The Impact of Self-Assessment and Reflection on Student Learning Outcomes

Iouri Belski
Royal Melbourne Institute of Technology, Melbourne, Australia
iouri.belski@rmit.edu.au

Abstract*: It has been found that the Task Evaluation and Reflection Instrument for Student Self-Assessment (TERISSA) has helped many educators to improve student satisfaction with educational feedback by engaging students in self-assessment and reflection. This paper evaluated the impact of TERISSA, used in formative assessment on student learning outcomes. It presents the results of an experiment conducted in semester 1 of 2009, where those students engaged in self-assessment and reflection with TERISSA achieved better grades than students who did not use TERISSA.*

Keywords – self-assessment, reflection, educational feedback, learning outcomes, student satisfaction.

Introduction

It is a well known that students can produce valid educational feedback both for themselves and for their peers (Heron, 1981). This can be achieved by engaging students in self-assessment and/or peer-assessment; and this has been demonstrated as a valid approach to self-improvement in engineering education (Boud & Holmes, 1981). The outcomes of student self- and peer-assessment can be manifold. It has been shown by a number of authors that well implemented self- and peer-assessment can notably improve educational feedback and simultaneously save significant time and effort for educators. Self- and peer-assessment can also result in better engagement and improved students' course performance. The following are just a few recent accounts of effective implementation of self- and peer-assessment.

'Traditional' Implementation of Self- and Peer-assessment

McDonald and Boud reported on teachers effectively developing students' self-assessment skills (McDonald & Boud, 2003). O'Shea and Bigdan engaged students in peer-assessment by devising an academic version of the Biggest Loser competition (O'Shea & Bigdan, 2008). Belski, Harlim and De Silva reported on statistically significant improvements in teaching quality indicators for numerous units that engaged students in self-assessment and reflection by means of the Task Evaluation and Reflection Instrument for Student Self-Assessment (TERISSA) (Belski, 2007, 2009b; Harlim, De Silva, & Belski, 2009).

Implementation of Self- and Peer-assessment Using IT/web Resources

Most of the recent studies actively used IT/web resources to engage students in self- and peer-assessment. Turns reported on the effectiveness of the *Reflective Learner* web-based environment in enhancing the traditional practice of self-assessment, by writing 'learning essays' on learning experiences (Turns, 1997). Smith and Kampf achieved effective peer-assessment by supporting informal cooperative student learning groups using the *WebCT* as a peer-review system (Smith & Kampf, 2004). McGourty reported on the effectiveness of the *Team Developer* computer-based survey system in engaging students in generating multisource feedback (McGourty, 2000). Willey and Gardner used the web-based tool *SPARK*^{PLUS} to facilitate learning orientated peer-assessment (Willey & Gardner, 2009). Kay, Li and Fekete proposed an exciting two-stage process of self-assessment and reflection (Kay, Li, & Fekete, 2007) that engaged students in reflection-on-action, and in reflection-in-action as suggested by Schön (Schön, 1987). There are also many reports on successful usage of Wikis and ePortfolios for self- and peer-assessment (Fielke & Quinn, 2009; Molyneaux, Brumley, Li, Gravina, & Ward, 2009).

Impact of Self- and Peer-assessment on Student Learning Outcomes

Although many authors have published on the successes of different strategies of self- and peer-assessment in providing students with valuable individual feedback on their learning, there is little empirical data on the impact of self-and peer-assessment onto the outcomes of individual learning. Interestingly, the impact of self-and peer-assessment on student grades was only reported by the authors who deployed the 'traditional' approach.

In the McDonald and Boud study, high school teachers were trained in how to develop student self-assessment skills. These teachers formally trained 256 students in self-assessment. End of the year examinations showed that this group of 256 students trained in self-assessment outperformed the 'control group' of 259 students in all areas of academic studies (McDonald & Boud, 2003).

Success of peer-assessment in impacting student learning has been reported by O'Shea and Bigdan while teaching a large first year Electrical Engineering class (O'Shea & Bigdan, 2008). The authors found statistically significant improvement in the mean grade of the 146 students who participated in the Biggest Loser initiative compared to the control cohort of 199 students.

Harlim, De Silva and Belski reported considerable difference in student examination performance due to their usage of TERISSA (Harlim, et al., 2009). The final exam results of 171 students who used TERISSA in three tutorial tests (summative) were compared against exam results of 256 who had never used TERISSA. The average exam marks for the former were 50.81 (SE=1.27) and 47.08 (SE=1.10) for the latter. An Independent Sample t-test showed this difference as statistically significant ($p=0.015$).

This paper analyses the impact of self-assessment and reflection using TERISSA in formative assessment on student academic performance. It presents the outcomes of the experiment conducted by the author in his third year class on Electronic Engineering in the first semester of 2009.

The TERISSA procedure

TERISSA provides a procedure for students to follow while resolving problems, conducting project work and preparing assignments. It requires students to conduct two evaluations of a tasks complexity on a Likert scale from 1 to 5 (Belski, 2009b). These evaluations are undertaken when the task is first presented and after the task has been resolved. Students then reflect on each of these evaluations and on the reasons for any discrepancy between them. TERISSA also requires students to devise and plan the actions needed to improve learning outcomes. Once the use of TERISSA has been perfected, it normally requires around five minutes to conduct. The following is a general procedure for a student using TERISSA while involved in active learning:

Step 1. (*To be conducted before you start work*)
Evaluate and record the complexity of the question, problem, assignment, etc. using the following scale: 1-very simple; 2-simple; 3-so-so; 4-difficult; 5-very difficult.
Give reasons (in writing) why you have not evaluated it as one level **less** difficult.

Step 2. (*To be conducted after the work has been concluded*)
Evaluate and record the complexity of the question, problem, assignment, etc. once again using the scale from Step 1.
Reflect (in writing) why you have not evaluated the question as one level **more** difficult this time.
Also reflect on the reasons for any discrepancy between the original (Step 1) and the final (Step 2) evaluation. Decide on the actions you need to undertake to become more confident with a similar task next time and write them down.

Over the past nine years, TERISSA has been successfully used by many academics in tutorial classes, home and class activities, individual and group exercises, various home assignments and practical laboratory work and resulted in notable improvement in student opinions on educational feedback

(Belski, 2007, 2009b). The results of the 2007 TERISSA trial at RMIT have shown that TERISSA can be effectively used by educators from different disciplines in both formative and summative assessment. Moreover, student opinions recorded in 2008 have further supported its effectiveness and have shown that its use can result in sustainable and significant improvements in student satisfaction with educational feedback (Belski, 2007, 2009b). The study of 2009 showed that significant improvement in student satisfaction can also be achieved in large classes (Harlim, et al., 2009).

The TERISSA experiment: semester 1 of 2009

Experimental setup

In semester 1 of 2009 (12 teaching weeks from March to May) the author coordinated the third year unit on Electronic Engineering with 63 students enrolled. The unit's weekly activities consisted of two one-hour lectures, one two-hour tutorial and one two-hour laboratory session. All lectures and tutorial classes were conducted by the author. Identical problems were resolved by all tutorial groups in the same week. All laboratory sessions have been supervised by the same laboratory tutor. The unit assessment consisted of two tutorial tests (week 7 and week 11), two home assignments, four practical laboratories and the final examination. During the first tutorial, which was conducted in week 1, students sat a 20 minutes test. This test was intended to evaluate their prior knowledge of electronics basics. The mark for this test did not influence the total mark.

Two weeks before the start of semester the author randomly, and without any knowledge of tutorial enrolments, selected one of the two tutorial groups as 'experimental'. The experimental group used TERISSA in their tutorials from week one (immediately after the introductory test). Students of the other group did not use TERISSA until week seven; hence this group played the role of the 'control group'. In week one the author discovered that the experimental group had a total of 24 students enrolled and the 'control group' – 39.

During the first six weeks of the semester students in the experimental group were using TERISSA during all tutorial sessions. Every time students were presented with a problem to resolve, they were asked to evaluate the complexity of the problem, as outlined in the general TERISSA procedure, and to record the complexity score together with their reflections on their reasons behind this score. In addition, they were asked to raise their hands, indicating the complexity score they recorded, when a tutor called out the appropriate score. One or two students were involved in counting hands of the 'voters'. The voting results were recorded on a tablet PC, which was used by the author to conduct tutorials in both groups. The actual recordings of both tutorials were printed in pdf as one file and were available to students a day or two after tutorial sessions. The actual recordings of the problem considered during tutorials in week 6 are depicted in Figure 1.

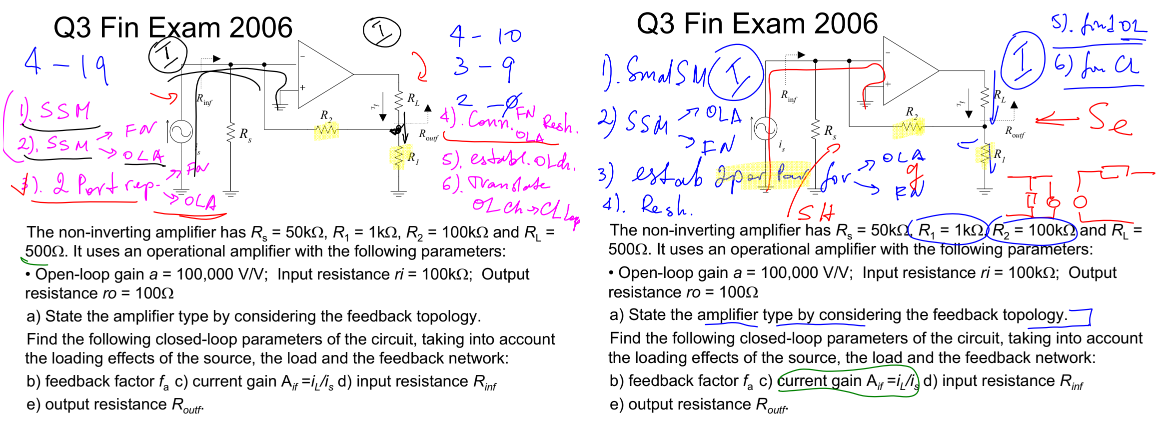

Figure 1: The recordings of the tutorial problem in week 6: experimental group (left), control group (right).

Uploading a single pdf file that contained both weekly tutorials to RMIT Learning Hub ensured that the students of both tutorial groups were able to analyse the solutions derived in both tutorials. As presented in Figure 1, the recordings of the experimental group contained the evaluation of the complexity of the problem before and after it was resolved. All of the 19 students, who attended the experimental group tutorial in week 6, evaluated the problem as difficult (4) before (the score of 4 – 19 recorded in upper left corner of the left slide in Figure 1). After the problem was resolved, 10 students still considered it as difficult (4) and 9 students downgraded their complexity opinion to 'so-so' (3) (the scores of 4 – 10 and 3 – 9 recorded in upper right corner of the left slide in Figure 1).

Experimental results

Table 1 presents a comparison in performance between students from the experimental group (TERISSA Group) and the control group in tutorial tests and final examination. All the tests' marks were out of 10; the examination mark was out of 100. The mean value (Mean), the standard error (SE) and the *p* value of statistical significance between the groups have been evaluated by means of the Independent Samples t-test.

TABLE 1: Performance of Experimental and Control groups in tutorial tests and final examination (Independent Samples t-test)

	Test Week 1		Test Week 7		Test Week 11		Final Exam	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
TERISSA Group	4.6	0.36	**6.1**	**0.66**	7.2	0.61	52.3	4.1
Control Group	4.5	0.21	**4.1**	**0.44**	7	0.39	52.6	3.8
	p = 0.475		**p = 0.007**		p = 0.385		p = 0.478	

The mean values for both the experimental and control groups in the test of week 1 were very close. This implies that the students' prior knowledge in electronics basics were similar. Thus, if the students of these two groups were involved in identical unit experience, it would be normal to expect more or less similar performance of both groups in all forms of summative assessment.

Despite such expectations, groups' results for the tutorial test in week 7 were vastly different (6.1/10 vs. 4.1/10). Moreover, this difference is statistically significant (p = 0.007). This fact indicates that unit learning experiences of two groups in the first six weeks of a semester were vastly different. What were the differences in the learning experiences of these two groups? Both groups had the same lecturer, tutor and laboratory demonstrator. They attended lectures together, resolved the same tutorial problems and performed the same laboratory experiments. The only real difference they experienced was the application, or lack there of, TERISSA during six tutorial classes. Students of the experimental group were engaged in self-assessment and reflection during tutorials for six weeks; students of the control group were not using TERISSA throughout. This leads to a conclusion, that *the notable difference in students' test 1 performance can only be attributed to self-assessment and reflection that the students of the experimental group were engaged in during the first six weeks of the semester*.

The TERISSA procedure was introduced to the control group directly after the test in week 7 and was used by both tutorial groups for the rest of the semester. The results of the tutorial test in week 11, presented in Table 1, show that the difference in groups' performance became smaller (and statistically insignificant). The outcomes of the final examination, which was held four weeks after the second test, were also very similar for both groups and have not shown any statistical difference in the groups' performance.

Discussion & Conclusion

Through analysing the results of the tutorial test held in week 7, it has been concluded that the difference in the groups' performance was due to the usage of TERISSA by the experimental group. In order to comprehend the impact of TERISSA on student learning, the following two questions need to be answered: i) How could TERISSA make such profound impact on student learning outcomes in just six weeks? ii) Why the difference in groups' performance registered in week 7 had almost entirely disappeared by the end of the semester?

The data collected during the TERISSA trial in semester 2 of 2007, suggested that the reason for TERISSA's success relates to the fact that the majority of students recorded a significant discrepancy between the difficulty levels in pre- and post-solution problem evaluations and reflected on their learning needs meaningfully (Belski, 2009a). Table 2 presents the discrepancies between the original and final evaluations of task difficulty levels recorded for three tasks (denoted as T1, T2 and T3) in student home assignments presented in (Belski, 2009a).

TABLE 2. Discrepancies in student evaluation of task complexity: home assignments
(source: Belski, 2009a)

Discrepancy (original – final)	T1	T2	T3
Under-evaluated [-2]	**5.6%**	**0.0%**	**20.0%**
-1	**38.9%**	**14.3%**	**13.3%**
No difference [0]	5.6%	28.6%	33.3%
1	**33.3%**	**42.9%**	**33.3%**
Over-evaluated [2]	**16.7%**	**14.3%**	**0.0%**

Only a minority of students undertaking these three problems were correct in their judgment of the problem's complexity before it had been resolved. In the case of task 1 (T1 in Table 2), for example, only 5.6% of students did not change their opinion of the complexity score; whilst over 94% recorded discrepancies. Moreover, 22.3% of the students made vast errors in their judgment of complexity by over- or under-estimating its difficulty level by 2 units of difficulty.

As suggested previously (Belski, 2009a), the fact that students' pre- and post-solution evaluation of task complexity differ, is fundamentally important for the success of TERISSA in engaging students in reflecting on their learning. Students usually do not expect any discrepancy in their judgments. When they apply TERISSA and discover that they were unable to evaluate the degree of difficulty accurately, they experience what Dewey referred to as "a state of doubt, hesitation, perplexity, mental difficulty, in which thinking originates" (Dewey, 1933). They naturally want to explain to themselves the reasons for the inaccuracy and are involved in "an act of searching, hunting, inquiring, to find material that will resolve the doubt, settle and dispose of the perplexity" (Dewey, 1933). In other words, students become engaged in reflection and, as a result, provide valuable feedback on their learning to themselves.

The above-mentioned reasons for student engagement in self-assessment and reflection using TERISSA are further supported by student opinions. Over 54% of the students from the 2009 class on Electronic Engineering either agreed or strongly agreed with the statement in the formal RMIT Course Experience Survey (CES) "*I will use the TERISSA's approach for my individual learning in the future*". Such positive student response suggests that many of the students enrolled in the 2009 class on Electronic Engineering had realised that TERISSA helped them in their learning.

The effectiveness of TERISSA in self-assessment is also supported by the opinions of 132 students enrolled in six courses coordinated by different academics in semester 2 of 2007. The latter cohort responded positively to the survey statement: "*TERISSA has helped me to identify the learning area which required my immediate consideration*" (205 students participated in the survey) (Belski, 2009b).

Thus, it can be concluded, that students of the experimental group, were able to identify their subject weaknesses in the first six weeks of the semester much earlier than the students of the control group.

Moreover, some of the students of the experimental group acted on their TERISSA reflections and addressed their weakness areas promptly. Students of the control group did not reflect during tutorials, so it is unlikely that they were able to act on their learning weaknesses as effectively as their experimental group colleagues. This resulted in significantly better performance of the experimental group in test 1. Once all students deployed TERISSA from week 7, the groups' performance gradually settled at the week 1 level.

Overall it is clear that the application of TERISSA in student tutorials not only increased the academic performance of participating students, but also benefited in the self-assessment ability and reflection skills of these students for further academic challenges.

References

Belski, I. (2007, 9-13 December). *Using Task Evaluation and Reflection Instrument for Student Self-Assessment (TERISSA) to Improve Educational Assessment and Feedback.* Paper presented at the Proceedings of the 18th Conference of the Australasian Association of Engineering Education (AaeE), Melbourne, Australia.

Belski, I. (2009a). *Improving Student Satisfaction with Feedback by Engaging Them in Self-Assessment and Reflection*. Paper presented at the ATN Assessment Conference.

Belski, I. (2009b). *TERISSA: 'Plug and Play' to Improve Assessment and Feedback.* Paper presented at the 20th Annual Conference for the Australasian Association for Engineering Education, Adelaide.

Boud, D., & Holmes, W. H. (1981). Self and peer marking in an undergraduate engineering course. *IEEE Transactions on Education, E-24*(4), 267-274.

Dewey, J. (1933). *How We Think*. Boston: D C Health.

Fielke, J. M., & Quinn, D. (2009). *Supporting self-assessment using e-portfolios*. Paper presented at the The 20th Annual Conference for the Australasian Association for Engineering Education.

Harlim, J., De Silva, A. D., & Belski, I. (2009). *Improving Feedback in Large Classes: Application of Task Evaluation and Reflection Instrument for Student Self-Assessment (TERISSA) in a Unit on Business Statistics.* Paper presented at the ATN Assessment Conference.

Heron, J. (1981). Assessment revisited. In D. Boud (Ed.), *Developing student autonomy in learning* (pp. 55-70). London: Kogan Page.

Kay, J., Li, L., & Fekete, A. (2007). *Learner Reflection in Student Self-assessment*. Paper presented at the Ninth Australasian Computing Education Conference (ACE2007).

McDonald, B., & Boud, D. (2003). The Impact of Self-assessment on Achievement: the effect of self-assessment training on performance in external examinations. *Assessment in Education, 10*(2), 209-220.

McGourty, J. (2000). Using multisource feedback in the classroom: a computer-based approach. *IEEE Transactions on Education, 43*(2), 120-124.

Molyneaux, T., Brumley, J., Li, J., Gravina, R., & Ward, L. (2009). *A comparison of the use of Wikis and ePortfolios to facilitate group project work*. Paper presented at the The 20th Annual Conference for the Australasian Association for Engineering Education.

O'Shea, P., & Bigdan, V. (2008). The Biggest Loser Competition. *IEEE Transactions on Education, 51*(1), 123-130.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass Publishers.

Smith, K., & Kampf, C. (2004). *Developing writing assignments and feedback strategies for maximum effectiveness in large classroom environments.* Paper presented at the Proceedings of the International Professional Communication Conference (IPCC)

Turns, J. (1997). *Learning essays and the reflective learner: supporting assessment in engineering design education.* Paper presented at the Proceedings of the 27th Annual Frontiers in Education Conference.

Willey, K., & Gardner, A. (2009). *Investigating the potential of self and peer assessment to develop learning oriented assessment tasks*. Paper presented at the The 20th Annual Conference for the Australasian Association for Engineering Education.