How effective are self- and peer assessment of oral presentation skills compared to teachers' assessments?

Luc De Grez, Martin Valcke and Irene Roozen

Biographical notes

Luc De Grez is assistant professor in Psychology and Leadership Behaviour at the University College Brussels (PhD Ghent University). His research interests include the learning and teaching of communication skills, and entrepreneurial learning.
Address: University College Brussels, Stormstraat 2, B1000 Brussel, Belgium. Tel: 0032 2 6081416. Email: luc.degrez@hubrussel.be

Martin Valcke is full Professor 'Instructional Sciences' at Ghent University, Belgium and head of the department "Educational Studies". His main research field is linked to the innovation of higher education and the integrated use of Information and Communication Technologies (ICT). A full CV can be downloaded via http://users.ugent.be/~mvalcke/CV/CVMVA.htm
Address: University of Ghent, H. Dunantlaan 2, B9000 Gent, Belgium. Tel: 0032 9 2648675 Email: martin.valcke@ugent.be

Irene Roozen is assistant professor in Marketing at the HUBrussel. Her research interests include marketing communication, advertising effectiveness and assessment of communication skills.
Address: University College Brussels, Stormstraat 2, B1000 Brussel, Belgium. Tel: 0032 2 6081422. Email: Irene.roozen@hubrussel.be

Abstract

Assessment of oral presentation skills is an under-explored area. The study described here focuses on the agreement between professional assessment and self- and peer assessment of oral presentation skills and explores student perceptions about peer assessment. The study has the merit of paying attention to the inter-rater reliability of the teachers. Comparison of the teacher and peer assessment rubric scores points at a positive relationship, but also at critical differences. The lower intra-class correlation suggests that peers and teachers still interpret the criteria and indicators of the rubric in a different way. With regard to the comparison of self-assessment scores and teacher scores, we have to conclude that there are significant differences between these scores. Self-assessment scores are, for the most part, higher than the marks given by teachers. The results also reflect a very positive attitude of students towards peer assessment as a relevant source of external feedback.

Key words: oral presentation skills, assessment, peer assessment, self-assessment, student perception

Quality of assessment

Recent approaches towards assessment stress the learning potential of assessment (Taras, 2008). This is labelled as formative assessment and defined as "assessment that is specifically intended to provide feedback on performance to improve and accelerate learning" (Nicol and Milligan, 2006: p. 64). Some consider this as a key quality of assessment and regard this as the "consequential validity" of assessment (Gielen et al., 2003). Consequential validity is put next to the two other traditional psychometric qualities of an assessment: reliability and validity. According to Messick (1994) consequential validity is one of the

six aspects of his unified concept of validity. Involvement of students in assessment can be organized in two ways: self- and peer assessment. In peer assessment, according to Falchikov (2005: p.27), "(…) students use criteria and apply standards to the work of their peers in order to judge that work". Building on the latter, we state that in self-assessment students use criteria and apply standards to judge their own work. Both self- and peer assessment are expected to decrease the central role of the teacher in assessment activities. During the last decades, there has been an increase in the implementation of self- and peer assessment in higher education learning environments (Segers et al., 2003). Despite this increased interest, formative assessment in higher education is still largely controlled by the teachers (Nicol and Macfarlane-Dick, 2006).

The theoretical position of self- and peer assessment in a self regulated learning process

In this study a social cognitive theoretical perspective towards self-regulated learning is adopted as a theoretical basis for oral presentation skills instruction (Bandura, 1997; Schunk, 2001). This choice builds on the literature that links the instruction of oral presentation skills to observational learning (Bandura, 1997). Via observational learning, learners compare their performance or the performance of others with more or less explicit standards of a good oral presentation. The oral presentation skills will evolve by achieving a better match between these standards and the current performance level (Sadler, 1989). We adopt the term *calibration* to refer to the match between an internal evaluation and a standard (Winne, 2004). Both internal and external sources of feedback are helpful to foster the calibration process to attain higher performance levels in the context of productive self-regulated learning (Winne, 2004). The calibration activity can be fostered by providing opportunities for self-assessment.

External feedback from peers can play a comparable role (Topping, 1998). An accurate calibration of oral presentation performance and the standards suggests that a sufficient level of reliability can be attained when comparable assessment results are reported by a teacher/expert, by peers, or by the learner. Self- and peer assessment result in a more active involvement of students in their own learning process (Ozogul and Sullivan, 2007). A student who always expects teachers to present a judgment will develop to a lesser extent a self-assessment orientation (Boud and Falchikov, 2006). From a self-regulated learning point of view, it is however critical to develop self-observation skills that help to compare the information gathered via self-observation to a performance goal. Sub-processes related to self-observation and self-judgment are important. They are regarded as the steps in a learning monitoring process that helps learners to bring their behaviour in line with their performance and goals (Schunk, 2001). Next to self-assessment, peer assessment was also found to have positive effects on domain-specific and on peer assessment skills (Van Zundert et al., 2009). Topping (2009) explains this by linking peer assessment to the provision of immediate, individualized and richer feedback. Since this feedback is formative in nature, it has a clear potential of fostering the subsequent learning process (Hattie, 2009).

Analysis of the assessment of oral presentation skills mainly results in an overview of studies about self- and peer-assessment of individual (oral) presentation skills (AlFallay, 2004; Campbell et al., 2001; Cheng and Warren, 2005; Hafner and Hafner, 2003; Hughes and Large, 1993; Langan et al., 2005; Langan et al., 2008; Magin and Helmore, 2001; Oldfield and Macalpine, 1995; Patri, 2002; Sellnow and Treinen, 2004). In some studies, the research focuses only partly on self- and peer assessment (Fallows and Chandramohan, 2001). In a minor number of cases, group presentations have been assessed (Kerby and Romine, 2009).

In general, research about self- and peer assessment of oral presentation skills reveals under-explored areas and diverging views. Moreover, the use of very different samples and different assessment instruments makes it difficult to compare the findings of those studies. AlFallay (2004) for instance involved students in applied sciences enrolled in an English language programme, while Patri (2002) involved Chinese students and Campbell et al. (2001) American students. These results indicate that more research is needed regarding self- and peer assessment of oral presentation skills.

Benefits of self- and peer assessments

Falchikov (2005: p.16) hypothesizes that "involving students in the assessment of presentations is extremely beneficial" for developing self regulating skills. Students are expected to analyze their own behaviour and develop a better understanding of the nature of quality criteria. Cheng and Warren (2005) cite several studies that reported improved presentation performance due to peer assessment. Others adopt in this context videotaped feedback for self-assessments, and also report the attainment of improved oral presentation skills (Bourhis and Allen, 1998).

Topping (1998) dedicates part of his review of the literature about peer assessment to the assessment of oral presentation skills. He summarizes improvements in marks, perceived higher learning performance, higher presentation confidence (self-efficacy), and the development of appraisal skills. Topping (2003) additionally mentions economical benefits to adopt self- and peer assessment. Shifting part of the responsibilities for assessment and feedback from the teacher to the student has – next to educational benefits – also benefits in terms of reducing teaching workload.

Inter-rater reliability of self- and peer assessments

There is considerable debate about the inter-rater reliability of self- and peer assessments (Topping, 2009). Topping (2003) points at the widespread approach in the discussions about reliability of comparing self- and peer assessments to assessment by teachers. Topping (2003) stresses that the a priori assumption that assessment by a teacher is more reliable and more valid can be doubted in some contexts. This a priori assumption relates to a positivist epistemological perspective about assessment (Elton and Johnston, 2002). There is little research that actually tested this assumption, and sometimes disagreement between tutors is mentioned (Langan et al, 2008).

Freeman (1995) concludes that there is no significant difference in the overall mark averages given by peers or given by teachers. In contrast, Langan et al. (2005) report that peer marks are on average 5% higher than marks given by their tutors. Other correlational studies conclude that peer assessment can be a relevant substitute for assessments by teachers (AlFallay, 2004; Campbell et al., 2001; Hughes and Large, 1993; Oldfield and Macalpine, 1995; Patri, 2002). Nevertheless, Hughes and Large (1993) warn that a high correlation between marks of peers and teachers can still hide a considerable variation in the marks. Freeman (1995) reports only a moderate correlation between peer and teacher scores. He also reports that the standard deviation of marks given by peers was half the value of the standard deviation in scores given by teachers (see also Hughes and Large, 1993). Cheng and Warren (2005) add to this that student average marks are within one standard deviation of teacher marks, but they point out that students did not always assess the same elements or criteria as their teachers did. Kappe (2008) found that students were able to provide a reliable overall assessment but needed additional training to provide reliable marks on specific criteria of oral presentations. Hafner and Hafner (2003) adopted regression analysis showing a significant positive functional relationship between instructor and mean peer scores and add that students come to a strong agreement in the final ranking of their scores.

Fewer studies are found that compare teacher assessment with self–assessment of oral presentation skills. In addition, it is clear that results are not univocal. Some studies report lower correlations values between self and teacher assessments as compared to the correlation values between teacher and peer assessment (Campbell et al, 2001; Langan et al., 2008; Patri, 2002). Nevertheless, others consider self-assessments to be as valid as peer assessment (AlFallay, 2004; Hafner and Hafner, 2003).

Variables affecting the quality of self- and peer assessment of oral presentations

A continuous debate is observed about the reliability of self- and peer assessment in relation to the development of presentation skills. In the context of peer assessment, rating errors and the impact of student perceptions about peer assessment is a key topic for discussion. Rating errors are central in the study of Sluijsmans et al (2001) who refer to personal differences in standards and rating styles, and the extent to which peers distribute grades and have different opinions about the rating tasks. Student perceptions are stated to have a considerable influence on student learning (Struyven et al., 2003). Concerns have been raised about resulting difficulties in peer assessment contexts (Hanrahan and Isaacs, 2001). Results show that students were concerned about their inexperience in marking, they felt

uncomfortable critiquing others' work and remarked that their marking input was not taken seriously because it was not considered when calculating the final mark. Students also complained about the time-consuming nature of the activity and asked feedback as to their involvement in the assessment.

Only a small number of studies explore the views students hold about peer assessments of oral presentation skills. The findings of Cheng and Warren (2005) showed that students reflected a low level of comfort in a peer assessment situation, and a low degree of confidence in their personal peer assessment skills. This suggests that low self-efficacy levels for peer assessment skills can affect the nature and quality of peer assessment. Langan et al. (2005) point at obvious problems with anonymity when building on peer assessment of oral presentations. Lack of anonymity may lead to assessment bias. These authors also detected gender effects and found that peers rated students from the same university slightly higher than students from other universities. Falchikov (2005, p.154) cites a study of Lapham and Webster of 1999 who mention bias when peers are asked to mark seminar presentations. Lastly, Sellnow and Treinen (2004) report that neither the gender of the presenter, nor the gender of the assessor, affects overall peer ratings.

As to self-assessment, a meta-analysis of Falchikov (2005) indicates that some, but clearly not all, students are able to assess in the way teachers apply assessment criteria. This is confirmed by a study of Kruger and Dunning (1999) where novices and low performers overestimate their performance level and lack related metacognitive abilities (monitoring, evaluation). Rust et al (2003) and Langan et al. (2008) concluded that women are more likely to underestimate their performance, whereas males tend to overestimate the quality of their performance in a self-assessment context.

Improving the quality of self- and peer assessments

There is no unequivocal answer as to how the quality of self- and peer assessments of oral presentation skills can be improved. Earlier research focused on the critical value of assessment training, the feasibility of student based assessment, the nature of the assessment criteria and the scoring approach. Hafner and Hafner (2003) state that providing training is not sufficient. In addition, Carlson and Smith-Howell (1995) found hardly any differences in assessment practices between untrained and trained teachers. Others conclude that peers need training in view of peer assessment (AlFallay, 2004; Campbell et al., 2001; Freeman, 1995; Patri, 2002; Sluijsmans, 2002). Langan et al. (2005) found that marks awarded by students who participated in preliminary discussions about the assessment criteria were significantly lower than the marks of students who were not involved in these initial discussions. Compensating for low self-efficacy related to peer assessment during these discussions was a key point in the study of Fallows and Chandramohan (2001). Miller (2003) concluded that a larger number of items in the evaluation checklist resulted in an increase in variance in scores. This diminishes inter-rater reliability but, on the other hand, provides students with more detailed and thus better feedback. In contrast, Freeman (1995) suggests reducing the number of criteria in the checklist, diminishing the quality of feedback generated by the assessment. Langan et al. (2008) recommend adopting short sessions in order to diminish loss of concentration.

To conclude, many questions remain unanswered. The main problem is the lack of empirical research that direct specific practices in the field of self- and peer assessment (Sluijsmans, 2008). Building on the theoretical and empirical basis, outlined above, we put forward the following research questions:

- What is the level of agreement between undergraduate students' peer assessments and the assessments of university teachers in terms of oral presentations?
- What is the level of agreement between self-assessments and assessments by university teachers?
- What are the student perceptions about peer assessment?

Research Design

Participants

Oral presentations and answers on a questionnaire were collected from 57 university freshmen students (21 females and 36 males) enrolled for a Business Administration introductory course about psychology. The research was set up during the first semester. Administration of the questionnaire was carried out at the end of the semester (December). Initially, 73 students participated in the study but due to illnesses, incompatibility of roster, internships or other reasons among the students, 16 students dropped out. The reasons for drop out were deemed not to be systematic. The average age among the 57 students who participated was 18 years. Informed consent was obtained from all participants, but they were not informed about the nature of the research questions.

Research instruments

Assessment instrument for oral presentation performance

In a preliminary study (De Grez et al., 2009b), six existing assessment scales to judge the quality of an oral presentation were analysed by four experienced higher education teachers acquainted with the knowledge domain. On the basis of the results of semi-structured interviews with these experts, a rubric was developed consisting of nine oral presentation evaluation criteria: three content-related criteria (quality of introduction, structure, and conclusion), five criteria about the nature of the delivery (eye contact, vocal delivery, enthusiasm, interaction with the audience and body language), and a general quality criterion (professionalism). Descriptors and indicators were added to support the use of the assessment criteria in the rubric. These were improved after application of the rubric by trained assessors that judged the quality of more than 300 oral presentations (for a detailed description, see De Grez et al.,2009b).

On the basis of a factor analysis, two evaluation components can be distinguished when applying the nine criteria: content related criteria, and delivery related criteria. One item, labelled “professionalism” loads on both components (De Grez et al., 2009b).

(insert table 1 about here)

University teachers and peer assessors were asked to rate the quality of an oral presentation on the basis of the rubric. For each criterion a 5 point Likert scale was used. Descriptors and indicators are provided to support the assessment process.

As an example, we describe the assessment related to the criterion “quality of the introduction”. Assessors are invited to consider three indicators to score this criterion:

- Grasps the attention of the audience with the first sentences.
- Gives a goal or central idea of the presentation in the introduction.
- Gives an idea of the structure of the presentation in the introduction.

The score reflects the extent to which the quality of the introduction meets none, one, or more of the three indicators.

Perception of peer assessment and of the learning process

A subscale focusing on “perceptions of peer assessment” was adopted from the seven item questionnaire (α = .74) developed and validated by Sluijsmans (2002). One item that was context-specific for the study of Sluijsmans (2002) was omitted from the scale, and some words were changed in order to adapt the subscale to the specific oral presentation situation (for example ‘You can learn from the feedback of peers’ and ‘I think students should be able to assess each other’). The scale, as used in this study, reflects a good reliability (α = .80). Respondents were also asked how much they learned from seven instructional components (on a ten-point Likert scale).

Teachers and peer assessors

The oral presentations (recorded) were assessed by five different assessors (2 male; 3 female) on the basis of the assessment rubric as explained above. Four of these assessors were faculty members with at least 5 years of a language teaching background but who had not taught the students being assessed. The fifth assessor was a junior researcher. These assessors and their assessments are labelled as teachers in this article. Next to the teachers, 47 students were involved as peer assessors. These students did not belong to the same group as those being assessed. They were enrolled in the second year Business Administration (32 male) and participated in the study as a formal part of their course about communication skills. Both the teachers and the peer assessors were unaware of the nature of the research questions. All the teachers received some short training (45 minutes on average) about the nature and use of the assessment rubric. Peer assessors received, as part of their formal instruction programme, an introduction to oral presentation skills and the use of the evaluation rubric. They had extensive experience in using the assessment instrument, as part of their communication course.

Procedure

The research sample was, as a formal part of their psychology course, invited to deliver three short (on average three minutes) oral presentations about a prescribed topic. Topics were of similar difficulty level: my town, my high school, my university. All the presentations were recorded. In the research setting, next to the presenter, an audience was present consisting of two other participants (but always a different pair), and the first author. A camera, recording the session, was positioned in an unobtrusive location. Due to drop-out of a number of participants for the second or third presentation, the final number of recordings was 209 instead of 219 (73 students x 3 oral presentations).

After the first presentation, students participated individually in a computer-based multimedia training programme about oral presentations. After the second presentation, students received feedback about their first presentation, based on the scores for the nine assessment criteria (see below). The intervention was spread over nine lesson weeks.

Assessors and the assessment procedure

The evaluation of the oral presentations by the teachers and peer assessors was based on video recordings. None of the assessors was aware whether they assessed a recording of a first, a second or a third oral presentation. Recorded presentations were assigned randomly to assessors. Table 2 describes the allocation of specific assessors to specific oral presentations.

(insert table 2 about here)

Each teacher individually evaluated between 34 to 49 of the total number of 209 oral presentations. For each oral presentation, scores were required for the 9 rubric criteria. Student peers assessed 29 presentations. Each of these 29 presentations was assessed by six different peers. The choice of this specific number is based on the work of Hafner and Hafner (2003) who reported a large improvement in reliability when moving from a single rater to about five raters, and on the work of Dannefer et al. (2005) who concluded that at least six peers are needed to achieve a moderate reliability when assessing professional competences. As part of the research design, participants were also asked to rate their own presentation on the basis of the assessment rubric. One third of the participants rated their first presentation and all the participants rated their second presentation. The rubric was introduced as part of the multimedia instruction package. Therefore, we assume that students were sufficiently acquainted with the rubric criteria in view of the self-assessment activity.

The consistency between scores of different raters is central to the concept of reliability. There is much debate about ways to calculate or estimate the inter-correlations between assessors. As suggested by Cho et al (2006), we calculate intra-class correlation coefficients. These authors (ibid, p.896) state that we have to adopt measures that are not influenced by distribution features, so correlation measures are preferred to percentage agreement measures. However, also the use of Pearson product-moment

correlations can be criticized. Together with Shrout and Fleiss (1979) intra-class correlations (ICC), a common measure of reliability of either different judges or different items on a scale, are used in this study.

Luc, this paragraph is 'research methods', not 'literature review', so it belongs in this section, not the one above (where it was). Whether this paragraph should be right here, at the end, is up to you. If you want to move it to somewhere else in the 'research design' section, please feel free to do so!

Results

Initial analyses

Prior to the actual analysis of the research data, a quality control of the assessment process was carried out. This focused on the way teachers applied the assessment rubric. Analysis of variance was applied to test differences. Post hoc comparisons confirm that teachers did not differ significantly in applying the rubric criteria Introduction, Structure and Contact with audience. But significant differences were observed in view of the other six criteria. Additional analysis reveals that, for five of the six criteria, it was consistently the same teacher adopting a more lenient view as compared to the other assessors. Scores from this teacher were removed from the data set.

To detect gender bias, an analysis of variance was carried out to compare whether the gender of the teacher and the gender of the assessed participant resulted in significantly different oral presentation skill sum scores. The results indicate that there was no significant difference between the scores of male and female presenters when assessed by a male or a female teacher.

What is the level of agreement between peer assessments and teacher assessments?

After calculating the sum score of the nine rubric criteria sum scores of teachers and peer assessors were compared. It is important to keep in mind that, as indicated in table 2, peers only assessed "first" presentations. Table 3 summarizes the analysis results. We can conclude that we have achieved an acceptable but low reliability level considering the value of the intra-class correlation.

(insert table 3 about here)

The rubric sum score reported by teachers is significantly lower as compared to the peer assessments (t=6.210; p< .001). To detect possible gender effects, an analysis of variance was carried out with gender of the assessor and the gender of the assessed student as independent variables and the oral presentation sum score as the dependent variable. The analysis was repeated for teachers and for peer assessors. Results indicate that gender of the teachers ($F_{(1,205)}$= .03, p=.87) and of the peer assessors ($F_{(1,170)}$ = .85, p=.36) did not result in significant differences in scoring. This implies that male assessors were not more severe or lenient than female assessors, and this was the case for teachers and for peer assessors. The interaction effect gender of the assessor and gender of the assessed was not significant for teachers but was significant for peers $_{(F(1,\ 170)}$ = 4.17, p <.05), as can be seen in figure 1. Male peers attributed higher scores than female peers to female presenters. Female presenters did however obtain higher scores than male presenters.

(insert figure 1 about here)

What is the level of agreement between self-assessments and teacher assessments?

In view of this question, the scores of teachers were compared with the scoring results obtained via self-assessment. It is important to keep in mind that, as indicated in table 2, participants self-assessed first and second presentations. Table 4 summarizes the analysis results. We can conclude that we have achieved again an acceptable but low level of reliability considering the value of the intra-class correlation.

(insert table 4 about here)

The 'total' rubric score of teacher assessments is significantly lower as compared to self-assessment scores (t = 6.19; p< .001). The self-assessment scores of male and female participants are not significantly different ($F_{(1,75)}$ = .30, p=.58).

What are the student perceptions about peer assessment and the learning process?

The average perception score about peer assessment reflects a predominantly positive opinion about peer assessment. Comparison of the first (m = 3.67 and second administration (m = 4.11) of the scale points at a significant increase in the positive appreciation of peer assessment (t = 4.11; p = < .001).

Participants ranked the instructional components and results showed they believed they learned most from the feedback (M = 7.42) and least from their first presentation (M = 5.91). They indicated, however, that they learned more from the second (M = 6.79) and third (M = 7.02) presentation.

Discussion and conclusions

In this study, alternative assessment approaches were explored. Self- and peer assessment were positioned within a social cognitive perspective on self-regulated learning. The limited, and often contradictory, empirical evidence about self- and peer assessment of oral presentation skills prompted the design of a study in which self- and peer assessment was contrasted with the assessment by teaching staff.

Comparison of the teacher and peer assessment rubric scores points at a positive relationship, but also at critical differences. The lower intra-class correlation described here suggests that peers and teachers still interpret the criteria and indicators of the rubric in a different way. This can be explained by differences in the width and depth of their experience basis. Also, within the group of peers, not all students could have applied the same criteria in a comparable and/or consistent way. Lastly, the finding that peers report higher marks as compared to teachers is in agreement with the results of other studies (Langan et al., 2008).

With regard to the comparison of self-assessment scores and teacher scores, we have to conclude that there is a level of agreement and disagreement when assessing the oral presentations. Also the finding that the self-assessment scores are, mostly, higher than the marks given by teachers is consistent with the results reported in the literature (Patri, 2002).

As stated above, these scoring differences can be explained by the broader experience of teachers when judging the quality of oral presentations. They can retrieve from their memory a larger set of models that exemplify how oral presentations do or do not meet the criteria. Price and O'Donovan (2006) mention tacit knowledge that is experience-based and can only be made explicit through the sharing of experiences. This also implies that in an implicit way, teachers add criteria and/or indicators when judging the quality of an oral presentation. This adds to the unreliable, but often neglected, nature of teacher assessments. This study has the merit of paying attention to the inter-rater reliability of the teachers. As explained above, one of the teachers applied a number of the criteria in a more lenient way. This problem was tackled by removing the particular assessment from the data set. Nevertheless, in a normal instructional setting, teachers have to be aware of bias potentially caused by assessors approaching the

criteria in diverse ways. This should also be considered when setting up assessment related research (Topping, 2009).

With regard to the research question focusing on student perceptions of peer assessment, it can be concluded that the results reflect a very positive attitude towards the value of peer assessment. In addition, the results indicate that the actual process of carrying out self- and or peer-assessment affects this perception in a positive way. This is a promising finding in the light of the impact of perceptions on the outcomes of student learning (Struyven et al., 2003). We might assume that students' perceptions of peer assessments will influence their willingness to take into account the feedback generated by peer assessment and to actually do something with the feedback. This positive attitude is probably also reflected in the ranking of the instructional components when participants declared that they learned most from the feedback.

Gender was also studied as a potential source of bias. Neither the gender of the assessor nor the student being assessed seems to influence the assessment process or assessment marks. The interaction effect gender of the assessor and gender of the assessed was not significant for teachers but was significant for peers. Gender effects are also reported by some (Edens et al., 2000; Langan et al., 2005) but not by others (Sellnow and Treinen, 2004). It is possible that the gender effect, observed in peer assessments, interacted with the lower correlation between peer and teacher assessments. This gender effect can lower the quality of the peer assessments and the correlation with teacher assessments. The gender effect was caused by male assessors who gave higher marks than female assessors to female presenters. Male assessors might have been somewhat biased towards female presenters and were too generous with their marks. It is also possible that female assessors underestimated the female presenters. Further research should shed more light on this issue.

Although a large number of recorded oral presentation sessions was assessed by peers, teachers and students themselves, the study remains limited when it comes to sample size, duration of the instructional intervention, scope of the skills to be mastered and the complexity level of the competencies. An important limitation is the limited variation in nature and background of the participants. The study has to be replicated involving students from other domains and from other educational levels. Additional research could focus on the impact of assessment training and student collaboration in relation to defining assessment criteria. Future studies should also consider the nature of the target audience that could vary in knowledge domains and expertise levels. These studies should investigate the short term, middle term and long term effects of self- and peer assessment. It would also be interesting to investigate the impact of individual and interpersonal variables (Van Gennip et al., 2009). In this study, like in many other studies, teacher assessments were compared with peer and self-assessments. It might be interesting, in future research, to additionally compare self- and peer assessments.

Though a more in-depth and qualitative analysis of differences in scoring behaviour between teachers, peers, and the students giving the oral presentations is beyond the scope of the present study, we have to keep in mind that the requirement to attain a high level of reliability was not completely answered (Price and O'Donovan, 2006; Topping, 2003). This finding suggests that the training of assessors is very important, and is in line with ideas presented (Van Zundert et al., 2009). Such training could provide more examples and/or more concrete indicators. The suggestion is also very important for practitioners. They should provide extensive training to learners. Enriching the learning process with an explicit discussion of the assessment approach and criteria between peers and between teacher and learners could enhance the quality of the educational process.

Nevertheless, the results do not question the value of self- and peer assessment of oral presentation skills, and practitioners are advised to use both forms of assessment in order to provide the learner with a sufficient level of formative feedback. Also Langan et al. (2005) and Sluijsmans (2002) make it clear that the benefits of peer-assessment outweigh a certain degree of discrepancy between student marks, tutor marks, and peer markings. Boud (2007) refers in this context to the consequential validity of assessment. This means that the value of self- and peer assessment is also to be found in the impact on the acquisition process of the complex oral presentation skills. Some, such as Winne (2004), stress the importance of the accuracy of feedback in view of future learning outcomes. But others, such as Gibbs

(2006) and Yorke (2003), state that it is not only the quality of the feedback evolving from the assessment that is crucial but what a student does with the feedback. In our opinion, a combination of both views is needed. On the one hand we do not want students to take wrong actions based on low quality feedback but on the other we want students to do something with the feedback.

The question is therefore how to improve the quality of self- and peer assessment approaches. Falchikov (2005) recommends developing evaluation criteria in close collaboration with students. Price and O'Donovan (2006) warn that it is insufficient to concentrate on more detailed indicators for assessment criteria or standards because these indicators can become counterproductive if they are too comprehensive. These stress the importance of giving students sufficient practice and discussion to develop a shared understanding of the explicit and tacit assessment criteria. Part of the less positive results of the present study can therefore be explained on the basis of insufficient practice. The students did not get sufficient opportunities to practice with the assessment criteria. This conclusion also challenges the statements of Hafner and Hafner (2003) and Carlson and Smith-Howell (1995) that assessment training is not essential.

Our study revealed some interesting results about the, until now, under-explored self- and peer assessment field of oral presentation skills. Additional research could help to clarify the relationship between self- and peer assessment and our theoretical framework about self-regulated learning.

References

AlFallay, I. (2004) 'The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer – assessment', *System 32*: 407-425.

Bandura, A. (1997) *Self efficacy: the exercise of control*. New York: Freeman.

Boud, D. (2007) '*Assessment design for learner responsibility'*, http://ewds.strath.ac.uk/public/reap07/Boud-web/img0.html [Accessed 3 November 2007].

Bourhis, J., & Allen, M. (1998) 'The role of videotaped feedback in the instruction of public speaking: a quantitative synthesis of published empirical research', *Communication Research Reports, 15(3):* 256-261.

Campbell, K., Mothersbaugh, D., Brammer, C., & Taylor, T. (2001) 'Peer versus self-assessment of oral business presentation performance', *Business Communication Quarterly, 64(3)*: 23-42.

Carlson, R., & Smith-Howell, D. (1995) 'Classroom public speaking assessment: reliability and validity of selected evaluation instruments', *Communication Education, 44:* 87-97.

Cheng, W., & Warren, M. (2005) 'Peer assessment of language proficiency', *Language Testing, 22(1)*: 93-121.

Cho, K., Schunn, C. & Wilson, W. (2006) 'Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives', *Journal of Educational Studies, 95(4*): 891-901.

Dannefer, E., Henson, L., Bierer, S., Grady-Weliky ,T., Meldrum, S., Nofziger, A., Barclay, C., & Stein, R. (2005) 'Peer assessment of professional competence', *Medical Education, 39(7)*: 713-722.

De Grez L., Valcke M., & Roozen, I. (2009a) 'The impact of an innovative instructional intervention on the acquisition of oral presentation skills in higher education', *Computers & Education*, 53: 112-120.

De Grez, L., Valcke, M., & Roozen, I. (2009b) 'The Impact of Goal Orientation, Self-reflection and Personal Characteristics on the Acquisition of Oral Presentations Skills', *European Journal of Psychology of Education*, *24(3):* 293-306.

Edens, F., Rink, F., & Smilde, M. (2000) 'De studentenrechtbank : een evaluatieonderzoek naar beoordelingslijsten voor presentatievaardigheden. *[*Student court of justice: an evaluation of assessment instruments for presentation skills*], Tijdschrift voor Onderwijsresearch, [Journal for Educational Research], 24 (3-4)*: 265-274.

Elton, L., & Johnston, B. (2002) *Assessment in universities: a critical review of research.* York: Learning and Teaching Support Network Generic Centre.

Falchikov, N. (2005) *Improving assessment through student involvement. Practical solutions for aiding learning in higher and further education.* New York: RoutledgeFalmer.

Fallows S., & Chandramohan, B. (2001) 'Multiple approaches to assessment: reflections on use of tutor, peer and self-assessment', *Teaching in Higher Education, 6(2):* 229-246.

Freeman, M. (1995) 'Peer assessment by groups of group work', *Assessment & Evaluation in Higher Education. 20(3):* 289-301.

Gibbs, G. (2006) 'How assessment frames student learning', in C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education,* pp. 23-36. London: Routledge.

Gielen, S., Dochy, F., & Dierick, S. (2003) 'Evaluating the consequential validity of new modes of assessment: the influence of assessment on learning, including pre-, post-, and true assessment effects', in M. Segers, F. Dochy, & E. Cacallar (Eds). *Optimising new modes of assessment: In search of qualities and standards,* pp.37-54. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Hafner, J., & Hafner, P. (2003) 'Quantitative analysis of the rubric as an assessment tool: an empirical study of peer-group rating', *International Journal of Science Education, 25(12):* 1509-1528.

Hanrahan, S., & Isaacs, G. (2001) 'Assessing self- and peer-assessment: the students' views', *Higher Education Research & Development, 20(1):* 53- 70.

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 Meta-analysis relating to Achievement*. Milton Park, Oxon: Routledge.Hughes, I., & Large, B. (1993) 'Staff and peer-group assessment of oral communication skills', *Studies in Higher Education, 18(3*): 379-385.

Kappe, F. (2008) 'Hoe betrouwbaar is peer-assessment? Twee empirische studies naar studentbeoordelingen. [How reliable is peer-assessment? Two empirical studies about assessment by students]*', Tijdschrift voor Hoger Onderwijs [Journal of Higher Education]. 26(2).*

Kerby, D., & Romine, J. (2009) 'Develop oral presentation skills through accounting curriculum design and course-embedded assessment', *Journal of Education for Business, 85:* 172-179.

Kruger, J., & Dunning, D. (1999) 'Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments', *Journal of Personality and Social Psychology, 77(6):* 1121-1134.

Langan, A., Wheater, C., Shaw, E., Haines, B., Cullen, W., Boyle, J., et al. (2005) 'Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria', *Assessment & Evaluation in Higher Education, 30 (1)*: 21-34.

Langan, M., Shuker, D., Cullen, R., Penney, D., Preziosi, R., & Wheater, P. (2008) 'Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations', *Assessment & Evaluation in Higher Education*, *33(2):* 179-190.

Magin, D., & Helmore, P. (2001) 'Peer and teacher assessments of oral presentation skills: how reliable are they?', *Studies in Higher Education, 26(3):* 287-298.

Messick, S. (1994) 'Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning' Research Report RR 94-95. Princeton: Educational Testing Service.

Miller, P. (2003) 'The effect of scoring criteria specificity on peer and self-assessment' *Assessment & Evaluation in Higher Education, 28(4):* 383-394.

Nicol, D., & Macfarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice', *Studies in Higher Education, 31(2):* 199-218.

Nicol, D., & Milligan, C. (2006) 'Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice' , in C. Bryan & K. Clegg (Eds.). *Innovative assessment in higher education*. pp. 64-77. Taylor and Francis Group, London.

Oldfield, K., & Macalpine, J. (1995) 'Peer and self-assessment at tertiary level: an experiential report', *Assessment & Evaluation in Higher Education, 20(1):* 125-132.

Patri, M. (2002) 'The influence of peer feedback on self- and peer assessment of oral skills', *Language Testing 19(2):* 109-131.

Price, M., & O'Donovan, B. (2006) 'Improving performance through enhancing student understanding of criteria and feedback', in C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education,* pp. 100-109. London: Routledge.

Rust, C., Price, M., & O'Donovan, B. (2003) 'Improving students' learning by developing their understanding of assessment criteria and processes', *Assessment & Evaluation in Higher Education, 28(2):* 147-164.

Sadler, D. (1989) 'Formative assessment and the design of instructional systems', *Instructional Science, 18:* 119-144.

Schunk, D. (2001) 'Social cognitive theory and self-regulated learning',in B. Zimmerman, & D. Schunk (Eds.) *Self-regulated learning and academic achievement. Theoretical perspectives,* pp.125-151. Mahwah, NJ: Lawrence Erlbaum.

Segers, M., Dochy, F., & Cascallar, E. (2003) 'The era of assessment engineering: changing perspectives on teaching and learning and the role of new modes of assessment', in M. Segers, F. Dochy, & E. Cascallar (Eds) *Optimising new modes of assessment: In search of qualities and standards,* pp.1-12. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sellnow, D., & Treinen, K. (2004) 'The role of gender in perceived speaker competence: an analysis of student critiques', *Communication Education, 53(3)*: 286-296.

Shrout, P., & Fleiss, J. (1979) 'Intraclass correlation: uses in assessing rater reliability', *Psychological Studies, 56(2):* 420-428.

Sluijsmans, D. (2002) '*Student involvement in assessment. The training of peer assessment skills',* Unpublished doctoral dissertation, Open University of the Netherlands, Heerlen.

Sluijsmans, D. (2008) 'Towards (quasi-) experimental research on the design of peer assessment', in M. van den Heuvel-Panhuizen, & M. Lacher (Eds.), *Challenging assessment. Book of abstracts of the fourth biennial Earli/Northumbria Assessment Conference,* pp.46. Berlin: Humboldt-Universität.

Sluijsmans, D., Moerkerke, G., Van Merriënboer, J., & Dochy, F. (2001) 'Peer assessment in problem based learning', *Studies in Educational Evaluation, 27:* 153-173.

Struyven, K., Dochy, F., & Janssens, S. (2003) 'Students' perceptions about new modes of assessment in higher education: A review', in M. Segers, F. Dochy, & E. Cascallar (Eds) *Optimising new modes of assessment: in search of qualities and standards,* pp.171-223. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Taras, M. (2008) 'Summative and formative assessment: perceptions and realities', Active Learning in Higher Education, 9(2): 172-192.

Topping, K. (1998) 'Peer assessment between students in colleges and universities', *Review of Educational Research, 68(3)*: 249-276.

Topping, K. (2003) 'Self- and peer assessment in school and university: reliability, validity and utility', in M. Segers, F. Dochy, & E. Cacallar (Eds). *Optimising new modes of assessment: In search of qualities and standards,* pp.55-87. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Topping, K. (2009) 'Peer assessment', *Theory Into Practice, 48:* 20-27.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2009) 'Effective peer assessment processes: research findings and future directions', *Learning and Instruction, 20(4):* 270-279.

Winne, P. (2004) 'Students' calibration of knowledge and learning processes: implications for designing powerful software learning environments', *International Journal of Educational Research, 41:* 466-488.

Yorke, M. (2003) 'Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice' *Higher Education, 45:* 477-501.

Figure 1: Gender interaction effect peer assessors

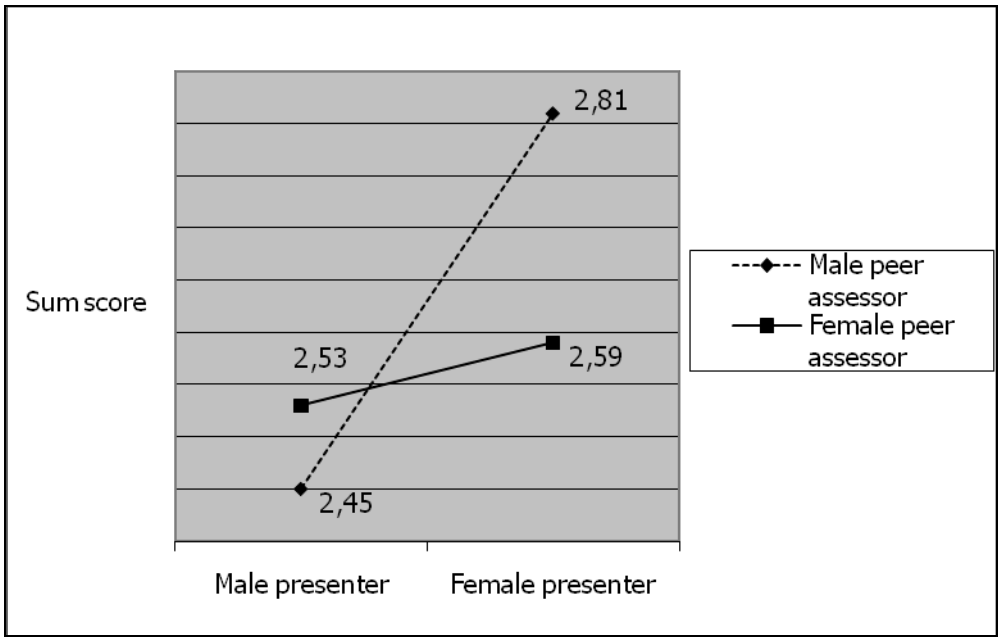

Table 1. Components found on the basis of the principal components analysis, and loadings

Component 1 Content	Component 2 Delivery
Introduction (.72)	Contact audience (.73)
Structure (84)	Enthusiasm (.76)
Conclusion (.77)	Eye contact (.70)
Professionalism (.59)	Vocal delivery (.54)
	Body language (.82)
	Professionalism (.71)

Table 2. Summary of the assessment procedure

Assessor	Presentation 1	Presentation 2	Presentation 3	Total number of assessments	Average number of assessed presentations for one assessor	Number of assessors for one presentation
Teachers	X	X	X	209	41.8	1
Peers	X			174	3.7	6
Presenters	X	X		79	1	1

Table 3. Teacher scores (= Teach.) versus peer assessment (= Peer) scores: descriptives and intra-class correlation (n=29)

	Teach. mean	*Teach. σ*	*Peer mean*	*Peer σ*	*Intra- class correlation*
Sum score	2.14	0.37	2.57	0.32	.45

Table 4. Teacher assessment (Teach.) versus self-assessment (Self): descriptives and intra-class correlation (n=79)

	Teach. mean	*Teach. σ*	*Self mean*	*Self σ*	*Intra- class correlation*
Sum score	2.46	0.53	2.70	0.49	.54