MOKSLAS – LIETUVOS ATEITIS **2010, 2 tomas, Nr. 2**
SCIENCE – FUTURE OF LITHUANIA 2010, Vol. 2, No 2

Verslas XXI amžiuje
Business in XXI Century

ANALYSIS OF STUDENT KNOWLEDGE EVALUATION APPLYING SELF-ASSESSMENT METHODOLOGY: CRITERIA, PROBLEMS AND RESULTS

Agnė Matuliauskaitė[1], Edmundas Žvirblis[2]
Vilnius Gediminas Technical University
E-mail: [1]agne.matuliauskaite@vgtu.lt, [2]e.zvirblis@ivpk.lt

Abstract. The article analyses research done by a number of authors on problems related to knowledge evaluation based on self-assessment. Self-assessment problems, self-assessment criteria, self-assessment methods, and integration of self-assessment data into the final results are considered in the article. This analysis of the researches is an attempt to reveal whether self-assessment corresponds to traditional knowledge evaluation and what kind of problems occur during such evaluation.

Keywords: self-assessment, student knowledge, evaluation criteria for assessment.

Introduction

The availability of multimedia technologies and distance learning makes traditional knowledge assessment rather time and money consuming. There are multiple methodologies and strategies of complementary knowledge evaluation (five factor model, intelligence quotient tests, etc.). Evaluation may be defined as the ability to judge the value of material for a given purpose (Žiliūkas *et al.* 2008). Self-assessment is usually presented as one of the alternatives of, or partial assistance to, traditional knowledge evaluation techniques. Wells *et al.* (1977) deliberated numerous descriptions of self-concept. Self-assessment is one of the most important formations of personality performing internal regulatory functions regarding behaviour and actions (Leary *et al.* 1995, Hogg and Abrams 1988). In academic literature, there are many different denominations of self-assessment, such as "self-reflection", "self-perception", "self-satisfaction" and others. Self-assessment may affect knowledge evaluation in different ways and various authors claim diverse impact of self-assessment on the final knowledge evaluation. Many authors give prominence to the role of self-assessment; self-assessment is a critical competency for both students and professionals (Zubin *et al.* 2007). The main aim of this article is to analyse the methodologies presented by various authors and the results obtained from practical self-assessment.

Self-assessment Understanding

The scale of current economic and social change, the process of globalisation, the rapid transition to a knowledge-based society and demographic pressure resulting from an ageing population in Europe are all challenges which demand a new approach to education (Kumpikaite 2008). Lately, distance studies, which are attempting to give the best possible education to students and to satisfy as many of their study needs as possible, are gaining wider popularity (Kaklauskas *et al.* 2009); but Radović Marković (2009) argues that online learning can assist in complementing studies when coupled with face-to-face learning and it is believed that online learning will not replace face-to-face learning.

In the past 25 years or so, self-assessment has become a more advocated and widespread assessment option in study process in foreign countries (White 2009). Self-evaluation is more useful for learners as a formative rather than a summative tool (Heather 1995).

Various authors give different definitions of self-assessment. Klenowski (1995) defines self-assessment as the evaluation or judgment of 'the worth' of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes. Brown (1998) defines self-assessment as any assessments that require students to judge their own abilities or performance. Cassidy (2007) defines self-assessment for students as the acceptance of responsibility for their own learning and performance. Self-assessment provides an approach in which learners typically rate themselves according to a number of criteria or dimensions (Bachman 2000).

Self-assessment in student evaluation is used for different reasons (Ross 2006):

1. Involving students in the assessment of their work, especially giving them opportunities to

ISSN 2029-2341 print / ISSN 2029-2252 online

doi:10.3846/mla.2010.034

contribute to the criteria on which that work will be judged, increases student engagement in assessment tasks.

2. Self-assessment contributes to variety in assessment methods, a key factor in maintaining student interest and attention.
3. Self-assessment has distinctive features that warrant its use.
4. Some teachers argue that self-assessment is more cost-effective than other techniques.
5. Students learn more when they know that they will share responsibility for the assessment of what they have learned.

Many academics are seeking to diversify assessment tasks, broaden the range of skills assessed and provide students with more timely and informative feedback on their progress. Others wish to meet student expectations for more flexible delivery and to generate efficiencies in assessment that can ease academic staff workloads. As more students seek flexibility in their courses, it seems inevitable that there will be growing expectations for flexible assessment as well (Formative Assessment 2009).

Next, we turn to the analysis of the evaluation criteria and methodologies presented by various authors. The results obtained from practical self-assessment are also analysed.

Problems and Criteria Analysis

Lindblom-Ylänne *et al.* (2006) focused in the study on comparing the results of self-, peer and teacher-assessment of student essays, as well as on exploring students' experiences of the self- and peer-assessment processes. Participants were 15 law students. The scoring matrix (Table 1) used in the study made the assessment easy, according to both teachers and students alike. Three people graded each critical essay. First, the student graded her or his own essay. Second, the student graded an essay of one of their peers. Third, the teacher graded all essays. The students were provided with ample opportunities to discuss and ask questions about the criteria during the course. They also used 'empty' versions of the matrix when assessing their own and a peer's essays. The self-, peer- and teacher-assessments were carried out independently of each other. Each criterion was scored on a four-point scale from "fail" to "excellent". The final grade was the mean score of self-, peer- and teacher-assessment (Lindblom-Ylänne *et al.* 2006).

Sung *et al.* (2009) conducted a study to determine the rating behaviours of teenagers in self- and peer assessments. The study involved 116 seventh graders (the first grade of middle school), where students individually playing musical recorders were subject to self- and peer assessments. The performance of the students playing a musical instrument (a recorder) was rated in a music class, which was a required item in the music classes at the school. Each student chose one of the two songs selected by the teacher and performed it for approximately 2 min. Each student was given a scoring card containing a form with the list of students' names in the class and five columns representing the following four criteria of the performance: proficiency (30 points), precision (30 points), interpretation (20 points) and posture on the stage (20 points). The fifth column was reserved for comments. The total score for each student, therefore, ranged from 0 to 100 (Sung *et al.* 2009).

Table 1. The scoring matrix and the criteria for self-, peer-, and teacher-assessment (Lindblom-Ylänne *et al.* 2006)

Assessment criterion	Excellent grade	Good grade	Satisfactory grade	Fail
Key issues and themes included	Relevant issues included	Most relevant issues included	Mistakes and irrelevant facts included	Severe mistakes and irrelevant facts
Coherent general picture	Thorough understanding of how events are linked	Understanding of how events are linked	Some understanding of how events are linked	No general picture formed
Independent thinking	Independent thinking and analytic approach	Some independent thinking	Little independent thinking	No independent thinking
Critical thinking	Critical evaluation and thinking	Attempts at critical evaluation	Very little effort in critical evaluation	No effort in critical evaluation
Use of literature	Several references, active search of references	Includes references other than "the main reference"	Only "the main reference"	No references, except discussions
Appearance	Tidy, accurate use of references	Tidy, inaccuracies in the use of references	Untidy, clear inaccuracies in the use of references	Untidy, inaccurate use of references
Length	9 – 11 pages	One page too long or short	Two pages too long or short	More than 2 pages shorter or longer

Brown (2005) reported the research on the development of an approach to learner self-assessment, which aims to facilitate the self-assessment of performance on extended-response writing tasks. The aims of the study reported here were to examine the usefulness of annotated samples of student work as a method for training students in assessing their own work reliably. The methodology was trialled on a small sample of 8 students. The end-of-module writing tasks require course participants to produce a short text such as a report or a letter. The criteria for assessment included content (30%), text features (30%), and style (15%), in addition to more narrowly defined linguistic features of grammar and vocabulary (25%).

The task of assessing and annotating the study participants' scripts in relation to the assessment criteria was undertaken jointly by two of the original course developers and tutors. For each task, they assessed the samples as low, medium or high. As they did this, they worked together to make explicit the basis for their judgements in terms of the various criteria, as this would form the basis of the annotations (Brown 2005).

In order to examine whether the annotated samples would help participants to evaluate accurately the overall quality of their own task performances, a peer- and self-assessment study was set up. Each participant was asked to assess their own and their peers' (i.e., those students who responded to the same tasks) performances as high, medium or low (Brown 2005).

Matsuno (2009) used Multifaceted Rasch measurement in the study with 91 student and 4 teacher raters to investigate how self- and peer-assessments work in comparison with teacher assessments in actual university writing classes.

Participants received instruction concerning essay writing such as essay format, mechanics, organization, and content. The participants practiced evaluating three essays together in class based on the essay evaluation sheet (Table 2). The students were then instructed to evaluate their own essay and the essays written by five peers at home, an assignment that was worth 10% of their final course grade (Matsuno 2009).

Ballantine *et al.* (2007) evaluated in the study the reliability of self-assessment as a measure of computer competence. This evaluation is carried out in response to recent research which has employed self-reported ratings as the sole indicator of students' computer competence. To evaluate the reliability of self-assessed computer competence, the scores achieved by students in self-assessed computer competence tests are compared with scores achieved in objective tests.

Table 2. Essay evaluation sheet (Matsuno 2009)

	Average					
	Too Many Mistakes Ineffective Very Poor					Very Few Mistakes Effective Very Good
1. Overall Impression	1	2	3	4	5	6
Content						
2. Amount	1	2	3	4	5	6
3. Thorough development of thesis	1	2	3	4	5	6
4. Relevance to an assigned topic	1	2	3	4	5	6
Organization						
5. Introduction and Thesis statement	1	2	3	4	5	6
6. Body and Topic sentence	1	2	3	4	5	6
7. Conclusion	1	2	3	4	5	6
8. Logical Sequencing	1	2	3	4	5	6
Vocabulary						
9. Range	1	2	3	4	5	6
10. Word/Idiom Choice	1	2	3	4	5	6
11. Word Form	1	2	3	4	5	6
Sentence Structure/Grammar						
12. Use of Variety of Sentence Structures	1	2	3	4	5	6
13. Overall Grammar	1	2	3	4	5	6
Mechanics						
14. Spelling	1	2	3	4	5	6
15. Essay Format	1	2	3	4	5	6
16. Punctuation/Capitalizasion	1	2	3	4	5	6
Comments						

The structure of this research instrument is as follows (Ballantine *et al.* 2007):

1. Section one comprised general questions to elicit background information on the subjects such as gender, whether they had studied IT before and the frequency with which they used a computer at home and at school prior to commencing university. When describing frequency of use at home and at school respondents were invited to choose from a five point Likert scale with verbal anchors, i.e., "daily" and "never".

2. Section two of the questionnaire consisted of 38 questions covering the six areas of computer competence. There were six questions each covering general information technology awareness, spreadsheets, word processing, databases and presentation software and eight questions in respect of e-mail/internet. Students were required to respond to statements such as "I feel comfortable opening and saving spreadsheet files" by selecting from a five-point Likert scale with a high positive anchor point at one end of the scale (5 representing "strongly agree") and a low negative anchor point at the other end (1 representing "strongly disagree"). Their responses represented their perceived level of knowledge in each of the areas of computer competence.

3. Section three of the questionnaire set out 18 multiple-choice objective tests. To be consistent with section two, multiple-choice objective tests represented each of the six areas of computer competence. Each of the multiple-choice questions had five possible answers, namely one correct, three deflectors and a fifth choice worded "I don't know". This fifth choice had been included to avoid the situation where respondents might be tempted to guess the answer to the questions.

Methods and Results of the Researches

In the research described by Lindblom-Ylänne *et al.* (2006), the evaluation of individual criteria was not integrated (importance evaluation of questions) into the general reduced criteria. The results of self-, peer-, and teacher-evaluation were compared only delivering a general schedule.

Lindblom-Ylänne's *et al.* (2006) comparisons among the results of self-, peer-, and teacher-assessment showed that they were quite similar to each other (Fig. 1).

Sung *et al.* (2009) used generalizability theory and criterion-related validity to obtain the reliability and validity coefficients of the self- and peer ratings. Analyses of variance were used to compare differences in self- and peer ratings between low- and high-achieving students. Low- and high-achieving students tended to over- and underestimate the quality of their work in self-assessment, respectively. The discrepancy between the ratings of students and experts was higher in group-work assessments than in individual-work assessments. The results have both theoretical and practical implications for researchers and teachers.

The results of Brown's (2005) study (Table 3) revealed that with the exception of one person, all participants demonstrated high agreement with each other and the tutors on the assessment of their own and their peers' performances. It was found to be both reliable and useful, allowing students not only to accurately evaluate their own performance but also to learn new language skills from the samples.

Because the multifaceted Rasch model does not require all raters to evaluate all essays, a sufficient degree of connectedness was achieved with the above rating. Multifaceted Rasch measurement was conducted using the FACETS computer program, version 3.22. In the analysis, writers, raters, and assessment criteria were specified as facets. The output of the FACETS analysis reported: (a) a FACETS map, (b) ability measures and fit statistics for each writer, (c) a severity estimate and fit statistics for each rater, (d) difficulty estimates and fit statistics for each assessment criterion, and (e) a bias analysis for rater writer interactions (Matsuno 2009).

The results indicated that many self-raters assessed their own writing lower than predicted. This was particularly true for high-achieving students. Peer-raters were the most lenient raters; however, they rated high-achieving writers lower and low-achieving writers higher. Most peer-raters were internally consistent and produced fewer bias interactions than self- and teacher-raters (Matsuno 2009).

Ballantine *et al.* (2007) used the Wilcoxon matched-pairs signed-ranks test, a non-parametric version of the paired difference t-test, because it was considered the most appropriate test for data analysis. The relative scores for the objective and subjective questions were paired within subjects and the differences analysed across all six areas of computer competence. The results of the test are presented in Table 4.

The results reveal a statistically significant overestimation of computer competence among the students surveyed. Furthermore, reported pre-university computer experience in terms of home and school use and formal IT education does not affect this result (Ballantine *et al.* 2007).

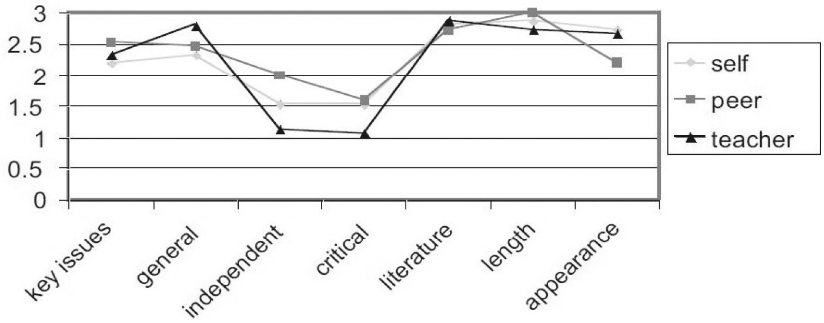

Fig. 1. Mean scores of self-, peer-, and teacher-assessment in each assessment criterion of the scoring matrix ($N = 15$) (Lindblom-Ylänne *et al.* 2006)

Table 3. Overall evaluation (Brown 2005)

Script	A1	A2	A3	B1	B2	B3	C1	C2
Teacher	H	M	L	H	M	L	H	L
Self-assessment	H	M	L	L	M	L	H	L
Other students' assessments	H (A2) H (A3)	M (A1) M (A3)	L (A1) L (A2)	H (B2) H (B3)	H (B1) M (B3)	M (B1) L (B2)	H (C2)	L (C1)

Key: H – high; M – medium; L – low.

Table 4. Wilcoxon matched-pairs signed-ranks test of the computer competence of entrants to an undergraduate business course using subjective and objective tests (Ballantine *et al.* 2007)

	Mean ranks[c]		Relobj[a] > Relsub[b]	Relobj[a] < Relsub[b]	Ties	Z score	P
	Relobj[a]	Relsub[b]					
General computing[d]	24.79	64.99	12	109	1	–8.778	0.000*
Spreadsheets	15.00	62.48	5	115	3	–9.313	0.000*
Word processing[e]	22.44	61.65	16	95	10	–8.092	0.000*
Databases	8.90	63.75	5	117	1	–9.475	0.000*
E-mail/Internet[f]	27.90	63.80	20	94	2	–7.690	0.000*
Presentation software[g]	33.16	63.88	22	93	2	–7.273	0.000*

[a] Relative score achieved in objective test.
[b] Relative score achieved in subjective test.
[c] Unless otherwise stated, N = 123.
[d] N = 122.
[e] N = 121.
[f] N = 116.
[g] N = 117.
* Indicate that differences are significant at the 1%.

Conclusions

The article analyses the research done by various authors on comparison of self-, peer-, and teacher-assessment (traditional evaluation methods) results. Some authors opted for other than mathematical methods to analyze the final results of their research. Different data needs different methods to obtain the final results.

The results of Brown's (2005) study revealed that, with the exception of one person, all participants demonstrated high agreement with each other and the tutors on the assessment of their own and their peers' performances. The research of Lindblom-Ylänne *et al.* (2006) showed that comparisons among the results of self-, peer-, and teacher-assessment were quite similar to each other. These results are similar to the results of Brown's (2005) research. Sung *et al.* (2009) indicated that low- and high-achieving students tended to over- and underestimate the quality of their work in self-assessment, respectively. Matsuno's (2009) results were similar to those of Sung *et al.* (2009) and indicated that many self-raters assessed their own writing lower than predicted. This was particularly true for high-achieving students. The results presented by Ballantine *et al.* (2007) reveal a statistically significant over-estimation of computer competence among the students surveyed.

The analysis of the research done by different authors revealed an absence of big differences between self-, peer-, and teacher- assessment. However, self-sufficiency or self-offence was observed in quite a few cases. It can be stated that, using this type of student knowledge evaluation, authors must take into consideration self-assessment inaccuracies.

Self-assessment criteria are chosen according to the form of knowledge. It must be noted that all evaluators should understand the evaluation criteria equally to maintain the final results unaffected.

References

Bachman, L. 2000. *Learner directed assessment in ESL,* In G. Ekbatani & H. Pierson (Eds.), New Jersey: Lawerance Erlbaum Associates, ix–xiii.

Ballantine, J. A.; McCourt Larres, P.; Oyelere, P. 2007. Computer usage and the validity of self-assessed computer competence among first-year business students, *Computers & Education* 49(4): 976–990. doi:10.1016/j.compedu.2005.12.001

Brown, A. 2005. Self-assessment of writing in independent language learning programs: The value of annotated samples, *Assessing Writing* 10: 174–191. doi:10.1016/j.asw.2005.06.001

Brown, J. D. 1998. *New ways of classroom assessment.* Alexandria, VA: TESOL Incorporated.

Cassidy, S. 2007. Assessing “Inexperienced” Students' Ability to Self-assess: Exploring Links with Learning Style and Academic Personal Control, *Assessment & Evaluation in Higher Education* 32(3): 313–330. doi:10.1080/02602930600896704

Formative Assessment in Computer Supported Learning. 2009. [cited 1 Januray 2010]. Available from Internet. <http://en.wikipedia.org/wiki/Formative_assessment>.

Heather, A. 1995. Student self-evaluations: How useful? How valid? *International Journal of Nursing Studies* 32(3): 271–276. doi:10.1016/0020-7489(94)00043-J

Hogg, M. A.; Abrams, D. 1988. *Social identifications*. New York. Routledge and Kegan Paul.

Kaklauskas, A.; Zavadskas, E. K.; Budzevičienė, R. 2009. Web-based Model of Multiple Criteria Ethical Decision-Making for Ethical Behaviour of Students, *Journal of Business Economics and Management* 10(1): 71–84. doi:10.3846/1611-1699.2009.10.71-84

Klenowski, V. 1995. Student self-evaluation processes in student-centred teaching and learning contexts of Austria and England, *Assessment in Education* 2(2): 145–163. doi:10.1080/0969594950020203

Kumpikaitė, V. 2008. Human Resource Development in Leraning Organizations, *Journal of Business Economics and Management* 9(1): 25–31. doi:10.3846/1611-1699.2008.9.25-31

Leary, M. R.; Tombor, E. S.; Terdal, S. K.; Downs, D. L. 1995. Self-esteem as a interpersonal monitor: The sociometer hypothesis, *Journal of Personality and Social Psichology* 68: 518–530. doi:10.1037/0022-3514.68.3.518

Lindblom-Ylänne, S.; Pihlajamaki, H.; Kotkas, T. 2006. Self-, peer- and teacher-assessment of student essays, *Active Learning in Higher Education* 7(1): 51–62. doi:10.1177/1469787406061148

Matsuno, S. 2009. Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms, *Language Testing* 26(1): 75–100. doi:10.1177/0265532208097337

Radović Marković, M. 2009. Education through E-learning: Case of Serbia, *Journal of Business Economics and Management* 10(4): 313–319. doi:10.3846/1611-1699.2009.10.313-319

Ross, J. A. 2006. The Reliability, Validity, and Utility of Self-Assessment, *Practical Assessment Research & Evaluation* 11(10): 1–13.

Sung, Y.; Chang, K.; Chang, T.; Yu, W. 2009. How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments, *Journal of Adolescence*, Corrected Proof, Available online 9 June 2009.

Wells, E. L.; Marwell, G. 1977. *Self Esteem. Its Conceptualizacion and Measurement.* Sage, Beverley Hills.

White, E. Assessing the Assessment: An Evaluation of a Self-assessment of Class Participation Procedure, *The Asian EFL Journal Quarterly* 11(3): 75–109.

Žiliūkas, P.; Katiliūtė, E. 2008. Writing and Using Learning Outcomes in Economic Programmes, *Inzinerine Ekonomika – Engineering Economics* (5): 72–77.

Zubin, A.; Gregory, P. A. M. 2007. Evaluating the Accuracy of Pharmacy Students’ Self-Assessment Skills, *American Journal of Pharmaceutical Education* 71(5): 2–8.

STUDENTŲ ŽINIŲ VERTINIMO TAIKANT SAVĘS VERTINIMO METODOLOGIJĄ ANALIZĖ: KRITERIJAI, PROBLEMOS, REZULTATAI

A. Matuliauskaitė, E. Žvirblis

Santrauka

Darbe analizuojami skirtingų autorių tyrimai, kuriuose nagrinėjama žinių vertinimo problematika taikant savęs vertinimą atsižvelgiant į sprendžiamą problemą, kriterijus, taikomus savęs vertinimui, jų integravimą į galutinius rezultatus. Atliekant tyrimų analizę, siekiama išsiaiškinti, ar savęs vertinimas atitinka žinių įvertinimą naudojant tradicinius vertinimo metodus, su kokiomis problemomis susiduriama atliekant tokį vertinimą.

Reikšminiai žodžiai: savęs vertinimas, studentų žinių vertinimas, vertinimo kriterijai.