How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in Flanders.

Tijs Rotsaert[1], Ernesto Panadero[2], Eduardo Estrada[3] & Tammy Schellens[1]

Author note

[1] Department of Educational Studies, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

[2] Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Madrid, Spain

[3] Departamento de Psicología, Universidad Camilo José Cela, Madrid, Spain.

Recommended citation:

Rotsaert, T., Panadero, E., Estrada, E., & Schellens, T. (2017). How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in Flanders. Studies In Educational Evaluation, 53, 29-40. doi:http://dx.doi.org/10.1016/j.stueduc.2017.02.003

Correspondence concerning this manuscript should be addressed to: Tijs Rotsaert. Department of Educational Studies, Ghent University, Henri Dunantlaan 2, BE9000 Ghent, Belgium. E-mail: Tijs.Rotsaert@UGent.be

Funding:
Second author funded by the Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad) under the Ramón y Cajal program (Reference number: RYC-2013-13469).

Abstract

This study explores the relationship between students' perceptions of peer assessment (PA) and its social nature. A quantitative survey study (N = 3680) was conducted in secondary education in Flanders, examining the students' perceptions of PA interpersonal variables and their beliefs on the educational value of PA. The structural equation modeling (SEM) results show that the educational value students attribute to PA was positively predicted through trust in their own and their peers' evaluative capabilities, awareness of negative interpersonal processes (e.g. fear of disapproval and friendship marking), and beliefs about PA accuracy. The importance attributed to anonymity appeared to be a negative predictor of PA conceptions. Tests of mean latent differences were performed to explore the differences between educational levels, PA experience and gender.

Keywords: peer assessment; interpersonal processes; anonymity; Structural Equation Modelling

How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in Flanders.

Current perspectives on assessment lead its goals away from end-of-course (i.e. summative) testing to in-course (i.e. formative) improvement-oriented interactions between learners and instructors (Black & Wiliam, 1998; Havnes, Smith, Dysthe, & Ludvigsen, 2012). This formative view blurs the strict distinction between instruction and assessment (Cowie, Moreland, & Otrel-Cass, 2013) and is often referred to as the 'Assessment for Learning' position (Assessment Reform Group, 2002). AfL/Formative assessment strategies stress that active involvement of students in assessment processes is necessary. For this reason peer assessment (PA) has been embraced as an innovative method of formative assessment and attributed significant educational value for learning (e.g., Topping, 2010).

In a PA activity, a student generates feedback that might be useful to the assessee, and potentially gives the peer assessor insights as to how their own work might be improved (Dochy, Segers, & Sluijsmans, 1999; Panadero, 2016; Reinholz, 2015). In this study PA is defined as an interpersonal collaborative learning arrangement in which students assess their fellow peers' performance by providing feedback (peer feedback – PF), which can be quantitative (i.e., grades or ratings across assessment criteria) and/or qualitative (written or oral comments) (Bolzer, Strijbos, & Fischer, 2015; Topping, 2010). Despite the benefits of PA, it remains a challenging assessment method to implement. Its social nature as being a fundamentally interpersonal process has been relatively overlooked and has only been studied in relatively small-scale interventions studies, mainly within vocational and higher education (Panadero, 2016; van Gennip, Segers, & Tillema, 2010).

Furthermore, previous research shows that teachers and students' conceptions about the purpose of assessment largely influence its implementation (Brown, Lake, & Matters,

2011; Segers & Tillema, 2011). To this end, this study aims to explore the relationship between secondary student perceptions related to PA and its social nature.

Conceptions of peer assessment

Classroom assessment is always a social experience: it does not happen in isolation because students define and practice assessment from their own point of view and in relationship with others (e.g. teachers, peers) (Brown, McInerney, & Liem, 2009). These student conceptions represent ideas, beliefs, propositions and preferences that fundamentally describe how students experience educational matters such as assessment practices (Brown et al., 2009). In formative assessment, students' conceptions are essential because it is students who must use the assessment to learn (Cowie, 2009). Unfortunately, the students' conceptions are not yet explored in sufficient detail even in formative assessment literature, a reason why they will be explored in the present study.

A small number of studies have been conducted on students' conceptions of assessment indicating their significant contribution to students' learning behavior and (future) learning (e.g., Harris, Brown, & Harnett, 2014; Struyven, Dochy, & Janssens, 2005). In a series of studies Brown and colleagues have reported on how to measure student and teachers' conceptions of the purpose of assessment (e.g., Brown & Hirschfeld, 2008). Building on Ajzen's (2005) theory of planned behavior, which suggests that personal intentions or beliefs about what others think shape their behavior, Brown and colleagues argue that students' responses to assessment depend on their appreciation of the process as well as its aims. Additionally, drawing on Zimmermans' (2001) self-regulation theory, they state it is important for students to have a 'personally meaningful purpose of assessment', as self-regulated learners often need to use feedback from educational assessment (Brown & Harris, 2011, p 46).

Therefore, more insight on students' perspectives about PA is essential. A recent large scale survey study generated a robust picture of teacher thinking about the use of PA (Panadero & Brown, 2016). This study came to the conclusion that teachers overall like the instructional use of PA, struggle with inherent difficulties (e.g. peer pressure) and their self-reported PA use depends largely on previous positive experiences. Interestingly, primary and secondary teachers reported higher values of PA implementation and certainty about its educational value, in contrast with the higher education teachers. Similar results were found in two previous studies also reporting teachers' conceptions about PA with smaller sample sizes and with only higher education teachers or secondary teachers (Lynch & Golen, 1992; Noonan & Duncan, 2005). Nevertheless, the field lacks knowledge about the conceptions of students as the most important actors in this fundamentally interpersonal assessment process, which will be explored in the present study.

Peer assessment and its social nature: Six interpersonal variables of interest

Most research on assessment has always been aware of its inherent social and emotional nature, and this is especially the case for PA (Boud, 1995). Previous research has questioned the accuracy and/or validity and/or reliability of PA (for a detailed discussion see Panadero, Romero, & Strijbos, 2013) because of the presence of possible reciprocity effects caused by interpersonal processes such as friendship marking or psychological unsafety (Harris & Brown, 2013; van Gennip et al., 2010; Vanderhoven et al., 2015). The limited number of studies on the social nature of PA show that students' perceptions of these interpersonal processes might be related to their conceptions towards PA (Cheng & Tsai, 2012; Harris & Brown, 2013; van Gennip et al., 2010). Attention to social and human factors is thus needed because well implemented PA should decrease negative social problems, assure accuracy and lead to positive learning outcomes (Panadero & Brown, 2016; Topping, 2010).

Six interpersonal variables that are frequently referred to in research, are discerned in this study (for a detailed discussion see Panadero, 2016). These six variables were chosen because are the most relevant when it comes to their possible effects on PA. (1) *Friendship marking* due to friendship bonds has been mentioned as a source of potential scoring bias. However, only a small number of studies have directly addressed this topic (Panadero et al., 2013). Recent research on the diminishing effect of rubrics on over- and underscoring by peers in PA shows that for low and medium friendship levels, a rubric does reduce the friendship bias, but for high-level friendship, the rubric even seems to amplify the potential friendship bias (Panadero et al., 2013). Cheng and Tsai (2012) found that anonymity was preferred for the reason of avoiding the pressure of friendships. (2) *Fear of disapproval* refers to the assessors' fear of negative comments from the assessee if they give them a low score or negative feedback (recrimination) (Cartney, 2010). To decrease this type of fear, it has been argued that anonymity might play a role. For example, in Vanderhoven et al. (2015) students in an anonymous face-to-face PA setting experienced significantly less fear of disapproval compared to students in a non-anonymous setting. (3) *Psychological safety*, refers to a situation in which students have a shared belief about taking interpersonal risks in a group. People that feel psychologically safe tend to perceive differences in opinions as opportunities rather than conflicts (Nicol, 2010; Yu & Sung, 2015). This is important as several authors state that creating a safe environment is a precondition for accurate and thus valuable PA activities (Harris & Brown, 2013; van Gennip, Segers, & Tillema, 2009). (4) *Value congruency* refers to the importance of unanimity on both the goals and criteria of the PA activity (Cheng & Tsai, 2012). Rubrics hold the potential of augmenting the value congruency within a PA-activity as they provide the assessment criteria in a structured format and might thus enhance the perceived fairness and comfort with PA (Panadero et al., 2013). (5) *Trust in the self as assessor* refers to the assessors' beliefs about their skills when assessing a peer

(van Gennip et al. 2010). Previous research has indicated that the higher the trust in self as assessor, the deeper the learning approaches towards PA (Cheng & Tsai, 2012) and might be increased through intensive practice and interaction (Panadero et al., 2016). (6) *Trust in the other as assessor* refers to the confidence in the reliability and validity of the assessment and feedback received from a peer. Students will only act on the basis of trustworthy information: if they believe that comments are capricious, they will not act on the basis of them (Carless, 2013).

Two crucial aspects in peer assessment: anonymity and accuracy

As this interplay of interpersonal variables influences the assessment outcome, it has often been stated that decreasing negative social effects via anonymity is desirable (Ballantyne, Hughes, & Mylonas, 2002; Vickerman, 2009) or should at least be explored (Howard, Barrett, & Frick, 2010). Topping (1998) indicated that privacy is an important structural feature of PA, in that disclosing the identity of the assessor or assessee, seems to matter to students. Vanderhoven et al (2015) found that students have more positive attitudes towards PA when anonymity for the assessor was assured, while the participating teacher suggested that identity revelation towards him might be desirable as a means to control for undesirable social effects. Yu & Sung (2015) state that anonymity might offer more psychological safety for students, but at the same it might lead to misbehavior, for example more positive marking towards friends. A recent survey study on Spanish teachers' reasons for PA use by Panadero and Brown (2016) revealed that they predominantly believed in the use of anonymous versions of PA, although it was not found to be a significant determinant to PA frequency of use, except for university teachers. In conclusion, anonymity needs further research especially students' conceptions towards the different anonymity modes that can be manipulated in a PA setting (i.e. anonymity towards the assessor, the assessee and the teacher).

Another crucial aspect related to PA is the concerns about the (perceived) validity and/or accuracy and/or reliability of student PA. The problem for some is that students, as novices and learners, may not be sufficiently competent in a field to make an accurate estimation of another's' work quality. Empirical research shows that students can be reliable sources under appropriate conditions (Falchikov & Goldfinch, 2000; Topping, 2003) such as being accompanied by the use of rubrics, involving students' in the discussion about the criteria and/or considering the level of expertise of the students (for a detailed discussion see Panadero et al. 2013). In the aforementioned Spanish survey study on teachers' reasons for PA use, the results clearly demonstrated that teachers were concerned about the accuracy of PA, although it was not a significant predictor of the self-reported use of PA (Panadero & Brown, 2016). In the present study the broad term accuracy will be used as it refers to the overall quality combining the reliability and validity indicators.

In sum, interpersonal variables, anonymity and accuracy play an important role in how students perceive the educational value of PA. However, our knowledge about these fundamental issues is often based on small sample studies. Therefore, there is a need for a larger study that explores these issues.

The Flemish context and its assessment practices

To understand the results of the study in this paper, it is important to get acquainted with the assessment context in Flanders (the Dutch community in Belgium). In general, there are four type of schools: General Secondary Education, Technical Secondary Education, Arts Secondary education and Vocational Secondary education. The government only imposes a minimum time table, defined as core-curriculum subjects, which depend on the educational level. Schools are distributed across educational networks: GO! (18.4%), official subsidized education (7.4%), subsided private education (75.5%) who function independently of the Flemish Ministry of Education (Flemish Governement, 2015). From an international point of

view secondary education in Flanders is known for its high quality (e.g., within top 10 for mathematical reasoning in PISA 2012) (Department of Educational Studies- Ghent University, 2013).

At a Belgian level, national tests do exist but these are exclusively concerned with the compulsory attainment and developmental targets of the curriculum and have no public accountability element. The main goal of this type of national testing is to monitor and evaluate schools and/or the education system as a whole. National test results are thus used as indicators of the quality of teaching and the performance of teachers, but also to gauge the overall effectiveness of education policies and practices. It is expected that this specific context of national testing could have an influence on students' conceptions of peer assessment.

On the meso and micro level, for decades the Flemish government has emphasized the importance of autonomy and trust in the policy-making capacity of schools. As a part of this autonomy, teachers and teacher councils are, as a rule, solely responsible for the majority of pupil learning and classroom assessment. All assessment practices are thus voluntary and low stakes. As a consequence, implementing formative assessment is the responsibility of individual teachers.

Regardless of the high amount of policy autonomy that has been attributed towards school boards, the Flemish government has taken some initiatives to promote formative assessment in the curriculum. For example, through the development of a toolkit *Breed Evalueren* ('broad' assessment) to support teacher to assess Dutch competences in both primary and secondary education. PA is explicitly mentioned herein (De Backer & Philips, 2013). Furthermore, the two biggest educational networks have written out a clear vision on assessment which mirror a balanced vision on assessment: both AfL as assessment *of* learning practices are promoted (GO!, 2012; Katholiek Onderwijs Vlaanderen, 2013).

As these recent initiatives promote AfL practices and give concrete suggestions on how to implement them, it could be assumed that AfL practices have found their way into classroom practice. Therefore, it could be expected that Flemish secondary education students would understand the educational value of assessment practices and already have had some PA experience. However, up-to-date no empirical evidence could be found confirming the effects of the aforementioned initiatives.

Aim and research questions

The present study explores Flemish secondary education students' self-reported experiences with PA. The goals of the study were to determine whether the concerns raised in the literature exist and affect students' attitudes towards PA. The research questions are:

RQ1 - Do Flemish secondary education students report having experienced PA, with what frequency, and in what format?

RQ2 – What variables predict students' PA conceptions?

Based on the previously described theoretical framework, it is hypothesized that (1) students' perceptions towards negative interpersonal variables (e.g., friendship marking,) would negatively affect their belief in the educational value of PA. (2) Positive interpersonal variables (e.g., psychological safety, trust and value congruency) would positively affect students' PA conceptions. (3) Perceived validity/accuracy would increase their belief in the educational value of PA, and (4) high perceived importance towards use of anonymous forms of PA would decrease positive belief in the educational value of PA.

RQ3 – What differences exist between gender, educational levels and students with differing PA experience?

Method

Participants

A total of 3680 Flemish High School students participated in this study. A subsample of 3066 students with PA experience was used for this study. In this subsample 32.2% of the participants were in grades 7-8 (henceforth *Level 1*); 34.4% grades 9-10 (henceforth *Level* 2) and 33.5% grades 11-12 (henceforth *Level 3*). In terms of demographic information, the students had an average age of 15.10 (*SD= 1.94*), range 11 - 21. The percentages of males and females were 45.3% (N=1398) and 54.6% (N=1675). The distribution of the collected data over the four educational types was 53.1% in General Secondary education, 29.2% in Technical Secondary education, 11.7% in Vocational Secondary Education and 5.6% in Arts Secondary education. Additionally in Table 1 the distribution by Flemish regions can be seen. The majority of the data was collected in East Flanders because the university that organized this survey is situated there and it was easier to arrange the data collection in such region.

Table 1
Distribution of the sample by regions and type of school

Region	*n*	*%*
Antwerp	386	10.50
East Flanders	1702	46.30
West Flanders	1278	34.70
Flemish Brabant & BCR	216	5.90
Limburg	98	2.70
Type of school		
General Secondary	1135	48.63
Technical Secondary	676	28.96
Arts Secondary	182	7.80
Vocational Secondary	341	14.61

Procedure

Survey-conductors were bachelor students in educational studies enrolled in a methodology course. In groups of four students they were asked to get permission from eight teachers to conduct a survey in their classes. In order to reduce social desirability, the pen-and-paper survey was designed to be filled in anonymously. Participants were asked to also fill in an informed consent which contained information about the purpose, confidentiality assurances and the possibility to withdraw. The informed consents were collected separately so to ensure confidentiality. The survey conductors received detailed instructions and documents how the survey session should take place (e.g. how to guarantee confidentiality through predefined ID-codes) as well as a detailed coding manual (e.g. how to handle missing values) including a predefined Excel sheet to enter the data of the written surveys. As a control mechanism, the first author went through the data file together with the survey conductors when they handed in their paper versions. A double check was performed when merging the data into one file.

Instrument

A self-report survey instrument was designed with 4 blocks: (a) demographic information, (b) descriptive questions about experienced PA-activity (yes/no/sometimes), (c) specific questions about conceptions towards PA, its social nature, and anonymity and accuracy within PA (same as b-block). Two different versions of the instrument were created: one with 49 questions concerning PA for students' with PA experience and one with 36 questions concerning PA for students without PA experience. Next we explain blocks b and d content in more detail.

The items in this questionnaire use a 6 point positively packed rating scale in order to elicit more variance in responses (Strongly disagree, mostly disagree, slightly agree,

moderately agree, mostly agree, and extremely agree) (Brown, 2004). Regarding the block c, the following definition of PA was presented to ensure students have a shared understanding: "In a peer assessment-activity students judge each other's tasks/presentations/group assignments. The judgement can be expressed in scores, oral or written feedback or a combination of both". The PA-section for students with PA experience focused on the following topics (details in Appendix 1):

a) PA activity description. The first eight questions established an overview of the kind of PA activity students had experienced.
b) Interpersonal processes in PA. The next six questions explored students' beliefs towards interpersonal processes within PA.
c) Importance towards anonymity in PA. Three items explored the degree to which students considered anonymity to be important in PA. It was asked how they valued being anonymous (towards fellow peers as well as towards the teacher).
d) PA accuracy. Two questions were used to establish whether students think PA is an accurate assessment method.
e) PA conceptions. Three questions explored students' attitudes towards PA in terms of usefulness, involvement in the assessment process and perceived learning gain.

A pilot of this survey was conducted before administration. An expert in formative assessment filled out the questionnaire and that input was used to revise some of the items. The revised survey was then evaluated by a teacher in year 9 and 10 (e.g., comprehension problems, length, etc.). After his input the survey was then evaluated with a think-aloud procedure by a pupil in grade 10 and one in grade 13. Finally, the survey was piloted with 16 10^{th} grade students based on all the input of the preceding try-outs. On average they completed the questionnaire within 20 minutes.

Analysis

To answer RQ1 concerning students' experience and attitude towards PA, beliefs towards, we computed the descriptive statistics by level.

The goal of RQ2 was to understand how students' perceived educational value of PA is influenced by self-reported beliefs on interpersonal processes, anonymity and accuracy within PA. Therefore, we identified the relationship of several latent factors and manifest item variables to each other and their contribution to the perceived educational value of PA. For this, Structural Equation Modelling (SEM) was used. SEM allows studying the fit of a formally defined theoretical model to an empirical dataset. It generates multiple parameters for examining particular hypothesis and incorporates latent factors which are helpful for isolating error of measurement out of the path analysis.

The following fit indices were calculated for every model: first, *RMSEA*, with values between 0 and .06 indicating very good fit, values between .06 and .08 indicating reasonable fit. Second, the *SRMR* with values between 0 and .08 indicating very good fit. Third the *CFI* and *TLI* indices. Acceptable values must be larger than .90. Excellent values must be above .95. Finally, the χ^2/df (Chi Square/Degrees of Freedom) ratio is considered. Values showing a good fit must not exceed 2.0 (Schreiber, Nora, Stage, Barlow, & King, 2006; Schweizer, 2010).

Tests of measurement invariance (configural, metric and scalar) were performed to answer RQ3 about differences between educational levels, PA experience and gender. To determine measurement invariance across subgroups in large samples, it is preferable to report the change in *CFI* and *RMSEA* between the unrestricted and restricted models instead of the difference in Chi-square statistics (Chen, 2007); Cheung & Rensvold 2002; Meade, Johnson & Brady, 2008; Kline, 2015). Cheung and Rensvold (2002) recommend using a ΔCFI value higher than .01 to indicate a significant drop in fit. Additionally, Chen (2007) suggests using

Δ*RMSEA* to test for evidence of invariance. The criteria for invariance are Δ*CFI* ≤ .01, Δ*RMSEA* ≤ .015. The tests of latent mean differences were conducted for the groups in which scalar invariance was observed. Assessment of latent mean differences is based on the critical ratio (*CR*) index, where *CR* ≤ or ≥ 1.96 indicates significant differences in the means. The Cohen's *d* effect size index was also calculated to interpret the magnitude of the mean differences (.20=small differences, .50=medium differences, .80=large differences; Cohen 1988). MPlus 7 (Muthén and Muthén, 2007) was used for the SEM, the tests of measurement invariance and the tests of latent mean differences. The models' parameters were estimated through robust maximum likelihood (MLR).

Results

RQ1 - Do Flemish secondary education students report having experienced PA and in what format?

Overall, the majority of students in all levels have experienced a PA activity at least once (Table 2). Furthermore, it follows logically that the majority of the students with no experience are situated in level one. Surprisingly the frequency of experience with PA was generally high, with "more than thrice" being the most chosen option in the first (27,76 %), second (35,15 %) and third (59,91 %) level.

Table 2
Students' experience with PA

Response category	Level 1 (*N*=1239)		Level 2 (*N*=1229)		Level 3 (*N*=1105)	
	N	%	*N*	%	*N*	%
None	*263*	21.23	*169*	13.75	*75*	6.79
Once	*209*	16.87	*217*	17.66	*134*	12.13
Twice	*227*	18.32	*223*	18.14	*135*	12.22

Trice	*196*	15.82	*188*	15.30	*99*	8.96
> Trice	*344*	27.76	*432*	35.15	*662*	59.91

Looking at the subsample of students with PA experience (*N* = 3066) the majority of students in all levels mentioned that the PA activities they experienced are mostly not or sometimes taken into account for the monthly report in terms of grades (see Table 3). Together with the fact that the PA activities took place during a lesson, it is possible that the current use of PA in Flemish secondary education is formative in nature, although this is a highly speculative explanation.

Regarding student training to perform PA current findings are in high contrast with current research guidelines: on all educational levels (74,61% Level 1, 78,76% Level 2 and 80,66% Level 3) students reported not having received PA training. On the other hand, two-thirds of the student population with PA experience reported that they were involved in defining the PA criteria. Surprisingly, the percentage of active involvement decreases from Level 1 (50,64%) to Level 3 (38,83%).

Table 3
Format of PA-usage for students with PA-experience

Question and Response category	Level 1 (*N*=957)		Level 2 (*N*=1015)		Level 3 (*N*=1006)	
	N	%	*N*	%	*N*	%
Grades: Was result of the PA activity mentioned on the monthly report?						
Yes	*140*	14.63	*210*	20.69	*273*	27.14
No	*444*	46.39	*402*	39.61	*286*	28.43
Sometimes	*373*	38.98	*403*	39.70	*447*	44.43
Time: The PA activity took place						
…during a lesson	*574*	59.54	*552*	54.33	*470*	48.25
…at the end of series of lessons	*390* *(N= 964)*	40.46	*464*	45.67	*504*	51.75
Training: Were you trained to perform PA?						

Yes	*128*	13.43	*86*	8.30	*75*	7.48
No	*711*	74.61	*816*	78.76	*809*	80.66
Sometimes	*114*	11.96	*134*	12.93	*119*	11.86
Involvement in defining PA criteria: Were you involved in defining the PA criteria?						
Yes	*593*	50.64	*453*	43.77	*393*	38.83
No	*345*	29.46	*231*	22.32	*302*	29.84
Sometimes	*233*	19.90	*351*	33.91	*317*	31.32

Table 4 provides details of students' responses of guaranteed anonymity within the experienced PA activities. Generally, anonymity was not provided for the assessor, the assessees nor towards the teacher. More specifically, if anonymity is used, anonymity for the assessor is the mostly used anonymous mode. This anonymous mode was reported 10 to 15% at all educational levels. The fact that about one-third indicated 'sometimes' as chosen option in all anonymous modes, possible indicates that teachers are exploring with different anonymous modes depending on the nature of the PA activity.

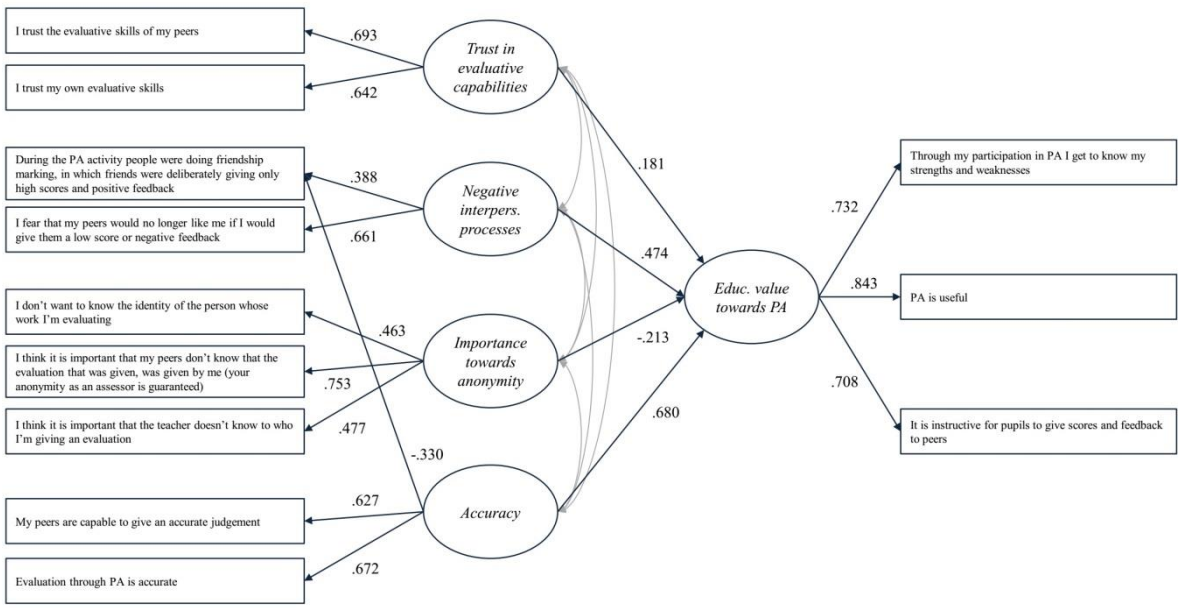
Table 4
Anonymity within PA-activity

Response category	Level 1 (*N*=976)		Level 2 (*N*=1040)		Level 3 (*N*=1010)	
	N	%	*N*	%	*N*	%
Anonymity for the assessor						
Yes	*111*	*11.48*	*170*	*14.57*	*148*	*14.65*
No	*576*	*59.57*	*573*	*49.10*	*548*	*54.26*
Sometimes	*289*	*29.89*	*297*	*25.45*	*314*	*31.09*
Anonymity for the assessee						
Yes	*57*	*5.83*	*44*	*4.25*	*33*	*3.26*
No	*801*	*81.99*	*887*	*85.70*	*879*	*86.94*
Sometimes	*119*	*12.18*	*104*	*10.05*	*99*	*9.79*
Anonymity towards the teacher						
Yes	*80*	*8.34*	*56*	*5.38*	*45*	*4.45*
No	*658*	*68.61*	*821*	*78.87*	*839*	*82.99*
Sometimes	*221*	*23.04*	*164*	*15.75*	*127*	*12.56*

RQ2 – What variables predict students' educational value towards PA?

For studying the latent constructs and their relationship, we fitted a structural equation model. In this model, we proposed five latent constructs or factors for explaining the responses to the survey's items. The first factor is '*Educational value towards PA*' and refers to the extent to which student' thinks PA is a valuable assessment method. The second factor is '*Trust in evaluative capabilities',* which measures the extent to which students believe in their own and peers' evaluative capabilities. The third factor is '*Negative interpersonal processes'* and refers to the extent to which students esteem that friendship marking is going on and to the amount of fear of disapproval they have for giving low scores and/or negative feedback. The fourth factor is '*Importance towards anonymity'* and evaluates the amount of attributed importance towards guaranteeing anonymity to assessors, assessees and/or teachers within PA activities. The fifth factor is '*Accuracy',* and measures students' perceived accuracy of PA and whether they think their peers are capable of giving accurate judgments. In order to correctly identify the model, the variance of all five latent variables was fixed to one. Since we wanted to know what variables predict '*Educational value towards PA*', we forced this latent construct to be explained (i.e. receive loadings) from the other four. Therefore, the model was conceptually equivalent to a multiple regression with one latent dependent variable and four latent predictors. Five items showed low factor loadings, which indicated a low communality with the rest of the scale. We decided to remove these items in order to gain conceptual clarity. The full SEM model (with standardized loadings) is shown in Figure 1.

Figure 1
*Baseline SEM model with standardized loadings**



* *Correlations between latent predictors reported in Table 5 for clarity*

The correlations between the four latent predictors are shown in Table 5. The total amount of variance of '*Educational value towards PA*' explained by the predictors was R^2 = .548 (*std. error* = .046). Note that the latent predictors '*Trust in evaluative capabilities'* (β = .181, *std. error* = .081, *p* = .022), *'Negative interpersonal processes'* (β = .474, *std. error* = .126, *p* < .001) *and 'Accuracy'* (β = .680, *std. error* = .038, *p* < .001) had positive loadings on '*Educational value towards PA*', meaning that higher values in the former tend to be associated with higher values in the latter. On the other hand, '*Importance towards*

anonymity' (β = -.213, *std. error* = .077, *p* = .005) appears to be a negative predictor of students' '*Educational value of PA'.* All four regression loadings were statistically significant.

This model achieved an acceptable to good fit: χ^2 / df = 120 / 43 = 2.79 (note that chi square values are highly inflated by large sample sizes such as ours), *CFI* = .987, *TLI* = .980, *SRMR* = .018 and *RMSEA*= .024, with a 90% confidence interval between .019 and .029. All the items loaded significantly onto their corresponding latent factors ($p < .001$). The item loadings ranged between -.33 and .84.

Table 5
Correlation between latent predictors

N = 3044 *p* < 0.001 in all cases	Neg. interp. processes	Importance towards anonymity	Accuracy
Trust in evaluative capabilities	-.700	-.498	.490
Negative interpers. processes		.692	-.318
Importance towards anonymity			-.298

RQ3 – What differences exist between educational levels, PA experience and gender?

In order to enable comparison of the mean scores of the latent constructs across educational level, differing PA experience and gender, first we tested the scalar invariance between the different groups. If scalar invariance holds, differences in means of the observed items can be interpreted as a consequence of the differences in the means of the latent constructs.

Table 6

Results of the tests of measurement invariance by (1) Level, (2) PA Experience, and (3) Gender

Multi Group CFA		Model Fit Indices				Model Comparisons		
		X^2 (df)	CFI	TLI	RMSEA	Comparison	ΔCFI	ΔRMSEA
(1) Level	Model 1 (configural invariance)	224.0 (129)*	.984	.976	.027			
	Model 2 (metric invariance)	267.3 (155)*	.980	.974	.028	Model 2 vs. Model 1	-.004	.001
	Model 3 (scalar invariance)	327.8 (169)*	.974	.969	.030	Model 3 vs. Model 2	-.006	.002
(2) PA Experience	Model 1 (configural invariance)	246.2 (173)*	.987	.981	.024			
	Model 2 (metric invariance)	302.5 (211)*	.984	.980	.024	Model 2 vs. Model 1	-.003	.000
	Model 3 (scalar invariance)	322.7 (232)*	.984	.982	.023	Model 3 vs. Model 2	.000	-.001
(3) Gender	Model 1 (configural invariance)	179.4 (86)*	.984	.976	.027			
	Model 2 (metric invariance)	194.6 (99)*	.984	.978	.025	Model 2 vs. Model 1	.000	-.002
	Model 3 (scalar invariance)	247.4 (106)*	.976	.970	.030	Model 3 vs. Model 2	-.008	-.005

* <.05 / With (1) 987 students in Level 1, 1053 students in Level 2 and 1026 students in Level 3; (2) 561 students with one PA experience, 586 with two, 481 with three and 1438 with more than three; (3) 1675 Girls and 1389 Boys – Criteria for invariance model comparison: ΔCFI ≤.01, ΔRMSEA ≤.015

As Table 6 shows, scalar invariance holds for all three comparisons, i.e. comparing for educational level, PA experience and gender. This means that students across the different groups interpreted our measurement instrument in a consistent way and that the mean scores of the latent constructs can be compared.

After scalar invariance was verified, we studied the differences between latent means. For doing so, the mean of one of the groups (the reference group) is fixed to zero in all latent variables, so the means of the other groups can be freely estimated and interpreted as a difference between those groups and the reference group. In the present study, students from level 1 (1), with only one PA experience (2), and males (3) were the reference groups in each comparison. Note that, since the metric of all latent variables is arbitrary, the comparisons of latent mean differences do not enable estimation of the absolute mean in each group but rather study differences in the latent variables between groups. Table 7 shows the results of the test of the latent mean differences for each latent variable. Since we fixed all the variances to one, the estimated mean differences can be interpreted as Cohen's standardized differences (Cohen 1988).

Table 7
Test of latent mean differences

	Latent construct	*d*	*S.E for difference*	*CR* (*z* = Estimate / S.E.)	*p*	*d*	*S.E for difference*	*CR* (*z* = Estimate / S.E.)	*p*	*d*	*S.E for difference*	*CR* (*z* = Estimate / S.E.)	*p*
		Second vs. First level				Third vs. First level							
(Comparison 1) Level	Trust	.025	.057	.432	.666	-.182	.058	-3.124	.002				
	Negative interpersonal processes	.007	.068	.107	.915	-.100	.066	-1.523	.128				
	Importance towards anonymity	.027	.057	.471	.638	-.016	.057	-.289	.773				
	Accuracy	-.335	.060	-5.619	<.001	-.588	.061	-9.654	<.001				
	PA Educ. Value	.014	.047	.295	.384	-.251	.046	-5.504	<.001				
		Twice vs. Once				Thrice vs. Once				More than thrice vs. Once			
(Comparison 2) PA Experience	Trust	.144	.078	1.858	.063	.150	.084	1.776	.076	.281	.066	4.235	<.001
	Negative interpersonal processes	.026	.082	.321	.748	.104	.091	1.141	.254	-.025	.071	-.357	.721
	Importance towards anonymity	-.131	.074	-1.776	.076	-.182	.080	-2.286	.022	-.235	.064	-3.648	<.001
	Accuracy	.070	.075	.941	.347	.150	.079	1.898	.058	.122	.063	1.952	.051
	PA Educ. Value	.280	.061	4.585	<.001	.318	.068	4.659	<.001	.488	.040	12.244	<.001
		Females vs. Males											
(Comparison 3) Gender	Trust	.024	.046	.506	.613								
	Negative interpersonal processes	.336	.053	6.285	<.001								
	Importance towards anonymity	.110	.046	2.402	.016								
	Accuracy	.047	.047	1.009	.313								
	PA Educ. Value	.177	.039	4.495	<.001								

Students in the second (n = 1046) and third level (n = 1017) have significantly lower score for *Accuracy* compared to students in level one (n = 981). The effect sizes of the mean differences are however moderate, respectively d = .34 and d = .59. Students in the third level show a significantly lower mean in *Trust* than students in level one. Nevertheless, the effect size is small (d = .18).

Students that had experienced PA three times (n = 478) attributed significantly less '*importance towards anonymity within PA'* than students with only one experience (n = 554) (small effect d = .18). Students with more than 3 PA experiences (n = 1430) had a significant lower mean concerning *Importance towards anonymity within PA* than group once (small effect d = .28), and a significantly higher mean in *Trust* than group once (small effect d = .24) and attributed *Educational value of PA* (medium effect d = .49).

Girls (n = 1665) show a significantly higher mean in *Negative interpersonal processes* (d = .34) and *Importance towards anonymity within PA* compared to boys.

Regarding the attributed *educational value of PA,* students in the third level attributed significantly lower value towards PA than students in the first level (d = .25). On the contrary, the more experience students have, the more positive attitudes towards PA they report (d = .28; d = .32; d = .49). Overall, females reported higher educational value of PA compared to males (d = .18). All the remaining comparisons were not significant.

Discussion

This cross-sectional survey study explored Flemish secondary education students' self-reported experiences with PA, with a specific focus on the inherent social nature of the activity. The majority of this 3066 students sample had previous experience with PA activities. This is in high contrast with previous studies (Noonan and Duncan, 2005; Panadero et al. 2016). More specifically, here PA was frequently used in all levels of secondary education, with the majority of students indicating that they experienced it multiple times. The

nature of the PA activities was relatively similar: most PA activities took place during the lesson and, if grades were awarded, they were not accounted for in a summative manner. This possibly mirrors a formative PA approach.

Students were in most cases not trained to perform PA, which goes against frequently mentioned PA practice guidelines (Sluijsmans 2002; Panadero et al. 2016). However, students on all levels were frequently involved in defining assessment criteria, an assessment scaffold which previous research has shown to be conducive for the overall quality of PA activities as it clarifies expectations and recognizes students' as valuable and active actors in the assessment process (Panadero & Romero, 2013). Regarding anonymity the great majority of students indicated that PA activities were conducted in a non-anonymous mode.

In this work, we used Structural Equation Modelling as a tool for understanding the relationship between the latent constructs of interpersonal variables, anonymity, perceived accuracy and students' attributed educational value towards PA. This approach allowed us to find relevant relations between the studied constructs, and to study the differences between the groups. However, it should be noted that this is the first large-sample based exploration of this understudied problem, and the model presented here should not be considered as a universal model. More research is needed with samples from different countries and ages for checking whether the latent structure and differences found here are consistently found in other scenarios. Our results lead us to think that four latent factors which are relevant to predict students' educational value towards PA: (1) *Trust in evaluative capabilities*, (2) *Negative interpersonal processes*, (3) *Accuracy* and (4) *Importance towards anonymity within PA.* In contrast to a previous study by van Gennip et al. 2010, *Value Congruency* and *Psychological Safety* were not needed for explaining the students' conceptions. A possible reason could be the fact that in Van Gennips' (2010) quasi-experimental design students' perceptions were measured directly after the intervention they conducted, causing a stronger

effect of these variables. According to our multiple regression SEM model, 54.8% of the variance in *Educational value towards PA* can be explained solely by the other four factors. We conducted several tests for measurement invariance, providing support for scalar invariance for all groups of interest, that is, educational level, PA experience and gender. The analysis revealed that the three factors (1) *Trust in evaluative capabilities*, (2) *Negative interpersonal processes* and (3) *Accuracy* were positive contributors to students' perceived educational value towards PA.

In line with previous research, reciprocal trust in the assessment skills and capability of peers to give a judgement on your work has proven to be important within the PA process (van Gennip et al., 2010), as it may lead to deeper learning approaches (Cheng & Tsai, 2012). Our model showed that the amount of *trust in your own and peers' evaluative capabilities* positively contributes to students' perceived educational value of PA. However, the fact that students in the third level had a significantly lower amount of trust compared to level one raises concerns. A possible explanation is that current PA implementation in the higher levels of education is still predominantly aimed at scoring, although it is not summatively accounted for, instead of pursuing a 'deep' approach with much more attention to the peer feedback component, as was also reported by Panadero & Brown (2016). This might also explain the significant lower amount of *perceived educational value to PA* between level three and one.

Contrary to our expectations, *negative interpersonal processes* were found to be a positive predictor of students' perceived *educational value of PA*. Possibly the items in our questionnaire were posed rather descriptively, in the sense that they rather indirectly refer to a value concept stronger than 'fear' and 'unfairness' as currently used. However, as the correlations between the latent predictors indicate a correct interpretation of the items, it is up to future research to deepen our understanding on the interrelationship between these variables. As a consequence, current findings suggest that students' awareness level about the

fact that these processes – including their potential undesirable effects - are possibly present in PA, leads to greater value on peer assessment as valuable learning activity. Test of mean latent differences revealed that girls rated significantly higher on this variable.

Perceived *accuracy* proved to be the most important positive predictor in the structural model. However, as students in level 2 and 3 reported significantly lower rates on perceived accuracy compared to level 1, again a predominant approach towards summative scoring alone instead of stimulating interactive feedback processes could be an explanation for this finding. One possible way to counter this decrease might be to consequently train students how to participate in PA tasks in order to increase (perceived) accuracy even when they are in the higher levels of secondary education.

Regarding the relation to *students' educational value of PA* and the factor *importance towards anonymity*, a negative relationship was found. This confirms previous theoretical work on PA by Topping (1998) who indicated that privacy is an important structural component of PA and confirms findings of small scale studies by for example Raes et al. (2013) and Vanderhoven et al. (2015) who found that students preferred anonymous modes of PA within face-to-face settings. The use of anonymous modes of PA was in general very low in the studied sample. Apparently Flemish teachers are either not aware that attributed importance towards anonymity can influence the outcomes of a PA activity (Vanderhoven et al., 2015) or, as found by Panadero et al. (2016), teachers believe in using anonymous PA formats but currently do not implement such formats. Mean differences between the experience categories revealed that students' amount of importance towards anonymity was lower for students with three or more than three PA experiences. This finding suggests that practice leads to more willingness to participate in non-anonymous PA settings. In these settings assessors and assessees can interact and preferably provide rich feedback. Interestingly, girls attributed higher importance towards anonymity. As a consequence, one

could advocate that offering anonymity can be used as a temporary catalyst to a) accelerate the creation of a safe PA environment and b) acknowledge the time that students need to get accustomed to the interpersonal processes that are evoked through participation in PA and which might possibly be mitigated through anonymity (Vanderhoven et al.,2015).

Implications

This study adds to our understanding of the impact of students' perceptions on interpersonal processes, accuracy and anonymity and their relation with the perceived educational value of PA. It has important consequences for educational practice. First of all, trust building in own and others' evaluative capabilities should be seen as an essential step when implementing PA activities. Second, teachers should discuss the inherent social nature of PA (Panadero, 2016), its educational value and the potential negative interpersonal processes (e.g., friendship marking, hostile behavior) before the start of the actual PA task in order to raise students' awareness on these issues as they positively influence students' PA conceptions. Moreover, teachers should give suggestions how to counter these negative effects, for example by exploring students' attributed importance towards anonymity, and create a safe learning environment in their classrooms. Third, it is important that teachers with mostly girls in their classes pay extra attention to the creation of a positive atmosphere that counters negative processes. Additionally, as girls seem to prefer anonymous PA modes and active guidance towards more interactive non-anonymous modes is needed with them.

It is up to future research to deepen our insights in these topics, but it is clear that implementation of different assessment scaffolds (e.g., active involvement in criteria development, training and trust building) to improve the educational potential of PA will take a considerable amount of time. Therefore, thoughtful implementation of PA by teachers is needed. Evidently, as also advocated by Brown and Panadero (2016), pre-service and in-service teachers should be trained in PA implementation and more importantly have

repeatedly experienced this themselves, in order to achieve insight in the effects of interpersonal processes that are inherent to this assessment method.

Limitations and future lines of research

This cross-sectional study's major limitation is its self-reported nature. This implies that students' answers may have been influenced by social desirability, as it is a risk with any form of subjective data collection (Desimone, 2009). However, throughout the process of survey development and administration, several steps were taken to reduce social desirability bias. This included extensive piloting, critical reviews and pretesting by an expert in formative assessment. Moreover, confidentiality for respondents was assured. Another matter of concern is the vague quantity 'sometimes' in the response scale of the items regarding RQ1. Evidently, there are possible memory problems within individuals when recalling what specific characteristics the PA activities they have experienced had and there is variability between individuals in how such a vague frequency is understood.

In this study we did not take into account students' general conceptions of assessment and their relation or impact on our structural model, although a reciprocal and influencing relationship between them is plausible. It is up to future studies to look into this so far under researched relationship. Additionally, this survey study did not take into account the teachers' opinions on the researched variables. Again, future research could focus on the linkage between the assessment and PA perceptions of these two actors.

Conclusions

Even though the interest in the inherent social nature of PA processes has increased over the past decade (e.g. Panadero, 2016), most of the experimental studies have been small-scale studies. This large sample survey study has confirmed the complex nature of PA processes which can trigger powerful feelings in our students and have an impact in their

beliefs in the educational value of PA. The four identified predictors of students' perceived educational value of PA can be of guidance for program developers, instructional designers, teacher educators and teachers when designing and implementing PA tasks. The results confirm Panadero's (2016) theoretical work on the social nature of PA: PA does not happen in a vacuum and a shallow implementation of it might do more harm than good. Mitigating the interpersonal variables will ask for intensive, repeated, highly interactive PA tasks in which interpersonal processes are actively monitored. Additionally, sufficient classroom time for the improvement of students' peer feedback skills should be made. Therefore, a structural integration of PA activities within the curriculum is strongly recommended.

References

Ajzen, I. (2005). *Attitudes, Personality, and Behavior* (2nd ed.). New York: Open University Press.

Assessment Reform Group. (2002). *Assessment for Learning: 10 principles research-based principles to guide classroom practice*. Retrieved from: http://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment & Evaluation in Higher Education*, *27*(5), 427–441. doi:10.1080/0260293022000009302

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. doi:10.1080/0969595980050102

Bolzer, M., Strijbos, J. W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer-feedback via eye-tracking: role of feedback-characteristics, fixation-durations and transitions. *Journal of Computer Assisted Learning*, *31*(5), 422–434.

doi:10.1111/jcal.12091

Boud, D. (1995). Assessment and learning: Contradictory or complementary? In K. Peter (Ed.), (pp. 35–48). London: Kogan Page.

Brown, G., & Harris, L. (2011). Student Conceptions of Assessment by Level of Schooling: Further Evidence for Ecological Rationality in Belief Systems. *Australian Journal of Educational & Developmental Psychology, 12*, 46–59.

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice, 15*(1), 3–17. doi:10.1080/09695940701876003

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210–220. doi:10.1016/j.tate.2010.08.003

Brown, G. T. L., McInerney, D. M., & Liem, A. D. (2009). Students Perspectives of Assessment: Considering What Assessment Means to Learners. In D. M. McInerney, G. T. L. Brown, & A. D. Liem (Eds.), *Student Perspectives on Assessment: What Students Can Tell Us about Assessment for Learning* (pp. pp. 1–21). Charlotte, NC: Information Age Publishing.

Carless, D. (2013). Trust and its role in facilitating dialogic feedback. In *Feedback in Higher and Professional Education* (pp. 90–103). doi:10.4324/9780203074336

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education, 35*(5), 551–564. doi:10.1080/02602931003632381

Cheng, K. H., & Tsai, C. C. (2012). Students' interpersonal perspectives on, conceptions of

and approaches to learning in online peer assessment. *Australasian Journal of Educational Technology*, *28*, 599–618.

Cowie, B. (2009). My Teacher and My Friends Helped Me Learn: Student Perceptions and Experiences of Classroom Assessment. In D. M. McInerney, G. T. L. Brown, & A. D. Liem (Eds.), *Student Perspectives on Assessment: What Students Can Tell Us about Assessment for Learning* (pp. pp. 85 – 105). Charlotte, NC: Information Age Publishing.

Cowie, B., Moreland, J., & Otrel-Cass, K. (2013). *Expanding Notions of Assessment for Learning: Inside Science and Technology Primary Classrooms*. Springer Science & Business Media.

De Backer, F., & Philips, I. (2013). CTO/SDL Toolkit Competenties Nederlands Breed Evalueren. Retrieved from http://www.ond.vlaanderen.be/toetsenvoorscholen/toolkit_breed_evalueren/pdf/ToolkitInZijnGeheel.pdf

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher*, *38*(3), 181–199. doi:10.3102/0013189X08331140

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*, 331–350. doi:10.1080/03075079912331379935

Falchikov, N., & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, *70*(3), 287–322. doi:10.3102/00346543070003287

Flemish Goverment. (2015). *Flemish Education in Figures*. Retrieved from

https://www.vlaanderen.be/nl/publicaties/detail/flemish-education-in-figures-2014-2015

Ghent University – Department of Educatoinal Studies. (2013). *Wiskundige geletterdheid bij 15-jarigen - Overzicht van de eerste Vlaamse resultaten van PISA2012*. Retrieved from http://www.pisa.ugent.be/uploads/assets/106/1396273183438-KORTE BROCHURE PISA2012.pdf

GO! (2012). *Evaluaren in het secundair onderwijs*. Retrieved from http://pro.g-o.be/blog/Documents/GO!_Visietekst_evalueren in het SO_DEF.pdf

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 36,* 101–111. doi:10.1016/j.tate.2013.07.008

Harris, L. R., Brown, G. T. L., & Harnett, J. A. (2014). Understanding classroom feedback practices: A study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation and Accountability*, *26*(2), 107–133. doi:10.1007/s11092-013-9187-5

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*, 21–27. doi:10.1016/j.stueduc.2012.04.001

Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to Promote Peer Feedback: Pre-Service Teachers' Comments in Asynchronous Computer-Mediated Communication. *Journal of Educational Computing Research*, *43*(1), 89–112. doi:10.2190/EC.43.1.f

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501–517. doi:10.1080/02602931003786559

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. Practical Assessment, Research & Evaluation, 10(17).

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568. doi: 10.1037/0021-9010.93.3.568

Muthén, L., & Muthén, B. (2007). *Mplus user's guide (version 7)*.

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Human factors and social conditions of assessment. New York: Routledge* (pp. 1–39). New York, NY: Routledge.

Panadero, E., & Brown, G. T. L. (2016). Teachers' reasons for using peer assessment: positive experience predicts use. *European Journal of Psychology of Education*. doi:10.1007/s10212-015-0282-5

Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). Enhancing ownership, relinquishing control: Scaffolding students regulated learning through involvement via Assessment for Learning. In D. Laveault & L. Allal (Eds.), *Assessment for Learning: Meeting the challenge of implementation.*

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 133–148. doi:10.1080/0969594X.2013.877872

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005

Reinholz, D. (2015). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, 1–15. doi:10.1080/02602938.2015.1008982

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research, 99,* 323-338. doi:10.3200/JOER.99.6.323-338

Schweizer, K. (2010). Some Guidelines Concerning the Modeling of Traits and Abilities in Test Construction. *European Journal of Psychological Assessment, 26*, 1-2. doi:10.1027/1015-5759/a000001

Sluijsmans, D. M. A. (2002). Student involvment in assessment: The training of peer assessment skills. (Unpublished doctoral dissertation). Open University of the Netherlands, Heerlen.

Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation, 37*(1), 49–54. doi:10.1016/j.stueduc.2011.03.008

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review1. *Assessment & Evaluation in Higher Education, 30*(4), 325–341. doi:10.1080/02602930500099102

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), (pp. 55–87). Dordrecht, The

Netherlands: Kluwer Academic.

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 339–343. doi:10.1016/j.learninstruc.2009.08.003

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*, 41–54. doi:10.1016/j.edurev.2008.11.002

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*, 280–290. doi:10.1016/j.learninstruc.2009.08.010

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education, 81*, 123–132. doi:10.1016/j.compedu.2014.10.001

Vickerman, P. (2009). Student perspectives on formative peer assessment: an attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*, 221–230. doi:10.1080/02602930801955986

Vlaanderen, K. O. (2013). Evalueren en leren: de kip of het ei?, 10–13. Retrieved from https://pincette.vsko.be/VSKO/Documentatiedienst/Tijdschriften/Breedbeeld/Publiek/2012-2013/Breedbeeld 2012-2013-3 Dossier.pdf

Yu, F.-Y., & Sung, S. (2015). A mixed methods approach to the assessor's targeting behavior during online peer assessment: effects of anonymity and underlying reasons. *Interactive Learning Environments*, 1–18. doi:10.1080/10494820.2015.1041405

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-Regulated Learning and Academic*

Achievement: Theoretical Perspectives. Retrieved from

https://books.google.com/books?hl=nl&lr=&id=og4hVOcjcqMC&pgis=1

Appendix 1
Students' conceptions about PA

Definition: In a PA-activity students judge each other's tasks/presentations/group assignments. The judgement can be expressed in scores, oral or written feedback or a combination of both.		
#	Question	Response Format
1	Have you ever experienced PA?	Once-Twice-Thrice-> Thrice - None
2	Via what tool the PA was conducted?	Paper – Oral – Computer - Combination
3	Did the results of the PA accounted for the monthly report?	Y/N/Sometimes
4	At what moment the PA-activity took place?	During / End series of lessons
5	Anonymity was guaranteed towards: a) Assessors b) Assessees c) Teachers	Y/N/Sometimes in each of three categories
6	Were you trained to do PA?	Y/N/Sometimes
7	Were you involved in defining the assessment criteria?	Y/N/Sometimes
8	Was the result of the PA orally discussed with the teacher?	Y/N/Sometimes
9i	Negative interpersonal - friendship marking: During the PA activity people were doing friendship marking, in which friends were deliberately giving only high scores and positive feedback	6 point Likert-scale (2 negatively formulated, 4 positively formulated)

9ii	Negative interpersonal - fear of disapproval: I fear that my peers would no longer like me if I would give them a low score or negative feedback	Idem
9iii	Interpersonal trust peer : I trust the evaluative skills of my peers	Idem
9iv	Interpersonal trust own: I trust my own evaluative skills	Idem
9v	Interpersonal psychological safety: I feel comfortable giving my opinion on a peers' work for the whole class group	Idem
9vi	Interpersonal value congruency: Everyone was interpreting the evaluation criteria in the right way	Idem
10i	Anonymity: I think it is important that my peers don't know that the evaluation that was given, was given by me*. * your anonymity as an assessor is guaranteed	Idem
10ii	Anonymity: I don't want to know the identity of the person whose work I'm evaluating.	Idem
10iii	Anonymity: I think it is important that the teacher doesn't know to who I'm giving an evaluation.	Idem
11i	Accuracy: My peers are capable to give an accurate judgement	Idem
11ii	Accuracy: Evaluation through PA is accurate	Idem
12i	PA conceptions: Through my participation in PA I get to know my strengths and weaknesses	Idem
12ii	PA conceptions: PA is useful	Idem
12iii	PA conceptions: It is instructive for pupils to give scores and feedback to peers	Idem