Self- and peer assessment of oral presentation skills in higher education

LUC DE GREZa, MARTIN VALCKEb and IRENE ROOZENA

aDepartment of economics and management, University College Brussels, Brussels, Belgium;
bDepartment of educational studies, University of Ghent, Ghent, Belgium

Abstract

Assessment of oral presentation skills is an under explored area. The present study focuses on the agreement between professional assessment and self- and peer assessment of oral presentation skills and explores student perceptions about peer assessment. Results show that self- and peer assessment result in significantly higher marks as compared to professional assessment. Generalizability analysis revealed that combining assessment scores of four peers already helped to attain a sufficient level of reliability. The results also reflect a very positive attitude of students towards peer assessment as a relevant source of external feedback.

1. Problem statement

Oral presentation skills are recognised as a central professional skill (Campbell, Mothersbaugh, Brammer, & Taylor, 2001). Therefore, higher education should prepare students to develop and master these skills. This study concentrates on the assessment of oral presentation skills but this does not imply that knowledge is not important. We choose for skills because knowing how to present is still no guarantee for showing. We define oral presentation skills in this study as the complex of two clusters of related skills: (1) delivery related skills: eye-contact, vocal delivery, enthusiasm, interaction with
the audience, body-language and (2) content related skills: quality of introduction, structure, and conclusion.

The study focuses on the implementation of an assessment format for oral presentation skills and the reason for this innovation is briefly described hereafter. Traditional assessment approaches consisted of oral feedback provided by faculty, immediately after the delivery of the presentation by the student. Feedback remained short considering the large group of students to be supported ($N = 200$) while delivering three individual oral presentations in small group settings ($N = 16$). On the one hand there was a growing discontent with this time-consuming assessment approach. On the other hand this approach was pressurized, considering recent developments in relation to assessment procedures. In the present context, especially the switch in assessment responsibilities from teacher to students played a crucial role (Segers, Dochy, & Cascallar, 2003). This also builds on the claims of this reorientation as to student involvement and the impact on related performance (Falchikov, 2005). This reorientation inspired us to look for an alternative, student-oriented assessment approach.

1. Theoretical and empirical background

Qualities of assessment

Stressing the learning benefits of assessment leads us to formative assessment and this form of assessment is defined as “assessment that is specifically intended to provide feedback on performance to improve and accelerate learning” (Nicol & Milligan, 2006, p. 64). The quality aspect that checks the influence of the assessment on student behaviour and learning is sometimes called the ‘consequential’ validity (Gielen, Dochy, & Dierick, 2003). This description regards it as a part of validity and this concept is together with reliability traditionally used to evaluate the psychometric quality of assessment. Wiersma and Jurs (2005) define (1) reliability as the consistency of the instrument in measuring what it measures and (2) validity as the extent to which the instrument measures what it is designed to measure.
Involvement of students in assessment can be organised in two ways: peer assessment and self-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.

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

Oral presentation skills are partly learned through observational learning (Bandura, 1997). The latter is in line with the social cognitive perspective towards self-regulated learning (Schunk, 2001). Via observational learning, learners compare their performance or the performance of others with more or less explicit standards. The oral presentation skills will evolve by achieving a better fit between the standards and the current performance level (Sadler, 1989). 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). Both self- and peer assessment help 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 (e.g., Segers et al., 2003). Despite the latter, formative assessment in higher education is still largely controlled by the teachers (Nicol & Macfarlane-Dick, 2006). An accurate calibration of oral presentation performance and the standards suggests that a sufficient level of reliability can be achieved when the same assessment result are attained whether the performance is assessed by a teacher/expert, by peers, or by the learner.

Analysis of the literature about the assessment of oral presentation skills, results especially in an overview of studies about peer- and self-assessment of individual (oral) presentation skills (AlFallay, 2004; Campbell et al., 2001; Cheng & Warren, 2005; Hafner & Hafner, 2003; Hughes & Large, 1993; Langan et al., 2005; Magin & Helmore,
2001; Oldfield & Macalpine, 1995; Patri, 2002; Selinow & Treinen, 2004). In some studies, only part of the research focuses on peer or self-assessment (e.g., Fallows & Chandramohan, 2001). In other cases, group presentations are assessed (e.g., Miller, 2003).

Benefits of self- and peer assessments.

In the different studies, the potential of peer and self-assessment is clearly stressed. Falchikov (2005, p.16) posits that “(...) involving students in the assessment of presentations is extremely beneficial”. She explains that peers involved in assessment have to concentrate more on performance and that they develop the skills of analysis and application of criteria and also develop tact. Cheng and Warren (2005) cite several studies that reported e.g., improved presentation performance due to peer assessment. Other authors adopt in this context videotaped feedback for self-assessments, and also report attainment of better oral presentation skills (Bourhis & Allen, 1998). Topping (1998) dedicates part of his review of the literature about peer assessment to the assessment of oral presentation skills. He summarizes findings of several studies, referring to improvement in marks, perceived higher learning performance, higher confidence (self-efficacy), and the development of appraisal skills. Cooper (2005) adds that the focus on the presentation performance of others can be enhanced by the use of reflection sheets and Shaw (2001) argues that peer assessment connects students with each other in a learning community.

Topping (2003) mentions potential economical benefits as a second purpose of implementing 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 staff workload.

Inter-rater reliability of self- and peer assessments.

There is a considerable debate in the literature about the inter-rater reliability of self- and peer assessment but first a fundamental remark about the research. As indicated by
Topping (2003), it is very common in research about reliability to compare self- and peer assessments to assessment by professionals (teachers). Topping (2003) stresses that the a priori assumption that assessment by a teacher is reliable and valid, can be doubted in some contexts. This assumption relates to a positivist epistemological perspective upon assessment (Elton & Johnston, 2002) that believes in an objective, scientific measurement. It is therefore remarkable that no research was found that tested this assumption, and that the use of professional assessments is in many cases not approached with due caution as asked by some authors (e.g., MacAlpine, 1999).

But let us take a closer look at the literature about assessment of oral presentation skills. Freeman (1995) concludes in his study that there is no significant difference in the overall mark averages given by peers or given by professional assessors. In contrast, Langan et al. (2005) report that peer marks are on average 5% higher than marking by their tutors. Other studies based on correlations, conclude that peer assessment can be a relevant substitute for assessments by professionals (AlFallay, 2004; Campbell et al., 2001; Hughes & Large, 1993; Oldfield & Macalpine, 1995; Patri, 2002). Nevertheless, Hughes and Large (1993) warn that a high correlation between marks of peers and professionals can still hide a considerable variation in the marks given by students and professionals. Freeman (1995) reports e.g., only moderate correlations between peer and professional scores. He also reports that the standard deviation for peers was half that of the professionals (see also Hughes & Large, 1993). Cheng and Warren (2005) add to this that student mean markings are within one standard deviation of teacher markings, but they point out that students did not assess the same elements as their teachers did. 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.

In the literature, fewer studies are found that compare professional assessment with self-assessment of oral presentation skills. It is clear that results are not univocal. Some of these studies report lower correlations between self- and professional assessments than between professional and peer assessment (Campbell et al, 2001; Patri, 2002). Nevertheless, other authors consider self-assessments to be as valid as peer assessment (AlFallay, 2004; Hafner & Hafner, 2003).
The latter study adopted a generalizability analysis to study inter-rater reliability. Generalizability analysis allows disentangling measurement error into multiple error sources; in contrast to classical test theory that decomposes an observed score into a true score and error (Brennan, 2000). The results of a generalizability analysis in the study of Hafner and Hafner (2003) helped to conclude that approximately one-quarter of the total variance is due to the actual differences in oral presentations, another one-quarter is accounted for by differences between raters and approximately one-half of the total variance is residual error. In addition, generalizability analysis makes it possible to determine the number of peer raters needed to obtain reliable scores. In their study, Hafner and Hafner (2003) found a sufficient generalizability (0.80) when 10 peers rated the presentation. A strong increase in generalizability score was observed when moving from a single rater to five raters. Wood, Marks, & Jarbour (2005) needed 8 peers for a reliability of .80.

Variables affecting the quality of self- and peer assessments

In the context of peer assessment, authors discuss especially rating errors and the importance of student perceptions. Rating errors are central in the study of Sluijsmans, Moerkerke, van Merrienboer, and Dochy (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 also stated to have a considerable influence on student learning (Struyven, Dochy, & Janssens, 2003). Concerns have been raised about resulting difficulties in peer assessment contexts by Hanrahan and Isaacs (2001). Their analysis revealed eight general dimensions and twenty higher order themes, exemplified below. Results showed e.g., that students were concerned about their inexperience with marking, that they felt uncomfortable critiquing others’ work and remarked that the process was not taken seriously because it doesn’t count for marks. Students also complained about the time-consuming nature of the activity and asked feedback as to their involvement in the assessment (Hanrahan & Isaacs, 2001).
As to self-assessment, a meta-analysis of Falchikov (2005) indicates that some—but not all—students are able to assess in similar ways as compared to teachers. This is confirmed in e.g., the study of Kruger and Dunning (1999) where novices and low performers overestimate their performance level and even lack related metacognitive abilities (monitoring, evaluation). Rust et al. (2003) come to the conclusion that women are more likely to underestimate their performance, whereas males tend to overestimate the quality of their performance in a self-assessment context.

But let us take a closer look at the literature about assessment of oral presentation skills. 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. They 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 who reported collaboration over marks and mark fixing in the assessment by peers of seminar presentations. However, Sellnow and Treinen (2004) report that neither the gender of the presenter, nor the gender of the assessor, did affect overall peer ratings.

Only a small amount 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 that peer assessment.

To improve the quality of self- and peer assessments of oral presentation skills, evidence from the research literature is not univocal. Research focused on the value of training in assessment, initial discussions (about the assessment criteria or about the fact that students can perform assessments), and the length of the scoring criteria list.

Hafner and Hafner (2003) state that instruction and training is not sufficient. In contrast, Carlson and Smith-Howell (1995) hardly found differences in assessment practices between untrained and trained professionals. 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. The enhancement of low self-efficacy for peer assessment was a key point of the debate with students in the case study of Fallows and Chandramohan (2001).

Miller (2003) came to the conclusion that more items in the evaluation checklist resulted in an increased variance in scores. This could diminish inter-rater reliability but provide students with more detailed and thus better feedback. In contrast, Freeman (1995) suggests to reduce the number of criteria in the checklist, but this could diminish the quality of feedback generated by the assessment. Lievens and Conway (2001) make a large-scale evaluation of multitrait-multimethod studies in assessment centres and use the cognitive load theory to explain the effectiveness of assessment instruments where the relevant behaviours are listed. These relevant behaviours could reduce the number of inferences required from the assessors because they can use the listed behaviours as retrieval cues to guide the recall of observed behaviour and assessors also don’t need to categorize the behaviour (Lievens & Conway, 2001).

Concluding we can state that many questions about self- and peer assessment of oral presentation skills remain unanswered. Nevertheless if we can’t reassure teachers on these matters and they feel worried about these drawbacks, than this will influence their willingness to adopt these alternative evaluation approaches (Falchikov, 2005). The problem seems to be the lack of research, and certainly the lack of (quasi-) experimental research, to fill the gap between what we know and what we claim about peer assessment as Sluijsmans (2008) recently stated.

In general, the available research on peer and self-assessment of oral presentation skills shows underexplored areas at the one hand and divergent views on the other hand. It is clear that more research is needed in this field.

3. Research Questions
The inconsistencies in the research findings about self- and peer assessment of oral presentation skills put forward a clear agenda for future research. In the context of the present study, we centre on the following key research questions:

- What is the level of agreement between peer assessments and professional assessments?
- What is the level of agreement between self-assessments and professional assessments?
- What are the student perceptions about peer assessment?

4. Research Design

Participants

The participants involved in the study were university freshman enrolled for a Business Administration introductory course about psychology (age average 18 years). 73 participants took part at the start of the study. Only 57 of them finished all phases of the study (36 male). Reasons for drop-out were not systematic and rather related to illness, incompatibility of rosters, or internships of the students. 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 – based on the analysis of oral presentation literature - a rubric was constructed consisting of nine 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). Assessors are asked to rate the quality of a presentation according to these criteria on a 5 point Likert scale. Descriptors and
indicators are provided as a help to direct the assessment process. A factor analysis of the instrument indicated underpinned the validity and reliability of the instrument (see De Grez, Valcke, and Roozen, 2006; De Grez et al., in press).

As an example, we describe the assessment related to the criterion “quality of the introduction”. Assessors are invited to use a set of 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.

Building on their judgement, their score reflects the extent to which the quality of the introduction meets none, one, or more of the indicators put forward.

**Perception of ‘peer assessment’**

A subscale focusing on “perceptions of peer assessment” was adopted from the questionnaire developed and validated by Sluijsmans (2002). The original subscale consists of 7 items, with an alpha reliability coefficient of .74. One item was omitted from the scale, and a few words were changed in order to adapt the subscale to the specific oral presentation situation (e.g., ‘You can learn from the feedback of peers’). The scale was presented after the first and after the third oral presentation.

**Professional and peer assessors**

The recorded oral presentations were assessed – on the base of the assessment rubric - by five assessors (3 female). Four of these assessors were faculty members with at least 5 years of a language teaching background who did not teach the students to be assessed. The fifth assessor was a junior-researcher. These assessors and their assessments are labelled as *professional* in this article. Next to the professional assessors, 47 students were involved as peer assessors in the study. These students were enrolled in the second year Business Administration (32 male) and participated in the study as a formal part of a course about communication skills. Both the professional and the peer assessors were unaware of the nature of the research questions. All the professional assessors received a
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.

Procedure

Participants were – as a formal part of their psychology course - invited to deliver three short oral presentations about a prescribed topic. All the presentations were recorded. Due to drop-out of a number of participants for the second or third presentation, the final number of recorded recordings of oral presentations was 209 instead of 219.

After the first presentation, students participated individually in a computer-based multimedia training programme about oral presentations (see De Grez et al., submitted). After the second presentation students received feedback, based on the assessment rubric about their performance on the first presentation. For one third of the participants, this feedback was based on the self-assessment of their first presentation. The other participants received feedback either from peers or from professionals, based on the quantitative scoring of the nine assessment criteria (see below).

Assessors and the assessment procedure

The evaluation of the oral presentations - both for professional 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. Recordings were assigned at random to assessors.

Professional assessors evaluated individually the 209 recorded oral presentations. For each oral presentation, scores were determined for the 9 criteria in the rubric. Each professional assessor evaluated between 34 to 49 oral presentations.

Student peers assessed 29 presentations. Each of these 29 presentations was assessed by six different peers. This specific number is based on the work of Hafner and Hafner (2003) who reported a large improvement in generalizability from a single rater to
about five raters and on the work of Dannefer et al. (2005) who concluded that six peers were needed to achieve a moderate generalizability in assessing professional competence. As a result, in total 174 peer evaluations have been carried out.

As part of the research design, one third of the participants was asked to rate their own presentation with the assessment rubric. The assessment rubric was at the base of the design of multimedia instruction package. Therefore, we can assume that these students were also well acquainted with the rubric criteria in view of the self-assessment activity.

5. Research Results

Initial analyses

Before the research data were analysed in view of the research questions, quality control of the assessment process was carried out. This focused on uncontrolled differences in the way professional assessors applied the assessment rubric, despite the random distribution of recorded presentations to the professionals. The latter implies that no significant differences in average scores are expected. Analysis of variance was applied to test differences. Post hoc comparisons confirm that professional assessors do not differ significantly in applying the rubric criteria Introduction, Structure, and Contact with audience. But significant differences are observed in view of the other six criteria. Additional analysis reveals that - for five of the six criteria – it is consistently the same professional assessor that adopted a more lenient view as compared to the other assessors. An implication of the assessment bias - that could result from this finding - could have been that the mean evaluation scores obtained for the first, second, and third presentation are significantly different when we carry out the analysis with or without the scores of the too lenient assessor. Carrying out a general linear model analysis of the differences in evaluation scores between the first, second, and third presentation results in both cases in a comparable pattern. In both cases there is a significant progress from presentation one to presentation two and a non significant progress from presentation two to three.

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

*What is the level of agreement between peer assessments and professional assessments?*

After calculating the sum score of the nine rubric criteria ($\alpha = .85$), the score for each of the nine criteria and the sum score of the criteria were compared between professional- and peer assessors. Table 4.1 summarizes the analysis results.

Table 4.1
Professional assessment scores (= Prof.) versus peer assessment (= Peer) scores: descriptives, $t$-tests and correlation analysis ($n=29$)

<table>
<thead>
<tr>
<th>Assessment criteria</th>
<th>Professional mean (Standard deviation)</th>
<th>Peer mean (Standard deviation)</th>
<th>$t$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1.72 (0.80)</td>
<td>2.01 (0.55)</td>
<td>2.10*</td>
<td>.44*</td>
</tr>
<tr>
<td>Structure</td>
<td>2.07 (0.81)</td>
<td>2.41 (0.37)</td>
<td>2.32</td>
<td>.32</td>
</tr>
<tr>
<td>Conclusion</td>
<td>1.25 (0.44)</td>
<td>1.68 (0.39)</td>
<td>4.29**</td>
<td>.18</td>
</tr>
<tr>
<td>Interaction audience</td>
<td>2.59 (0.68)</td>
<td>2.89 (0.35)</td>
<td>2.67*</td>
<td>.45*</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>2.24 (0.64)</td>
<td>2.86 (0.60)</td>
<td>4.64**</td>
<td>.32</td>
</tr>
<tr>
<td>Eye contact</td>
<td>2.45 (0.87)</td>
<td>2.99 (0.57)</td>
<td>3.15**</td>
<td>.24</td>
</tr>
<tr>
<td>Vocal delivery</td>
<td>2.90 (0.67)</td>
<td>3.25 (0.44)</td>
<td>3.00**</td>
<td>.39*</td>
</tr>
<tr>
<td>Body language</td>
<td>1.79 (0.86)</td>
<td>2.37 (0.52)</td>
<td>3.26**</td>
<td>.10</td>
</tr>
<tr>
<td>Professionalism</td>
<td>2.10 (0.56)</td>
<td>2.61 (0.39)</td>
<td>5.56**</td>
<td>.52**</td>
</tr>
<tr>
<td>Sum score</td>
<td>2.14 (0.37)</td>
<td>2.57 (0.32)</td>
<td>6.21**</td>
<td>.45*</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$
The analysis results indicate that a positive— but not consistently significant - correlation can be observed between professional and peer assessment scores. The rubric sum score of professional assessments is significantly lower than peer assessments ($t=6.21; p< .01$). Also, for eight rubric criteria we can observe significantly lower scores from the professional assessors.

Above, analysis results were discussed about the quality of the professional assessors. Below, we report the results of an analysis focusing on the inter-peer agreement on the basis of a generalizability analysis. A two-facet generalizability study was carried out by adopting the analysis procedure of Mushquash and O’Connor (2006). The generalizability coefficient indicates the reliability pending the number of peer assessment scores and the amount of criteria used in the rubric. As not all the peers assessed all the oral presentations, the data are nested. The analysis of the variance components – as summarized in Table 4.2 - shows that the variance in scores related to the oral presentations is low (9.7 % of the total variance). The variance component for peers (21% of total variance) is large and the component for peers by criteria (7% of total variance) and peers by participants (3% of total variance) are moderate to small. Finally there is a zero variance component for the rubric criteria and a large residual variance (58.8% of the total variance).

Table 4.2
Estimation of variance components

<table>
<thead>
<tr>
<th>Source of variance</th>
<th>df</th>
<th>SS$^a$</th>
<th>MS$^b$</th>
<th>Variance</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral presentation scores</td>
<td>28</td>
<td>148,99</td>
<td>5,32</td>
<td>0,85</td>
<td>9.7</td>
</tr>
<tr>
<td>Peers</td>
<td>5</td>
<td>254,81</td>
<td>50,96</td>
<td>0,185</td>
<td>21.1</td>
</tr>
<tr>
<td>Peers x Criteria</td>
<td>48</td>
<td>115,20</td>
<td>2,40</td>
<td>0,065</td>
<td>7.4</td>
</tr>
<tr>
<td>Peers/participants</td>
<td>140</td>
<td>105,10</td>
<td>0,75</td>
<td>0,026</td>
<td>3.0</td>
</tr>
<tr>
<td>Residu</td>
<td>1344</td>
<td>693,02</td>
<td>0,52</td>
<td>0,516</td>
<td>58.8</td>
</tr>
</tbody>
</table>

$^a$ SS = Sum of Squares; $^b$ MS = Mean Square
In Table 4.3 estimation is given of the generalizability coefficients. The generalizability coefficient for the nine criteria and the six peers, points at a good reliability. The relative $G$-coefficient is equal to .85. This is higher than the $G > .80$ criterion for reliability (Mushquash & O’Connor, 2006). This cut off criterion is already reached when nine rubric criteria are applied by four different peers ($G = .80$). Alternatively, when six peers are involved, we are allowed to reduce the rubric to six criteria ($G = .81$).

Table 4.3
Estimation of the generalizability coefficients ($G$), based on the number of peers and amount of criteria

<table>
<thead>
<tr>
<th>Number of peers</th>
<th>Criteria 1</th>
<th>Criteria 2</th>
<th>Criteria 3</th>
<th>Criteria 4</th>
<th>Criteria 5</th>
<th>Criteria 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.13</td>
<td>.23</td>
<td>.31</td>
<td>.38</td>
<td>.43</td>
<td>.48</td>
</tr>
<tr>
<td>2</td>
<td>.23</td>
<td>.37</td>
<td>.47</td>
<td>.54</td>
<td>.59</td>
<td>.64</td>
</tr>
<tr>
<td>3</td>
<td>.29</td>
<td>.46</td>
<td>.56</td>
<td>.63</td>
<td>.68</td>
<td>.71</td>
</tr>
<tr>
<td>4</td>
<td>.35</td>
<td>.52</td>
<td>.62</td>
<td>.68</td>
<td>.73</td>
<td>.76</td>
</tr>
<tr>
<td>5</td>
<td>.39</td>
<td>.56</td>
<td>.66</td>
<td>.72</td>
<td>.76</td>
<td>.79</td>
</tr>
<tr>
<td>6</td>
<td>.43</td>
<td>.60</td>
<td>.69</td>
<td>.75</td>
<td>.79</td>
<td>.81</td>
</tr>
<tr>
<td>7</td>
<td>.45</td>
<td>.62</td>
<td>.71</td>
<td>.77</td>
<td>.80</td>
<td>.83</td>
</tr>
<tr>
<td>8</td>
<td>.48</td>
<td>.65</td>
<td>.73</td>
<td>.78</td>
<td>.82</td>
<td>.84</td>
</tr>
<tr>
<td>9</td>
<td>.50</td>
<td>.67</td>
<td>.75</td>
<td>.80</td>
<td>.83</td>
<td>.85</td>
</tr>
</tbody>
</table>

To detect possible gender effects, a two-way ANOVA was carried out with gender of the assessor and the gender of the assessed student as independent variables and the oral presentation skills sum score as the dependent variable. This was done separately for professional and for peer assessors. Results indicate that gender of the professional assessors ($F(1,205)= .03, p=.87$) and of the peer assessors ($F(1,170)= .85, p=.36$) did not
have a significant impact. The interaction effect gender of the assessor and gender of the assessed was not significant for professional assessors ($F(1,205) = .2.91, p=.09$) but was significant for peers ($F(1,170) = .4.17, p=.04$). Male peers give female presenters significantly higher scores than male presenters, but female peers do not make such a difference.

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

In view of this research question, the scoring results are compared of professional assessors and the self-assessment by students. Table 4.4 summarizes the analysis results. Overall, we find a positive – but not consistently significant - correlation between professional and self-assessment scores.

Table 4.4
Professional assessment (Prof.) versus self-assessment (Self):

descriptives, $t$-tests and correlation analysis ($n=79$)

<table>
<thead>
<tr>
<th>Assessment criteria</th>
<th>Professional mean (standard deviation)</th>
<th>Self mean (standard deviation)</th>
<th>$t$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Professional mean (standard deviation)</td>
<td>Self mean (standard deviation)</td>
<td>$t$</td>
<td>$r$</td>
</tr>
<tr>
<td>Introduction</td>
<td>2.30 (0.69)</td>
<td>2.71 (0.92)</td>
<td>4.67**</td>
<td>.56**</td>
</tr>
<tr>
<td>Structure</td>
<td>2.57 (0.94)</td>
<td>2.99 (0.96)</td>
<td>3.55**</td>
<td>.40**</td>
</tr>
<tr>
<td>Conclusion</td>
<td>1.91 (0.74)</td>
<td>2.21 (0.94)</td>
<td>3.20**</td>
<td>.53**</td>
</tr>
<tr>
<td>Interaction audience</td>
<td>2.72 (0.76)</td>
<td>2.62 (0.81)</td>
<td>0.88</td>
<td>.18</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>2.43 (0.79)</td>
<td>2.66 (0.80)</td>
<td>2.19*</td>
<td>.33**</td>
</tr>
<tr>
<td>Eye contact</td>
<td>2.72 (0.82)</td>
<td>3.15 (0.91)</td>
<td>3.51**</td>
<td>.19</td>
</tr>
<tr>
<td>Vocal delivery</td>
<td>3.07 (0.81)</td>
<td>3.03 (0.72)</td>
<td>0.43</td>
<td>.38**</td>
</tr>
<tr>
<td>Body language</td>
<td>2.09 (0.93)</td>
<td>2.47 (0.78)</td>
<td>3.07**</td>
<td>.18</td>
</tr>
<tr>
<td>Professionalism</td>
<td>2.47 (0.67)</td>
<td>2.57 (0.69)</td>
<td>1.31</td>
<td>.51**</td>
</tr>
<tr>
<td>Sum score</td>
<td>2.46 (0.53)</td>
<td>2.70 (0.49)</td>
<td>4.13**</td>
<td>.53**</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$
The ‘total’ rubric score of professional assessments is significantly lower than self assessments ($t = 4.13; p < .01$). With the exception of two criteria (Interaction with the audience and Vocal delivery), the same applies to the scoring of specific rubric criteria.

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?**

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

### 6. Discussion and conclusions

In the present study, alternative assessment approaches were studied. In this context, self- and peer assessment were positioned within a social cognitive perspective on self regulated learning. The limited -and often contradictory- available empirical evidence about self- and peer assessment of oral presentation skills in learning prompted the design of a study in which the psychometric quality of self- and peer assessment was contrasted to the assessment by professionals.

Comparison of the professional and peer assessment rubric scores points at a positive relationship, but also at critical differences. The positive and significant correlation of $\text{.45}$ is comparable to indices found by Cheng and Warren (2005) and by Patri (2002), but lower than the $\text{.83}$ correlation values reported by Hughes and Large (1993). The rather low correlation in the present study might suggest that peers and professionals interpreted the criteria and indicators of the rubric in a different way. This can be explained by differences in the wide and depth of their respective experiential base. Also, within the group of peers not all peers could have applied the same criteria in a comparable and/or consistent way. The inconsistency is suggested by the large proportion of the variance that is related to peers in the generalizability analysis results, and the small proportion of the total variance to be attributed to the quality of the oral presentations. But, this suggestion is in conflict with the observation of a zero variance
component for the assessment criteria in the rubric. This points at a strong internal consistency. The large residual variance, found in the present study, is comparable to the results reported by Hafner and Hafner (2005). Our results suggest that the combined scores of four peers can partly compensate for the differences. It is important to note that “four peer assessors” is a significant reduction in the number of assessors required to attain the $G$ criterion for reliability. Alternatively, the analysis results also suggest that we can reduce the number of evaluation criteria in the rubric. This is however not an option, because we want – as explained in our theoretical base - to provide learners with as much feedback as possible. Lastly, the finding that peers produce higher marks as compared to professional assessors, is in agreement with other studies (e.g., Langan et al., 2005).

With regard to the comparison of self-assessment scores and professional assessment scores, we can again conclude that there is a level of agreement and disagreement when assessing the oral presentations. Nevertheless, the high and significant correlation between both assessment scores ($r = .53$), is higher than values reported in other studies (e.g., AlFallay, 2004). Also the finding that the self-assessment scores are – nearly always - higher marks than professional marks, is confirmed by other studies (e.g., Patri, 2002).

As explained above, these nearly consistent differences can be explained by the more detailed and broader experience of professionals with oral presentations. They can retrieve from their memory a larger set of models that exemplify how oral presentation 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.

With regard to the research question focusing on student perceptions of peer assessment, it can be stated that the results reflect a very positive attitude towards the value of peer assessment. In addition, having experienced self- and or peer-assessment affects this perception again in a positive way. This positive perception towards peer assessment of oral presentation skills is a promising finding in the light of the impact of perceptions on student learning (Struyven et al., 2003).

The present study has the merit of paying attention to the inter-rater reliability of the professional assessors. As explained above, one of the assessors applied a number of the criteria in a more lenient way. Statistical solutions were found to tackle this problem.
in the present study. Nevertheless, in a normal instructional setting, teachers have to be aware of the bias caused by assessors approaching the criteria in diverse ways. This should also be considered when setting up assessment related research (Topping, 2003).

Gender was also studied as a potential source of bias. Gender of the assessor nor the student being assessed seems to influence the assessment process or assessment marks. This is in concordance with the findings of Sellnow and Treinen (2004), but different from what was reported by others (Edens, Rink, & Smilde, 2000; Langan et al., 2005). Analysis of the inter-rater reliability also reveals that specific assessment criteria pose hardly a problem, whereas other criteria more easily lead to disagreement between types of assessors.

These findings suggest that the training of assessors should take this into account, and more examples and or more concrete indicators should be provided to them, and this is consistent with the point of view of several authors (e.g., Sluijsmans, 2002).

Though a more in-depth analysis of the inter-rater reliability of professional, peer, and self-assessment is beyond the scope of the present study, we have to keep in mind that the requirement to guarantee a high level of reliability is not always met (see also e.g., Price & O’Donovan, 2006; Topping, 2003). The results do not suggest that we should stop applying self- and peer assessment of oral presentation skills. Also Langan et al. (2005) and Sluijsmans (2002) make it clear that the benefits of peer-assessment outweigh a certain degree of discrepancy between e.g., student marks, tutor marks, and peer markings. Boud (2007) refers in this context to the “consequential” validity of assessment. 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 authors, such as Winne (2004), stress the importance of the accuracy of feedback in view of future learning outcomes. But other authors, such as Gibbs (2006) and Yorke (2003), state that not the quality of the feedback evolving from the assessment is crucial, but the actual engagement of the student in a feedback generating process. In our opinion, we have to combine the views of the different authors. On the one hand, we do not want students to take the wrong actions based on low quality feedback. On the other hand we also don’t want to block the future behaviour that should evolve from feedback. The question is therefore especially how to improve the quality of self and peer assessment.
assessment approaches. Falchikov (2005) recommends developing evaluation criteria in 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 authors rather 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 be explained on the base of the latter. 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 that essential.

Although a large amount of recorded oral presentation sessions were assessed by peers, professional assessors, 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, the complexity level of the competencies, etc. These limitations are important when we consider e.g., the findings of Gao and Brennan (2001) that estimated variance components can vary from one generalizability study to another, depending on the sample sizes.

Our study revealed some interesting results about an – until now - under-explored instruction and assessment field. Additional research could focus on the impact of assessment training, student collaboration in relation to defining assessment criteria etc. Future studies should also consider the nature of the target audience that could vary in knowledge domains and expertise levels. Further research should investigate the short term, middle term and long term effects. In this context, the relationship between self- and peer assessment and our theoretical framework about self regulated learning will become more obvious.

References


De Grez, L., Valcke, M., Roozen, I. (in press). The impact of goal orientation, self-reflection, and personal characteristics on the acquisition of oral presentations skills. Accepted for publication in the *European Journal for the Psychology of Education*.


