT) and attempts to build logical “if-then-else” rules that can help predict the scores of the term paper and the summative exam results. The PCRT requires the investigator to identify the relevant categories of the dependent variable that we should be able to predict. In this study we are primarily interested in a binary classification with two categories (Pass and Fail). For each student there are two binary, dependent variables. Of course, it is possible that a student passes one test and fails the other.

The PCRT is built by the use of the J48 algorithm as implemented in the RWeka package (Hornik et al. 2009; Witten et al. 2005). The algorithm uses an iterative approach to select the most important variables that allow us to predict whether a student belongs to the Pass or Fail category. Each selected variable is represented in one or several “if-then-else” nodes that may be connected to other decision nodes in a hierarchical tree. The end-nodes of each branch represent a prediction value (Pass/Fail).

The use of the final exam scores deserves a word of caution because, as has been often observed, exam or test scores have unexpected characteristics with respect to their validity to test student’s analytical skills. For example, some questions may be poorly understood by students because of unusual wording or grammar. Another example is when a question relates to concepts that are poorly treated in the course or accompanying course materials. In both cases, it is highly likely that students will not be able to find the right answer, even if they have acquired a deep understanding of most of the statistical concepts. The use of such questions introduces noise in the prediction models and should thus be avoided. We applied a statistical model to determine the optimal weights of the

individual questions of the final examination in order to obtain a total exam score which can be shown to be rationally predictable by the independent variables in the model (Wessa 2009b). The optimal exam score transformations are applied before using the J48 algorithm to build the PCRT.

4.1 Prediction models

In the first model we predict the outcome of the term paper based on purely objective information that is available *before* the actual paper is submitted (Figure 3).

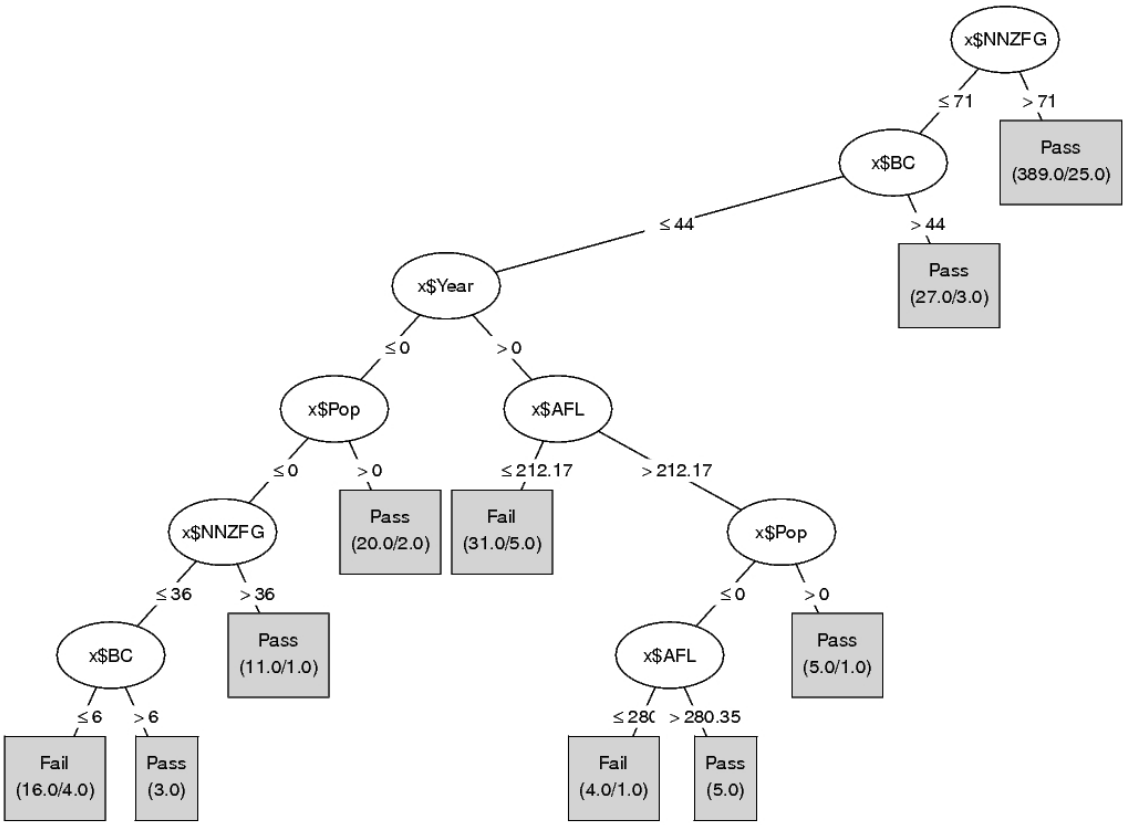

Figure 3: PCRT (prediction of PSCORE without term paper properties) R command: source ("http://www.freestatistics.org/blog/index.php?v=date/2010/May/25/t1274792047ryasw6b110c9xey.htm&rcode=T")

The most important predictive variable is the number of submitted feedback messages (prediction rule: IF NNZFG > 71 THEN PASS). There are three activity-based variables (NNZFG, BC and AFL) in the PCRT which implies that these variables must be somehow related to the learning process of writing a term paper.

In the second model we predict the outcome of the term paper based on objective information that is available *after* the actual paper is submitted (Figure 4).

It is perhaps surprising that the word count per author is the only activity-based variable in the PCRT. Writing more words seems to be a good predictor of how an educator will perceive the quality of the paper's content. One may wonder if this observation is caused by laziness on the part of the educator or if quantity as measured in words really indicates quality of content. A simple test to determine which of these is true is to examine whether the variable PSCORE, i.e. the score for the term paper, is contained in the PCRT that predicts final exam scores (Figure 5).

It is clear that the variable PSCORE is an important predictor, and hence, that the term paper scores are likely to reflect quality of content and the student's grasp of the underlying statistical concepts. The other variables of importance are related to quantitative properties of computations and feedback messages: BC, NNZFG, and AFL.

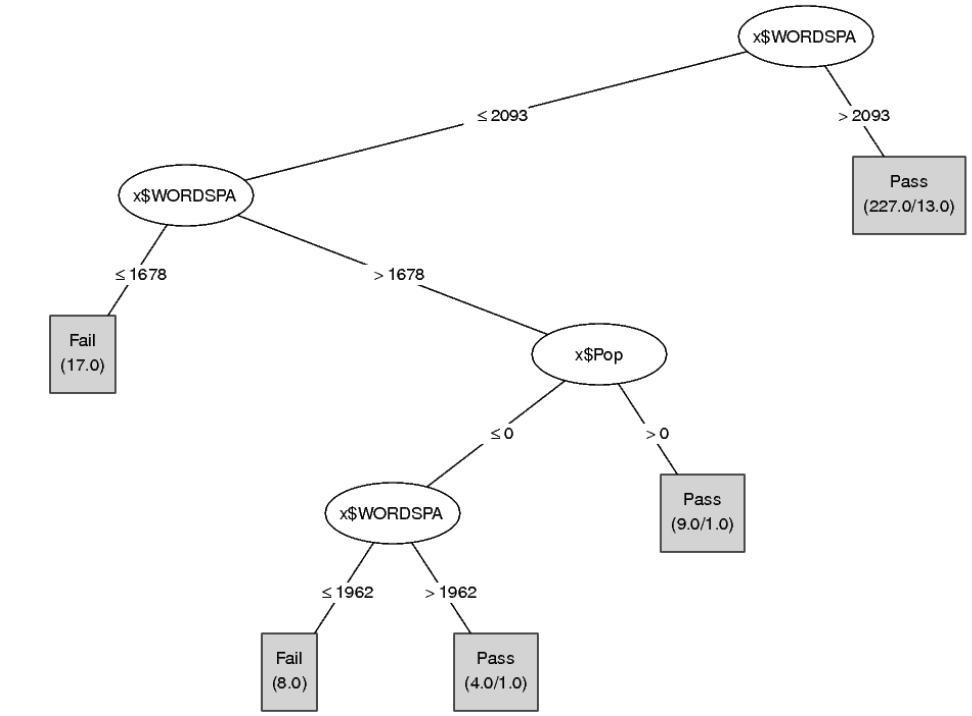

Figure 4: PCRT (prediction of PSCORE with paper properties) R command:source ("http://www.freestatistics.org/blog/index.php?v=date/2010/May/25/t1274792072iv1ziqxftwtzpso.htm&rcode=T")

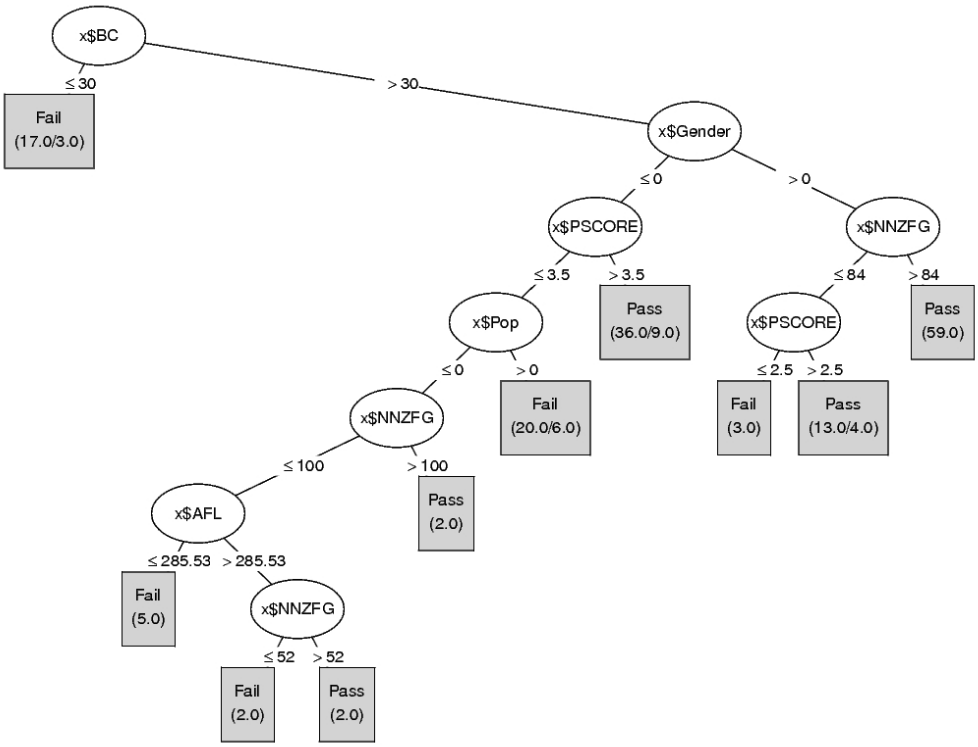

Figure 5: PCRT (prediction of optimally weighted exam scores) *R command:* source ("http://www.freestatistics.org/blog/index.php?v=date/2010/May/25/t1274792092n28z7yyfkimoye4.htm&rcode=T")

Even though there may be important differences between the various subgroups, it is remarkable that final exam scores can be predicted by the use of purely quantitative measures. The most likely explanation is that PRR motivates a student to take the reviewing process seriously, which can be taken to imply that feedback messages are only submitted after careful consideration. In this sense it is very plausible that "quantity" is an indication of "quality" of feedback. Of course, the same argument may apply to the term papers – students know that the term paper is graded by the educator; therefore they will try to include relevant and high-quality information only.

Note that there is one variable in Table 2 that does not appear in any of the above PCRTs. It is the average Levenshtein Distance (LPM), which measures the degree of variation in the feedback messages that are submitted by a student. The variation is likely to be high when the student provides feedback that is specific for the Compendium under review. If the variable PSCORE is excluded from the last PCRT model, then the variable LPM seems to have relevance – note that the tree is not shown. In other words, in statistics courses in which students are not required to write a term paper, LPM is likely to play an important role.

4.2 Prediction performance

Table 3 shows some summary statistics of the three PCRT models discussed above. The percentage of correctly classified students is very high, especially when compared to other predictive models reported in the recent literature (Wessa 2009b). For each model the statistics were computed for the entire sample (interpolation) and for a so-called 10-fold cross validation test that can be interpreted as an out-of-sample (extrapolation) performance.

Table 3: Prediction quality of PCRTs

Prediction of:	PSCORE				Exam Score			
	Without paper		With paper		Without PSCORE		With PSCORE	
	Entire Sample	10-fold CV	Entire Sample	10-fold CV	Entire Sample	10-fold CV	Entire Sample	10-fold CV
Correctly Classified Students	91.78%	86.30%	94.34%	88.30%	74.79%	68.49%	84.03%	78.99%
Kappa	0.6162	0.3844	0.7389	0.5291	0.2914	0.1608	0.6222	0.5028
RMSE	0.2709	0.3472	0.229	0.3306	0.4333	0.4700	0.3479	0.4105

The so-called Kappa statistic is a measure that takes into account the probability that a classification is correct by chance. Negative values indicate a disagreement between predicted and observed classification values. Positive Kappa values may range from 0 to 1.

The analysis allows us to draw two important conclusions. First, there is strong empirical evidence for all of the predictive relationships that were identified in Figure 2. Secondly, within the framework of our self-imposed principles (see Section 2), we are able to implement a set of *decision rules* that are based on predictable relationships that allow us to compare students during all stages of the PRR process.

5. Significance of the Peer Review-Review model to eLearning

In this section we illustrate how a rule-based approach to PRR can be implemented. We selected a few real-life cases that represent different PRR grades and for each of them we stored accompanying screenshots of our peer review software on the Compendium Platform server (the hyperlinks are included). The various categories of students require different remedial and other interventions on the part of the educator. It is our contention that what holds true for the statistics course discussed here is also significant to other courses that adopt a similar PRR framework within a rich eLearning environment.

5.1 Lowest PRR grade

The first category is made up of students that did not submit a large number of feedback messages (example: http://www.freestatistics.org/ICEL2010/screen1.jpg; NNZFG = 25). As discussed in Section 4, these students are highly unlikely to write a good term paper (Figure 3) or to pass the final exam (Figure 5). Our PRR is primarily useful in identifying these vulnerable cases early on and to help us focus on finding out the reason for the low performance. It is clear that the student in question did not actively participate in the course activities (8 out of 11 assignments were not submitted; the number of

reviewed assignments is 2 and only 5 Compendia were reviewed for them). The software will alert educators on time and help them take appropriate action to avoid dropouts and course failures. Note that educators may also attribute a penalty for not submitting assignment papers.

5.2 Low PRR grade

This category includes students with a sufficiently high number of feedback messages but an otherwise poor track record of reviews (example: http://www.freestatistics.org/ICEL2010/screen2.jpg; NNZFG = 114, AFL = 167 and LPM = 86). They are not likely to have submitted feedback messages of sufficiently high quality. As can be seen in Figures 3 and 5, an appropriate AFL level is 280 characters; the student in question here only manages 167. In addition, his or her Levenshtein Distance is relatively low, which suggests that the feedback messages are very similar to each other.

The peer review software selects a sample of the feedback messages produced over the course of the workshops. Note that the first three workshops are not included as we allow for a "learning effect" in review quality and that also the last workshop is left out because it corresponds to the term paper. For each student, and for every selected workshop, the peer review software shows only two sets of feedback messages, i.e. the longest and shortest ones, and automatically selects the longest feedback messages from other students who were required to review the same author (example: http://www.freestatistics.org/ICEL2010/screen3.jpg). This particular approach allows the educator to compare the actual content of the messages and assess the following aspects:

- Did the student provide relevant information, given the pre-specified question in the rubric?
- Did the student neglect to provide important and/or specific details that were mentioned by the other student (which is shown for comparison purposes)?
- Is the feedback constructive and helpful for the author?

As the third screenshot shows, the student mentions that the author has made "many errors because of badly chosen parameters". The other student, however, explicitly lists the various problems with the statistical techniques that have been used (VRM, spectral analysis, etc.).

Careful reading of selected feedback messages reveals that this student provided low-quality reviews only, as predicted by the statistics. In this case, the educator's job is to verify the predicted quality of the feedback, without the need to read too many reviews. However, it should also be noted that we strongly advise against a purely automated grading mechanism which excludes human intervention. A rule-based approach is intended to increase our grading efficiency; it is not a replacement for the educator's judgement.

5.3 Medium PRR grade

Some students may have a sufficiently high number of feedback messages while other properties (AFL, LPM) are about average. Typically, these students have submitted feedback messages that have adequate content but lack detail. Sometimes these peer reviews are too short to be meaningful, however. The sample screenshot (http://www.freestatistics.org/ICEL2010/screen4.jpg) shows two messages: one has an average length (482 characters) and the second one is very short (45 characters). Analysis of a fair sample of these messages confirmed the predicted medium-quality of the feedback. The educator's grade is documented through a rubric that lists the quality-related properties found in the messages (e.g. relevance, completeness, helpfulness, etc.).

5.4 High PRR grade

A relatively small number of students show extremely favourable review statistics (example: http://www.freestatistics.org/ICEL2010/screen5.jpg). It is generally sufficient for educators to read a very small sample of feedback messages to confirm that these students performed very well, resulting in their high grades on all relevant variables.

6. In conclusion

As Henze (2008) puts it, teaching embraces a great deal more than telling, instructing and treating the learner as "an empty vessel to be downloaded with knowledge". This kind of *cognitive dumping* is not likely to provide learners with the right skills mix to organize and conceptualize vast amounts of complex information in any useful way. The alternative, as he argues, lies in acknowledging that

learning involves individual as well as social processes: thinking but also assimilating, accepting and building on past knowledge and shared experience. It is this general approach that Henze (2008) refers to as "constructivist learning".

More than six years ago we started to sketch out plans for an approach to university-level statistics education that would incorporate constructivist-learning principles and would harness the right innovative technology to make it happen. We were told that it would be impossible to achieve this within the context of a large student population. Experts suggested, however, that we should introduce constructivist activities in small groups within a computer lab setting. The present paper shows that this approach need not be expensive or unrealistic. As the PCRT analysis of our peer review and student assessment data has shown, it is perfectly feasible to support constructivist learning efficiently and within an empirically verified, theoretically sound framework. The only two requirements are, first, the availability of free, innovative software that supports reproducible computing and peer review, and secondly, a formative Peer Review-Review model based on rules that can be shown to predict objectively measured learning outcomes.

Acknowledgements

This project was funded by the Catholic University of Leuven Association (Research Grant OOF2007/13).

References

Benos, D.J., Kirk, K.I. and Hall, J.E. (2003), "How to Review a Paper", *Advances in Physiology Education*, Vol 27, pp 47-52, doi:10.1152/advan.00057.2002.

Dabbagh, N. (2005) "Pedagogical Models for ELearning: A Theory-Based Design Framework", *International Journal of Technology in Teaching and Learning*, Vol 1, No 1, pp 25-44.

Henze, M. (2008) "Demystifying 'constructivism': Teasing unnecessary baggage from useful pedagogy", [online] (cited 27 May 2010) Available from <URL:http://www.markhenze.com/pdf%20files/Demytifying%20Constructivism.pdf>

Hornik, K., Buchta, Chr. and Zeileis, A. (2009) "Open-Source Machine Learning: R Meets Weka", *Computational Statistics*, Vol 24, No 2, pp 225-232.

Lundstrom, K. and Baker, W. (2009) "To Give is Better than to Receive: The Benefits of Peer Review to the Reviewer's Own Writing", *Journal of Second Language Writing*, Vol 18, pp 30-43.

Noël, S. and Robert, J.-M. (2004) "Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like?", *Computer Supported Cooperative Work*, Vol 13, No 1, pp 63-89.

Strijbos, J.-W., Narciss, S. and Dünnebier, K. (2009a) "Peer Feedback and Sender's Competence Level in Academic Writing Revision Tasks: Are they Critical for Feedback Perceptions and Efficiency?", *Learning and Instruction*, Vol 20, No 4, pp 291-303.

Strijbos, J.-W. and Sluijsmans, D. (2009b) "Unravelling Peer Assessment: Methodological, Functional, and Conceptual Developments", *Learning and Instruction*, Vol 20, No 4, pp 265-269.

Van Gennip, N.A.E., Segers, M.S.R. and Tillema, H.H. (2009) "Peer Assessment for Learning from a Social Perspective: The Influence of Interpersonal Variables and Structural Features", *Educational Research Review*, Vol 4, pp 41-54.

Van Rooyen, S., Black, N. and Godlee, F. (1999) "Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts", *Journal of Clinical Epidemiology*, Vol 52, No 7, pp 625-629.

Wessa, P. (2009a) "How Reproducible Computing Leads to Non-Rote Learning within Socially Constructivist Statistics Education", *Electronic Journal of eLearning*, Vol 7, No 2, pp 173-182.

Wessa, P. (2009b) "Quality Control of Statistical Learning Environments and Prediction of Learning Outcomes through Reproducible Computing", *International Journal of Computers, Communications & Control*, Vol 4, No 2, pp 185-197.

Wessa, P. (2009c) "Reproducible Computing: A New Technology for Statistics Education and Educational Research" *in* Rieger, B., Amouzegar, M.A., and Ao, S.L. (eds.), *IAENG Transactions on Engineering Technologies*, American Institute of Physics, pp 86-97.

Witten, I.H. and Eibe, F. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco.