



## PEER ASSESSMENT IN PROBLEM BASED LEARNING

Dominique M. A. Sluijsmans\*, George Moerkerke\*,  
Jeroen J. G. van Merriënboer\* and Filip J. R. C. Dochy\*\*

*\*Open University of the Netherlands, Heerlen, The Netherlands*

*\*\*University of Leuven, Belgium;*

*\*\*University of Maastricht, The Netherlands*

### Introduction

In many institutions for higher education problem based learning has become the educational concept. The aim of problem based learning is to improve students' ability to work in a team to solve new, complex and ill-structured real-life problems, showing their co-ordinated abilities to access information and turn it into viable knowledge. Knowledge, then, would not be something possessed only for the learners' own sake, but rather something accessed and constructed when needed to solve a problem or design something useful (Segers & Dochy, 1999). These key elements should be transferred to the design of assessment. However, the tests in problem based learning are not always in line with the goals of problem based learning. Teachers often develop assessments that test content knowledge, rather than areas like self-directed learning, problem solving and skills as a group member. Progress tests for example, are applied in medical problem based learning. These are multiple-choice tests with true/false questions about all content areas of a specific profession (Van der Vleuten, Verwijnen, & Wijnen, 1996). Although Van der Vleuten et al. claim that progress tests are in line with the curricular goals of problem based learning, the danger still exists that students develop "test behavior": they only invest in what is required in the assessment (Lockwood, 1995). And if this is content knowledge, they only learn content knowledge. In these cases, assessment is not congruent with instruction, since the goal is to go beyond memorisation. New forms of assessment, such as Overall tests (Segers, 1997) and peer- and co-assessment (Sluijsmans, Dochy, & Moerkerke, 1999) provide such possibilities.

It is important to plead for an assessment system that requires students to use higher-order thinking skills to solve and analyse problems instead of memorising facts

and solving well structured, decontextualized problems. Two of these higher-order skills, which are important in professional organisations, are that students be able to reflect on their own behaviour (self assessment) and that of their peers (peer assessment). The assumption that self and peer assessment are important skills in order to work on complex problems is widely acknowledged in education (e.g., Birenbaum & Dochy, 1996; Boud, 1995; Sambell & McDowell, 1998). Problem-based learning should moreover occur in a clear operationalization of a constructivist learning environment, characterised by co-operative learning and self-directed learning. In such a learning environment the responsibility for the learning process is partly given to the student. In order to enlarge the educational congruence, students should also receive responsibility in the assessment. Peer-assessment provides such an opportunity (Dochy, Segers, & Sluijsmans, 1999).

In problem based learning students work in groups. The size of the group varies from 7 to 14. When students work together as a team on a particular problem, each student has to take his or her responsibility for a certain part of the task. Tutors often find it difficult to determine what each individual has contributed to the group product. Introducing peer assessment can be a way to force students to take the responsibility to make a judgement about the actual contribution of each of their peers in the group discussion.

Prior analysis of 62 studies showed that self and peer assessment can be effective tools to develop the skills needed in the working field (Sluijsmans, Dochy, & Moerkerke, 1999). But assessing one's process or product is not a simple task. Because students often are novices in assessing the work of a peer, rating errors can occur. A number of these errors can be identified, five of which are well documented. Although these errors are described in general terms, we can conclude that these errors are also applicable to peer assessment.

First, there are many personal differences among raters in their standards and their rating styles (Coffman, 1971; De Groot, 1975). Raters may differ in their severity or leniency. Some raters consistently tend to give high grades (lenient raters), while others consistently tend to give low grades (severe raters; see also Lunz, Wright, & Linacre, 1990). Second, raters differ in the extent to which they distribute grades on the score scale. Some raters tend to distribute scores closely around their average; others will spread scores much more widely. In other words, some raters avoid giving extreme grades while others prefer to use them. A third effect is the so-called halo effect. This is the tendency of human raters to base distinctive aspects of the rating on an overall impression created by one single dominating aspect. This may indicate that raters cannot differentiate among distinct aspects of one product or procedure (Borman, 1975). Fourth, the significant effect refers to the fact that raters may have different opinions about the rating tasks. According to Voss and Post (1990), this problem is not so much related to the divergent views of an individual, but rather to the diverging opinions of groups of individuals. Voss and Post argue that in particular in the assessment of "soft" or less "tangible" skills, objectivity is significantly decreased due to divergence of views among

raters of different schools. The fifth and last rating error is caused by so-called evaluation policy. Judges differ in the ways they employ criteria (Sadler, 1983). Every assessor has his or her own evaluation policy. According to some the performance must achieve a minimum qualifying level on a number of criteria. Other judges act conjunctively: While the performance is excellent on one criterion, it is weak on the rest of the criteria. One could also judge compensatorily: Poor showings on some criteria could be balanced by high performance on others.

Based on the rationale for introducing peer assessment in a problem based learning context, two exploratory studies were conducted to find answers to the following research questions:

1. Are peer ratings in problem based learning groups reliable?
2. Do students have idiosyncratic (i.e., personal) strategies in peer assessment?
3. What are students' experiences with peer assessment and problem based learning?

Studies I and II below describe projects in which the students themselves assessed the work process of each of their peers, while the product was assessed by the tutor.

### Study I

#### *Method*

##### *Participants*

The population consisted of 27 university students (9 male, 18 female) who were enrolled in a four-year course in educational sciences using problem based learning. The students were randomly distributed amongst two groups – Group I ( $N=13$ ) and Group II ( $N=14$ ). Twenty students graduated the first licentiate in educational sciences with satisfaction (approximately "B") and seven graduated with distinction (approximately "A"); 23 of the students entered university directly after secondary education, and 4 students had first been enrolled in higher vocational education before attending the university.

##### *Materials*

At the end of a predefined period, all students assessed the peers in their own group on four criteria, which were explained in detail on a peer assessment form. The criteria were defined by the students in negotiation with the tutor. These criteria were: (1) contribution to the group discussions, (2) quality of the contributions, (3) preparedness to be involved in tasks, and (4) actual contribution to the teamwork. Peers scored on a scale varying from *better than the group* (3), *mean of the group* (2), *slightly below the mean of the group* (1), *no help for the group* (0) to *hindrance for the group* (-1). This scaling was based on a comparable scaling method used by Boud (1995), with a positive contribution to the group yielding positive scores, and a negative contribution to the group yielding negative scores.

---

A two-part evaluation questionnaire was developed. The first part consisted of 28 closed items (5-point Likert scale) about different aspects of problem based learning, such as working in a team, problem solving, the learning process and the role of the tutor. These 28 items were reduced to four variables: the satisfaction of working in a group, the achievement of the goals of problem based learning, the instructional process and the role of the tutor. The second part included eleven items about peer assessment, seven yes/no-items and four open-ended questions.

### *Procedure*

The two groups worked for four consecutive periods of six weeks. Each period had several concrete, content-related goals, such as understanding different teaching and learning methods and being able to apply several alternative assessment tools. In each period the groups received one or more problem tasks which had to be solved in the group. At the end of each of the four periods, students had to report how they had solved the problem. Students shared the work load for this report and organized meetings in order to be able to make it a real group result.

The four peer assessment criteria were defined by the students in negotiation with the tutor in weekly two-hour discussion meetings. At the end of the fourth period, a session was organized for conducting the peer assessment and filling out the evaluation questionnaires. In Group I, for example, each student gave scores to his or her twelve peers.

In this study, the peer assessment score was a part of the final score. The tutor also rated each student on the four criteria.

### *Data Analysis*

In order to examine whether the peer ratings in problem based learning groups are reliable, the data of the peer assessments were analyzed. The reliability of the ratings was estimated within the framework of generalizability theory. This theory provides a mechanism for disentangling the error term into multiple sources (Brennan, 1995). Through generalizability analysis, the relative magnitude of variance caused by persons, raters, criteria, and their interaction can be estimated. In contrast to classical test theory, which treats only one error source at a time (e.g., inter-rater reliability or test-retest reliability), generalizability recognizes that there may be multiple sources of error variance which determine how accurately observed scores allow us to generalize about raters' behavior in a universe of situations. The *person variance* is an estimate of the variance across person's mean scores, where the mean is taken across all criteria and raters. The *criteria component* is the estimated variance of criteria mean scores, where each mean is taken across all persons and raters. The *rater component* is the variance of rater mean scores, where each mean is taken across all persons and criteria.

Decisions about students will generally not be based on the results of a single scoring of a single task. Important individual scores of a student will be based on average

scores over multiple criteria and/or raters. In a so-called decision study or D-study the reliability of the scores can be estimated on the basis of the variance components (Brennan, 1983; Feldt & Brennan, 1989). A decision study is designed to identify the number of raters that would be required to obtain acceptably small error variances or acceptably large reliability coefficients.

To investigate if raters have personal strategies a Q-analysis was applied. Using the so-called Q-analysis or profile analysis, it is possible to determine the similarities and differences among raters (McKeown & Thomas, 1988; Tucker, 1962). Our Q-analysis used inter-rater correlations as similarity measures. The aim of the analysis of ratings was to verify whether there were (groups of) students with idiosyncratic strategies on the peer assessment task. For each of the two groups, correlations were calculated. Each correlation matrix was analyzed with Principal Component Analysis using SPSS (1997).

The evaluation questionnaires were analyzed to measure the third research question about students' experiences with peer assessment and problem based learning. Descriptives were calculated for the four variables concerning problem based learning. Frequencies were calculated for the seven yes/no-questions about the peer assessment. The answers to the four open-ended questions about peer assessment were analyzed qualitatively and reduced to categories.

### Results

*Are peer ratings in problem based learning groups reliable?* In Tables 1 and 2 the estimated variance components are shown for the student ratings for Group I and Group II. A positive sign regarding the results of Group I is the fact that the largest variance in the scores was related to the performance of *persons* (40%). The variance related to the *raters* was relatively large (*raters*: 5%; *persons\*raters*: 11%; *raters\*criteria*: 3%) compared to other studies.

Table 1: Estimation of Variance Components Persons, Raters and Criteria of Group I

Source of variance	Sum of Squares	<i>df</i>	Means square	$\sigma^2$	% total variance
<i>Persons</i> (P)	178.531			0.232	
<i>Raters</i> (R)	32.352			0.032	
<i>Criteria</i> (C)	0.449			0.000	
<i>Persons*raters</i>	80.077	169	0.474	0.065	11
<i>Persons*criteria</i>	18.122	39	0.465	0.018	3
<i>Raters*criteria</i>	17.872	39	0.458	0.017	3
Error	108.556	507	0.214	0.214	37

Table 2: Estimation of Variance Components Persons, Raters and Criteria of Group II

Source of variance	Sum of Squares	df	Means square	$\sigma^2$	% total variance
Persons (P)	114.882	12	9.569	0.1429	15
Raters (R)	35.515	12	2.960	0.0399	4
Criteria (C)	2.822	3	0.941	0.000	0
Persons*raters	104.3313	144	0.725	0.127	13
Persons*criteria	57.947	36	1.610	0.196	20
Raters*criteria	13.562	36	0.377	0.227	24
Error	93.668	432	0.217	0.217	23

The results of Group II show that the variance in scores related to persons (15%) is not the largest component in this analysis. Moreover, variance components involving raters were high (*raters*: 4%; *persons\*raters*: 13%; *raters\*criteria*: 24%).

In Table 3 the estimates of the generalizability coefficients are given for both groups.

Table 3: Estimation of Generalizability Coefficient ( $\rho^2$ ) for the Pooled Peer Assessment Procedure

Number of rating students	G-coefficient for group I	G-coefficient for group II
1	.653	.408
2	.785	.550
3	.841	.622
4	.872	.665
5	.892	.695
6	.906	.716
7	.916	.731
8	.923	.743
9	.929	.753
10	.934	.761
11	.938	.768
12	.942	.774
13	.944	.778

Gronlund (1988) gives some rules of thumb for the acceptability of generalizability coefficients. He states that the generalizability coefficients for classroom assessments in education usually have values between 0.60 and 0.80; 0.60 is considered to be acceptable, but open to improvement, 0.80 is considered as a very reasonable value. The generalizability of the ratings in Group I is better than those of Group II. Within Group I,

there is hardly any need for pooling the scores. The generalizability of scores based on the work of one rating student is already acceptable (0.653). When final scores are based on three rating students the generalizability is good (0.841). In Group I there is, from the perspective of generalizability, no need to base the scores for the peer assessment on more than three students. The generalizability coefficients for Group II show that the quality of scoring is low. A somewhat acceptable level is reached when the final score is based on ratings of at least three students. Although the maximum value looks acceptable, one should bear in mind that this value (0.778) is only reached when the final score is a composite of 13 scores.

*Do raters have idiosyncratic strategies?*

The eigenvalues and percentages of explained variance of the first two principal components are presented in Table 4. The structures of the eigenvalues show, in both groups, a dominant first principal component. The amount of variance between respondents explained by the first factor is 69.3% for Group I and 63.7% for Group II. The amount of variance explained by the second factor is respectively 9.9% and 11.1%. The scree-test strongly suggests a final solution with two principal components. This means that the systematic variance in the peer ratings of the students could be accounted for by two latent variables.

Table 4: Eigenvalues and Percentages of Explained Variance for the First Three Principal Components for Groups I and II

Group	Principal component	Eigenvalue	% of explained variance
I	1	10.51	69.3
	2	1.48	9.9
	3	.86	5.7
II		8.91	63.7
		1.56	11.1
	3	1.13	8.0

Figure 1 presents the component plot for Group I. The factor loadings on the first component for the students ranged from 0.72 to 0.92, while the factor loading for the tutor was 0.70. This means that all respondents had a substantial and positive correlation with the first principal component. This component may be interpreted as a mutual understanding of the quality. Factor loadings on the second component ranged from -0.53 to 0.48, for the students and was 0.56 for the tutor. So, the tutor is somewhat at the extreme pole of the second principal component. The second principal component can be interpreted as the deviation of the mutual understanding of the quality of learning.

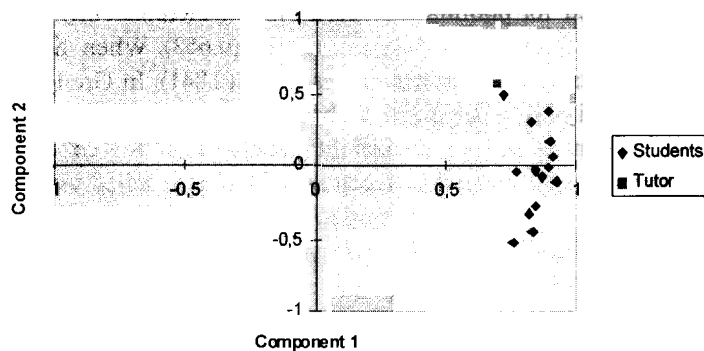


Figure 1: Component Plot Group I

Figure 2 presents the component plot for Group II. For the students the factor loadings on the first component ranged from -0.22 to 0.90, while for the tutor this was 0.83. One rater had a negative factor loading. The other raters had a substantial and positive correlation with the first principal component. Factor loadings on the second component ranged from -0.36 to 0.51. The factor loading for the tutor is 0.13. Inspection of the component plots suggests that one of the students was using an idiosyncratic rating strategy.

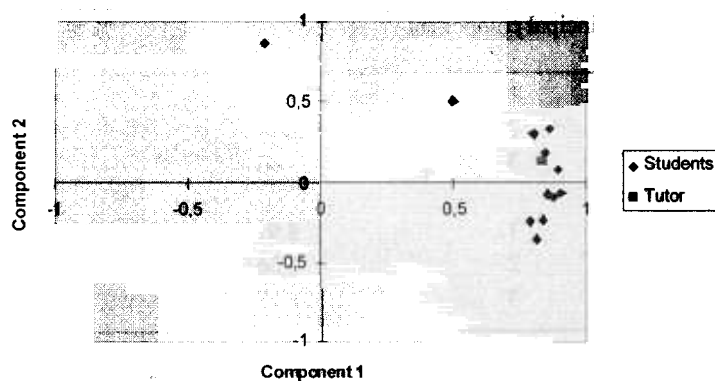


Figure 2: Component Plot Group II

*What are students' experiences with peer assessment and problem based learning?*

Students were asked to give their opinion on various items in the following areas: teamwork, the goals of problem based learning, the instruction and the role of the tutor.

In Table 5 the results on the problem based learning variables are presented.



Table 5: Descriptives of the Items about Problem Based Learning in Study I

Variables	<i>N</i>	Minimum	Maximum	Mean	<i>SD</i>
Working in a group	27	2.80	4.40	3.30	.39
Achieving goals of PBL	27	2.70	4.50	3.42	.42
Instruction process	27	2.82	4.09	3.47	.32
Role of the tutor	27	2.00	4.00	3.26	.51

Students were slightly positive about working in a group. Regarding the achievement of the goals of problem based learning, the majority of the students indicate that problem based learning contributed to the development of their problem solving skills and critical thinking abilities. As regards the instruction process, the situation is not substantially different. In contrast to traditional lectures, problem based learning brings the content of the knowledge domain to the surface and makes the students responsible for their own learning process. Prior knowledge is activated and sometimes the relationship with other knowledge domains is discussed. The authentic character of the problem tasks stimulates active participation in discussion and in working towards problem solution. The search for and selection of relevant information, either independently or in the group, and the integration of different topics were considered as very useful. The perceived role of the tutor varied from quite negative to very positive, with a mean of 3.26, which can be regarded as a neutral attitude.

The students feel that working in a system of problem based learning is very intensive and invokes a high level of responsibility. The learning effects though are very positive because of the active participation in the group process. A high level of co-operation in the group is regarded as conditional for an optimal effect of problem based learning. Dominant roles of certain students occurred, which hindered other students to contribute their input. The students perceived a need for more attention for the development of communication skills. The students sometimes felt that they did not receive enough feedback during the course periods. Especially because they were not used to learning and working in a problem based way.

In Table 6 the results on the peer assessment items are presented for Study I.

Table 6: Percentages of "Yes" Responses on the Items Peer Assessment in Study I

Items Peer Assessment	%
Students are capable of assessing each other	44
Students are capable of assessing each other in a fair and responsible way	19
I feel comfortable when assessing peers	
I knew what peer assessment was about	66
I am in favour of implementing peer assessment	74
Implementing peer assessment means a major change for our institution	82
Peer assessment can be used in other courses	26

In the open-ended questions, students were asked to write down their experiences with peer assessment. The most positive aspects were that the students had the opportunity to express their opinion about the contribution of each peer-student in the group. In this method, more persons make a judgement. The students felt that their scores could be helpful to the tutor. The involvement in the assessment was regarded as fair, although the majority of the students doubted the reliability of the method. Some students were stimulated to think critically about their own learning behavior. Students indicated that peer assessment is not only product-evaluation but also process-evaluation.

A more negative aspect of the peer assessment was that the contribution of the peers differed every period while a score had to be given for the average contribution of four periods. Many students indicated, moreover, that working with a score only was too simple and ineffective. There was no room for feedback. The criteria appeared to be difficult to interpret. One student suggested to weight the criteria, i.e., to indicate which criteria are important and which are of less importance. The peer assessment was not introduced sufficiently well. Some students experienced it as difficult and felt uncomfortable, because they had no prior experience in peer assessment.

### *Conclusions Study I*

In Group I, the largest variance in the scores was related to the performance of *persons* (40%). This means that most of the variance in scores can be attributed to individual effort. The *raters'* variance components suggest differences in leniency. The interaction between *persons* and *raters* (*persons\*raters*) indicates that interpersonal relationships are biasing the peer assessments. The magnitude of the variance components for these student ratings was in concordance with the pattern often found in studies on the generalizability of performance assessment (cf. Moerkerke, 1996).

The quality of ratings clearly differs between the two groups. This means that it is unclear if the peer assessment method applied leads to acceptable results. Probably, the robustness of the method can be improved (as outlined in the discussion below). The peer assessment score is based on the ratings of 12–14 students. Acceptable quality was reached for both groups with this number of raters: The generalizability coefficient was over .90 for Group I and over .75 for Group II.

When we look at the relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component it can be concluded that there is a high level of common strategy among the *raters* in Group I. The relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component also pointed to a high level of common strategy among the *raters* in Group II. The analysis of interrater correlations revealed that one rater can be regarded as odd.

The results of the problem based learning items in the questionnaire show that students stress cooperation, working on an authentic problem, constructing knowledge

---

and skills and active participation as the best features of problem based learning. They generally feel that implementation of this system requires a lot of time and that the ability to work together in a group demands particular skills, which need to be trained and monitored. The most striking result is that the students felt very uncomfortable in assessing their peers. The implementation of peer assessment would be a major change for the institution.

## Study II

### *Method*

#### *Participants*

Besides a study in a university setting, an identical study was organized within the setting of higher vocational education. The population in the second study consisted of 51 fourth year students of a primary teacher training college (13 male, 28 female) also using problem based learning. The students were randomly distributed amongst four groups ( $n1=12$ ;  $n2=13$ ;  $n3=13$ ;  $n4=13$ ).

#### *Materials*

The peer assessment form used in Study I was used for the actual peer assessment. The same evaluation questionnaire was used to measure students' perception on problem based learning and peer assessment.

#### *Procedure*

During the course "Developing Schools", students had to develop a school plan that represented the ideal elementary school. To do this, students had to be aware of current developments in elementary education. In this particular course, students had to be willing to make an active contribution to a group product, to work independently and to be responsible for their own learning.

Each of the four groups worked for a period of six weeks. The groups received a problem task, which concerned the development of a school plan. At the end of the sixth week, all students assessed the peers from their group on the same four criteria as in Study I. In contrast to the first study, the scores were not used as part of the final score. The teacher of the class decided that the peer assessment was an independent activity in the course. The individual peer assessment score was translated into an absolute score ranging from 1 to 10. In this situation a non-passing peer assessment score ( $< 5.5$ ) would lead to an additional study task. In this study the tutor was not one of the raters in the peer assessment.

#### *Data Analysis*

The data-analyses were similar to the analyses in Study I.

### Results

#### *Are peer ratings in problem based learning groups reliable?*

In Tables 7–10, the estimated variance components are shown for the student ratings for Groups I to IV, respectively. To enable the interpretation of the quality of the method, the data on the four groups should be compared. The four groups manifest two patterns. The pattern of Groups I and II corresponds with the pattern of Group II in the first study. This pattern consists of relatively low *person* variance (about 20%) and thus relatively high error variance and interaction variance. *Rater* variance (about 5%) and *criteria* variance (about 1%) are low.

Table 7: Estimation of Variance Components *Persons*, *Raters* and *Criteria* of Group I

Source of variance	Sum of Squares	df	Means square	$\sigma^2$	% total variance
<i>Persons</i> (P)	17.005	11	1.546	.0278	14
<i>Raters</i> (R)	10.130	11	0.921	0.011	6
<i>Criteria</i> (C)	0.880	3	0.293	0.000	0
<i>Persons*raters</i>	17.973	121	0.149	0.006	3
<i>Persons*criteria</i>	6.223	33	0.189	0.005	3
<i>Raters*criteria</i>	12.432	33	0.377	0.021	2
Error	45.214	363	0.125	0.125	64

Table 8: Estimation of Variance Components *Persons*, *Raters* and *Criteria* of Group II

Source of variance	Sum of Squares	df	Means square	$\sigma^2$	% total variance
<i>Persons</i> (P)	19.905	12	1.659	0.028	22
<i>Raters</i> (R)	6.231	12	0.518	0.006	5
<i>Criteria</i> (C)	2.071	3	0.690	0.002	2
<i>Persons*raters</i>	16.018	144	0.111	0.011	9
<i>Persons*criteria</i>	5.467	36	0.152	0.007	6
<i>Raters*criteria</i>	5.006	36	0.139	0.006	5
Error	28.456	432	0.066	0.066	52

Table 9: Estimation of Variance Components *Persons*, *Raters* and *Criteria* of Group III

Source of variance	Sum of Squares	df	Means square	$\sigma^2$	% total variance
Persons (P)	76.485	12	6.374	0.117	44
Raters (R)	17.562	12	1.464	0.022	8
Criteria (C)	2.402	3	0.801	0.003	1
Persons*raters	27.015	144	0.188	0.027	10
Persons*criteria	5.751	36	0.160	0.006	2
Raters*criteria	7.598	36	0.211	0.010	4
Error	34.749	432	0.080	0.080	30

Table 10: Estimation of Variance Components *Persons*, *Raters* and *Criteria* of Group IV

Source of variance	Sum of Squares	df	Means square	$\sigma^2$	% total variance
<i>Persons</i> (P)	45.524	12	3.794	0.069	43
<i>Raters</i> (R)	6.678	12	0.556	0.006	4
<i>Criteria</i> (C)	1.129	3	0.376	0.002	1
<i>Persons*raters</i>	27.899	144	0.194	0.038	24
<i>Persons*criteria</i>	2.121	36	0.059	0.001	1
<i>Raters*criteria</i>	3.121	36	0.087	0.004	3
Error	17.379	432	0.040	0.040	25

This pattern in variance leads to low generalizability coefficients as indicated in Table 11. A somewhat acceptable level of generalizability is reached when at least five ratings are considered. The second pattern can be found in Groups III and IV. This pattern corresponds with the pattern of group I in the first study. It consists of relatively high *person* variance (about 40%), low *rater* variance (about 6%) and low *criteria* variance (about 1%), and leads to acceptable generalizability coefficients. The mean score of one or two ratings lead to acceptable rating practice.

Table 11: Estimation of Generalizability Coefficient ( $\rho^2$ ) for the Pooled Peer Assessment Procedure

Number of rating students	G-coefficient for group I	G-coefficient for group II	G-coefficient for group III	G-coefficient for group IV
1	.419	.488	.708	.585
2	.582	.643	.825	.737
3	.669	.720	.873	.807

Table 11/Cont.

Table 11 (cont.)

Number of rating students	G-coefficient for group I	G-coefficient for group II	G-coefficient for group III	G-coefficient for group IV
4	.723	.766	.899	.847
5	.760	.796	.915	.873
6	.787	.817	.926	.891
7	.807	.833	.935	.904
8	.823	.846	.941	.915
9	.836	.856	.946	.923
10	.846	.864	.950	.930
11	.855	.870	.953	.935
12	.862	.876	.956	.940
13	.869	.881	.958	.944
14	.874	.885	.960	.947

*Do raters have idiosyncratic strategies?*

In order to investigate this question the same approach was used as in Study I. Using the interrater correlations matrix of each of the four groups a principal component analysis was conducted. The explained variance of the first two principal components is presented in Table 12.

Table 12: Eigenvalues and Percentages of Explained Variance for the First Two Principal Components for Groups I, III and IV

Group	Principal component	Eigenvalue	% of explained variance
I	1	5.54	46.2
	2	2.33	19.4
III	1	11.04	84.9
	2	.95	7.3
IV	1	9.98	76.6
	2	1.29	9.9

Due to lack of variance the principal component analysis for Group II cannot be performed. The pattern of the component plot of Group I has the same structure as group II in Study I (see Figure 2). The patterns of the component plots of Groups III and IV are comparable with Group I in Study I (see Figure 1).

*What are students' experiences with problem based learning and peer assessment?*

The students in this study also filled out the evaluation questionnaire. The descriptives of the four variables of the problem based learning environment (working in a group, achieving the goals, the instruction process, the role of the tutor) are presented in Table 13.

Table 13: Descriptives of the Items About Problem Based Learning in Study II

Variables	N	Minimum	Maximum	Mean	SD
Working in a group	51	2.92	4.62	3.99	.51
Achieving goals of PBL	51	2.70	4.50	3.67	.48
Instruction process	51	2.82	4.09	3.87	.42
Role of the tutor	51	1.80	4.20	2.56	1.08

The mean of the first variable indicates that problem based learning stimulates active contribution in a group. The students are also positive about the achievement of certain goals. It appeared that working according to the problem based approach is supportive in developing skills like defining instructional problems, analyzing problems, critical thinking, leading a discussion, interacting with peers, using prior knowledge, giving argumentations and presenting reports. The instruction process was evaluated very positively, resulting in a mean score of 3.87. Finally, the students were less positive about the role of the tutor.

In Table 14 the results on the peer assessment items are presented.

Table 14: Percentages of "Yes" Responses on the Items Peer Assessment in Study II

Items Peer Assessment	%
Students are capable of assessing each other	73
Students are capable of assessing each other in a fair and responsible way	61
I feel comfortable when assessing peers	59
I knew what peer assessment was about	31
I am in favour of implementing peer assessment	71
Implementing peer assessment means a major change for our institution	53
Peer assessment can be used in other courses	50

Results show that the majority feels that they are capable of making assessments and in favor of implementation of peer assessment practices, although they were not very well informed as to what the peer assessment was about.

The most positive element of the peer assessment for the students was that they felt involved in the assessment procedure. They appreciated that their opinion was taken seriously. The majority of the students experienced peer assessment as a method to force

them to think about the contribution of their peers in the group as well as about their own contribution. Some students stressed that a student at a primary teacher training college should be capable of giving critical comments on the work of peers. Peer assessment was regarded as an opportunity to express appreciation for the work done.

The most negative aspects of the peer assessment were that the students felt uncomfortable in making negative judgements. They stressed that the situation is too personal and that it is useless if there is no opportunity to give feedback. For certain students receiving a negative score can have serious consequences. Students indicate that they should not have the power to give negative scores if there is no evaluation or argumentation afterwards.

### *Conclusions Study II*

The question *Are peer assessments in problem based learning groups reliable?* cannot be answered affirmatively. The patterns in the data are scattered. In two of the four groups the peer assessment method does lead to acceptable generalizability coefficients. However, in the other two groups this is not the case. This result can be explained by a low *person* variance. In such a case, the scores do not discriminate between skilled and non-skilled students. An in-depth look at the data of Groups III and IV in Study II revealed that those categories of the rating scales that indicated incompetent behavior, were hardly used. When scales are not fully used, scores become homogeneous and thus non-informative.

Based on the relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component it can be concluded that there is a high level of common strategy among the *raters* in Groups III and IV. The relative magnitude of the eigenvalue for the first component and the magnitude of each of the loadings on the first principal component pointed to a lower level of common strategy among the *raters* in Group I of Study II. The component plot of the data of group I also showed one outlier.

The results of the questionnaire show that most students support the idea of problem based learning, but that they are less positive about the way the tutor functioned. Students are positive about the peer assessment, although they agree that filling out a piece of paper does not change the whole educational system. The method needs improvement; as it is the peer assessment is too subjective according to a lot of students. Many students agree that this is an innovation that needs further development, both for students and teachers.

### *Conclusion and Discussion*

Studies I and II describe the reliability of peer assessment in problem based learning groups. The generalizability of the ratings in Study I appeared to be better for



Group I than those of Group II. Within Group I, the generalizability is acceptable when final scores are based on three rating students (0.84). The generalizability coefficients for Group II led to the conclusion that the maximum value is acceptable, but that this value (0.78) is only reached when the final score is a composite of 13 scores. In Study II, an acceptable level of generalizability is reached in Groups I and II when at least five ratings are considered (0.76 and 0.80). In Groups III and IV an acceptable level of generalizability is reached when at least two ratings are considered (0.83 and 0.74). This pattern corresponds with the pattern of Group I in the first study. Although the groups are randomized, result of the first study leads in one group to acceptable results and in the other group to unacceptable results. The peer assessment method does not lead to enough score variance. This is also the case in the second study, where the results in Groups III and IV lead to acceptable results only.

It is remarkable that there seems to be a cultural difference between the students in Study I (university education for educational scientists) and Study II (higher vocational education for primary school teachers). In Study I extremely negative scores were used more often. The comments of the students in Study II indicate that they were not satisfied with the method. They found it unacceptable to give negative scores without having the opportunity to give informational feedback. In a primary teacher training college students should be supported to give constructive comments and not a mere score.

The results of the questionnaire about problem based learning of Study I and II are quite similar, although two aspects need some discussion. First, we see that the students in Study I are less positive about working in a group than the students in Study II. An explanation for this difference between the students of the two studies might be that students in Study I had no experience with problem based learning and working in teams, while students in Study II were more familiar with this kind of instruction. Second, students in the first study are less negative about the tutor than the students in the second study. This may be explained by the fact that the tutor in the second study was much less involved; students had to do the problem solving process all on their own.

The results of the peer assessment part of the questionnaire revealed that the students in the second study are more confident in their ability to assess than the students in the first study. They feel more comfortable about assessing than the group in the first study. An explanation could be that students in Study II have more experience with different kinds of instruction. The students in Study I had more prior knowledge, because the teacher informed the students better about the peer assessment. All students were positive about implementing peer assessment, but the students in the first study predicted many more implications for the institution. Half of all students can see possibilities for peer assessment in other courses with problem based learning.

On the whole it can be concluded that the peer assessment method applied in these studies needs improvement. One improvement could be that not only processes but also products be evaluated in a peer assessment. In the current study only the process was subject to assessment. Another major improvement would be that students be provided with a possibility to give informational feedback to benefit subsequent learning

---

processes. These types of improvements should lead to a full use of the assessment scale.

Peer assessment the way it was conducted in the studies does not prevent rating errors like *friendship marking*, resulting in over-marking; *collusive marking*, resulting in a lack of differentiation within groups; *decibel marking*, where individuals who dominate groups get the highest marks; and *parasite marking*, where students fail to contribute but benefit from groupmarks (Pond, Ul-Haq, & Wade, 1995). The rating errors, outlined in the Introduction, seem not to be eliminated. Severity and leniency, respectively, lead to under-marking and over-marking of peers. The halo effect occurs when students find one criterion the most important one, thus slanting the objectivity towards the other criteria. Students have different views on the quality of a performance. This was outlined as the significant effect. And finally, evaluation policy means that all students have their own interpretations of the importance/meaning of the established assessment criteria

Giving students opportunities to carry out peer assessments by means of a scoring form seems not enough. Our results underpin the need for instruction in peer assessment, in order for students to make reliable judgements. While Arter (1996) and Perkins (1986) already stressed the need for training students in assessment skills, schools have only recently been paying attention to the development of this type of skills. Students, but also teachers, seldom get training and practice in the development of assessment skills. This was also the case in the school settings we presented in the two studies. Moreover, little is known about training in assessment skills.

The basic goal in current and future research should be to control the strategies students use when they have to make a judgement about their own work or that of their peers. Novice-behavior in rating is characterized by rating-errors or the naive strategies learners exhibit in using peer assessment. When there is no training in assessment skills, rating processes will stay subject to a variety of measurement errors. Hogarth (1981) already stated that the literature shows a depressing picture of human judgmental ability.

Despite the occurrence of rating errors, several studies show that the ability of students to rate themselves improves in the light of feedback or development over time (Birenbaum & Dochy, 1996; Boud & Falchikov, 1989; Griffée, 1995). Moreover, students' interpretations are not just dependent on the form of the assessment process, but on how these tasks are embedded within the total context of the subject and within their total experience of educational life.

An overall conclusion is that adequate training in peer assessment strategies is necessary to eliminate rating errors. This training has to be embedded in the course domain, in order to ingrate assessment and the instruction (Frederiksen, 1994). To begin a process of designing instruction in which peer assessment strategies are included, it is important to reflect on the learning environment in which such training can achieve an optimal effect. A clear definition of the context is required. Criteria have to be elaborated and discussed during a course period. Students have to learn how to give feedback and how to write a peer assessment report. Peer assessment is not only a tool to provide a

peer with constructive feedback which is understood by the peer. Above all, peer assessment is a tool for the learner himself (Dochy & McDowell, 1997).

## References

- Arter, J. (1996). Using assessment as a tool for learning. In R. Blum & J. Arter (Eds.), *Student performance assessment in an era of restructuring* (pp. 1-6.). Alexandria, VA: Association for Supervision and Curriculum Development.
- Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum, & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-29). Boston, MA: Kluwer.
- Birenbaum, M., & Dochy, F. (Eds.) (1996). *Alternatives in assessment of achievement, learning processes and prior knowledge*. Boston, MA: Kluwer.
- Borman, W.C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 60, 556-560.
- Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of self-assessment in higher education: A critical analysis of findings. *Higher Education*, 18, 529-549.
- Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 271-302). Washington, DC: American Council on Education.
- De Groot, A.D. (1975). *Methodology* (9th ed.). 's-Gravenhage: Mouton.
- Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation*, 23 (4), 279-298.
- Dochy, F., & Moerkerke, G. (1997). The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research*, 27 (5), 415 - 432.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer-, and co-assessment in higher education: A review. *Studies in Higher Education*, 24 (3), 331-350.
- Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Frederiksen, N. (1994). The integration of testing with teaching: Applications of cognitive psychology in instruction. *American Journal of Education*, 102 (4), 527-564.
- Griffiee, D.T. (1995). Criterion-referenced test construction and evaluation. In J.D. Browne & S.O. Yamashita (Eds.), *Language testing in Japan* (pp. 20-28). Tokyo, Japan: The Japan Association for Japan Language Testing.

Gronlund, N.E. (1988). *How to construct achievement tests*. (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lockwood, F. (1995). Students' perception of, and response to, formative and summative assessment material. In F. Lockwood (Ed.), *Open and distance learning today* (pp. 197-207). London: Routledge.

Lunz, M.E., Wright, B., & Linacre, M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3 (4), 331-345.

McKeown, B., & Thomas, D. (1988). *Systematic data collection*. Newbury Park: Sage.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovations in Education and Training International*, 32, 314–323.

Sadler, D.R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education*, 54 (1), 60-79.

Sambell, K., & McDowell, L. (1998). The value of self and peer assessment to the developing lifelong learner. In C. Rust (Ed.), *Improving student learning — improving students as learners* (pp. 56–66). Oxford, UK: Oxford Centre for Staff and Learning Development.

Segers, M.S.R. (1997). An alternative for assessing problem-solving skills: The OverAll test. *Studies in Educational Evaluation*, 23 (4), 373-398.

Segers, M., & Dochy, F. (1999). Een nieuw onderwijsmodel voor het Hoger Onderwijs in theorie en praktijk [A new educational model for higher education in theory and practice]. In M. Lacante, & P. De Boeck, *Handboek leerlingenbegeleiding [Handbook student counselling]* (pp. 153-180). Dordrecht: Kluwer.

SPSS (1997). *SPSS for Windows*. Chicago: SPSS Inc.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self- peer- and co-assessment. *Learning Environments Research*, 1, 293-319.

Tucker, L.R. (1962). Factor analysis of relevance judgements: An approach to content validity. *Proceedings: 1961 Invitational Conference on Testing Problems* (pp. 29-38). Princeton, NJ: Educational Testing Service.

Van der Vleuten, C.P.M., Verwijnen, G.M., & Wijnen, W.H.F.W. (1996). Fifteen years of experience with progress testing in a problem based learning curriculum. *Medical Teacher*, 18 (2), 103-109.

Voss, J.F., & Post, T.A. (1990). On the solving of ill-structured problems. In N. Frederiksen, R. Glaser, A. Lesgold & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 261-285). Hillsdale, NJ: Erlbaum.

---

### The Authors

DOMINIQUE SLUIJSMANS has been a PhD student since 1998 at the Educational Technology Expertise Centre of the Open University of the Netherlands. Her main interests are student involvement in assessment and curriculum design.  
E-mail: dominique.sluijsmans@ou.nl

GEORGE MOERKERKE is an educational technologist and researcher at the Educational Technology Expertise Centre of the Open University of the Netherlands. His main topics are curriculum development and assessment.  
E-mail: george.moerkerke@ou.nl

FILIP DOCHY is Professor of Educational Innovation & Information Technology at the University of Maastricht, Department of Educational Innovation & IT and at the Research Centre for Teacher Education, Department of Instructional Science University of Leuven, Belgium.  
E-mail: Filip.Dochy@Ped.kuleuven.ac.be

JEROEN VAN MERRIËNBOER has a Master's degree in cognitive psychology and a PhD degree in instructional technology. He is now a Full Professor at the Educational Technology Expertise Center of the Open University of the Netherlands, where he is also heading the research program. He specializes in Instructional Design (ID) for complex learning, computer-based learning environments, and intelligent performance support for ID.  
E-mail: jeroen.vanmerrienboer@ou.nl

---