The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance

Ernesto Panadero [1 2], Margarida Romero [2], and Jan-Willem Strijbos [3]

Author Note

[1] Department of Educational Sciences and Teacher Education. Learning and Educational Technology Research Unit (LET), University of Oulu, Finland.

[2] Departament de Psicologia Bàsica, Evolutiva i de l' Educació, University Autònoma de Barcelona, Barcelona, Spain

[3] Department of Psychology. Ludwig-Maximilians-University, Munich, Germany

Recommended citation:

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

Correspondence concerning this article should be addressed to: Ernesto Panadero, Department of Educational Sciences & Teacher Education, Learning and Educational Technology Research Unit (LET Team) University of Oulu PO BOX 2000. Finland-90014. E-mail: ernesto.panadero.uam@gmail.com

Acknowledgements: Research funded through grant to Ernesto Panadero by the Alianza 4 Universidades and EUROCAT project (FP7 Program).

Abstract

Construct validity of peer assessment (PA) is important for PA application, yet difficult to achieve. The present study investigated the impact of an assessment rubric and friendship between the assessor and assessee on construct validity of PA. Two-hundred nine bachelor students participated: half of them assessed a peer's concept map with a rubric whereas the other half did not use a rubric. The results revealed a substantial reliability and construct validity for PA. All students over-score their peers' performance, but students using a rubric were more valid. Moreover, when using a rubric a high level of friendship between assessor and assessee resulted in more over-scoring. Use of a rubric resulted in higher quality concept maps for peer and expert ratings.

Keywords: Peer assessment; Reliability; Construct validity; Rubric; Friendship.

The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance

In the past decades peer assessment (PA) has been promoted in higher education as a valuable approach to formative assessment (Falchikov, 2003, 2005; Falchikov & Goldfinch, 2000). PA is an educational arrangement where students evaluate a peer's performance with scores, and/or with written or oral feedback (Topping, 1998). PA is also regarded as a specific type of collaborative learning (Boud, Cohen, & Sampson, 1999; Strijbos, Ochoa, Sluijsmans, Segers, & Tillema, 2009).

Despite positive effects of PA, such as increased 'perceived learning', essay and writing revision and presentation skills (Topping, 2003; Van Gennip, Segers, & Tillema, 2009), a consistent challenge for the acceptance of PA among students and teachers is the reliability and validity of PA (Cho, Schunn, & Wilson, 2006; Dochy, Segers, & Sluijsmans, 1999; Falchikov & Goldfinch, 2000; Van Zundert, Sluijsmans, & Van Merriënboer, 2010). Typical sources that can reduce reliability and validity, such as poorly designed assessment instruments or lack of assessor training, might be alleviated via structured assessment scaffolds, for example an assessment rubric. However, the reliability and validity of PA can also suffer from a friendship between the assessor and assessee. Although friendship is typically addressed in terms of 'reciprocity effects' in PA literature, it has remained an under-researched area. Furthermore, the value of rubrics to alleviate the potential scoring bias due to friendship has not been systematically investigated.

The following sections will elaborate on rubrics and evidence of their use in PA, the influence of friendship on PA (in collaborative learning in general and within PA specifically), the potential influence of a rubric on perceived fairness and comfort with PA, and the quality of the performance (concept map).

Rubrics for assessment purposes

A rubric articulates the expectations for an assignment by listing the assessment criteria and by describing levels of quality in relation to each criterion (Andrade & Valtcheva, 2009). Historically, research on rubrics has followed either a summative or a formative approach (Panadero & Jonsson, 2013). The summative approach aims to increase the inter-rater reliability and intra-rater reliability of assessors (Jonsson & Svingby, 2007; Stellmack, Konheim-Kalkstein, Manor, Massey, & Schmitz, 2009). The formative approach applies rubrics to enhance students' learning by promoting reflections on their own work (Andrade & Valtcheva, 2009; Panadero, Alonso-Tapia, & Reche, 2013) or the work by a peer (Sadler & Good, 2006). Irrespective of a summative or formative approach, indicators for reliability and validity must be must be examined for the use of rubrics in general, as well as for their use with PA in particular.

Indicators for reliability and validity of rubrics and for construct validity of rubrics in PA

Reliability can be expressed in general as inter-rater and intra-rater agreement, both of which in turn can be expressed as consensus agreement (often also referred to as objectivity) or consistency agreement. Validity typically addresses the degree to which a performance in an educational setting is captured by a specific measurement instrument; also referred to as the 'content aspect' (Messick, 1995) of construct validity. Finally, in the assessment literature the term 'accuracy' is also used (e.g., Brown et al., 2004), but refers to overall psychometric quality combining the reliability and validity indicators – as such accuracy is at present too ill-defined for practical application.

Inter-rater consensus. Inter-rater agreement in terms of consensus refers to different assessors awarding the same score to the same performance (Brown, Glasswell, & Harland, 2004), which is typically expressed as percent exact agreement,

adjacent agreement (same score plus or minus one category), Cohen's kappa and/or Krippendorff's alpha to correct for agreement by chance, and intraclass correlations (ICC) when each assessor also evaluates multiple performances (Cho et al., 2006). In their review on reliability and validity of rubrics, Jonsson and Svingby (2007) found that more than half of the studies (40 of 75) reviewed reported inter-rater reliability. In terms of percentage exact agreement with rubrics, Jonsson and Svingby (2007) found that the majority of consensus agreement varied from 55-75%, but in terms of adjacent agreement the consensus increased to 90%. In general, the 70% criterion is used for percent agreement, and for Cohen's Kappa and Krippendorff's Alpha, values between .40 and .75 represent a fair agreement beyond chance. Jonsson and Svingby (2007) concluded that the number of levels of a rubric directly affects consensus agreement.

Inter-rater consistency. Inter-rater agreement in terms of consistency refers to the pattern in the distribution of scores across a set of assessors and is expressed by a Pearson (or Spearman) correlation in the case of two assessors or by a Cronbach's alpha or intraclass correlations when each assessor evaluates multiple performances. Jonsson and Svingby (2007) observed in the case or rubrics that the majority of studies reported a Pearson's or Spearman correlation which varied between .55 and .75 (where in values of .70 and above are considered acceptable). With more than two assessors Cronbach's alpha is reported, and Jonsson and Svingby (2007) found eight studies with coefficients ranging .50 to .92 (with .70 as the threshold for an acceptable consistency). Overall, in the studies reviewed by Jonsson and Svingby, the inter-rater agreement (consensus and consistency) varied extensively.

Intra-rater consensus and consistency. Intra-rater consensus agreement refers to the reliability of a single assessor when scoring the performance of an assessee at two different occasions, and it can be expressed by a Pearson's or Spearman's correlation or

Cohen's Kappa and Krippendorff's alpha. However, this type is rarely reported – instead most studies on rubrics report intra-rater consistency (Jonsson & Svingby, 2007). Intra-rater consistency refers to the reliability of an assessor when scoring the performance of different assesses at a single occasion (Cronbach's alpha), for example a teacher scoring the performance of all students. In their review on rubrics, Jonsson and Svingby (2007) found only seven studies that reported Cronbach's alpha as a consistency indicator for intra-rater agreement; the majority reported a value over .70, considered as sufficient (Brown et al., 2004).

Construct validity. Another important aspect to consider is construct validity of assessment. In the case of scores on performance assessment, construct validity is based on what Kane (2001) refers to as 'observable attributes' and such scores typically reflect what Messick (1995) refers to as the content aspect of construct validity which includes content relevance and representativeness. In education it is not a property of a test or an assessment, but rather an interpretation of the outcomes (Jonsson & Svingby, 2007). It is typically determined via Pearson correlations of different raters (usually experts) that use the same instrument to measure the same construct.

In PA construct validity can be determined by comparing the peer score(s) with the teacher score(s), which results in construct validity indicators from both the teacher perspective and student perspective (see Cho et al., 2006). PA validity from the teacher perspective (TPv) (i.e., a pattern in the scoring distribution) is typically expressed as a Pearson correlation (for pairs) or Cronbach's alpha (with more than two peer assessors), where the teacher assessment serves as the unbiased baseline. Validity from the student perspective (SPv) is examined by the deviation between a peer (or multiple peers) and teacher assessment, which reveals the degree of under- or over-scoring when compared to the unbiased teacher score (Wang & Imbrie, 2010). Cho et al. (2006) advocate to take

"the square-root (...) of the sum of the squared differences divided by the number of peer ratings minus 1 (to produce an unbiased estimate) (...) the higher the score, the lower the perceived validity" (p. 896). Yet, this calculation provides the average deviation to the expert and not the expert-peer deviations at the individual level for each peer. As such variance within research conditions will be lost. Thus, the expert-peer deviation at the individual level constitutes a second student perspective of validity in PA, which also enables the determination of over-scoring and under-scoring by peer relative to the expert. The average deviation to the expert will be referred to as SPv1, and the expert-peer deviation at the individual level will be referred to as SPv2.

Measurement estimates. The application of measurement estimates spans both inter-rater reliability, intra-rater reliability, as well as (construct) validity aspects. Their application thus depends on the indicators reported as relevant for a specific study and research purpose, although in general these techniques (multi-faceted Rasch model or Generalizability Theory) are applicable to address most of the indicators.

Application of rubrics in PA

In PA it is crucial that the assessor knows the assessment criteria in order provide a reliable and valid assessment (Panadero & Jonsson, 2013). Rubrics are particularly helpful as they provide the assessment criteria in a structured format. Moreover, when conducting PA students are (a) reluctant to being assessed by a peer who is not an expert in the domain, and (b) they believe that assessment is the responsibility of the teacher (Ballantyne, Hughes & Mylonas, 2002). Rubrics hold the potential to alleviate both of these issues (Hafner & Hafner, 2003) and thus enhance perceived fairness and comfort with PA.

Nevertheless, research on the application of rubrics to PA is scarce and predominantly focused on inter-rater agreement between peers and validity from the

teacher perspective. The majority of studies revealed that rubrics increase the inter-rater reliability and validity of PA from the teachers' perspective (i.e., consistency in terms of a Pearson correlation or Cronbach's alpha; see for example Hafner & Hafner, 2003; Sadler & Good, 2006; Tsai & Liang, 2009, Tseng & Tsai, 2007), yet few studies have focused explicitly on the effect of a rubric-supported PA on students' performance (Panadero & Jonsson, 2013). Moreover, the potential of rubrics to decrease over-scoring due to friendship has not been systematically explored. Since PA is regarded as a specific type of collaborative learning (Boud et al., 1999; Strijbos et al., 2009), the following sections will elaborate on friendship in collaborative learning in general and within PA specifically.

Friendship in collaborative learning

Friendship has been of interest in collaborative learning with respect to group formation, organization of the activity and assessment of the collaboration (Barron, 2003; Jones & Issroff, 2005). Wang and Imbrie (2010) for example observed that students – when offered to assemble their own groups – choose peers with whom they were better befriended. In addition, they observed that group-member selection on the basis of friendship positively influenced interdependence, individual accountability and sense of community – leading to a warmer atmosphere and a better collaborative learning process. Wang and Imbrie (2010) attributed this effect partly to implicit rules of collaboration and trust that already exist between friends. Vass (2002) observed the same implicit collaboration rules between friends in the context of creative collaborative writing. Finally, Tolmie, Topping, Christie, Donaldson, Howe, Jessiman, Livingston and Thurston (2010) observed that friendship pairs performed better on collaborative assignments and showed a better ability to be critical and develop their peers' ideas when the task was challenging.

Friendship in PA and its relation to perceived comfort and fairness

Within the PA literature the term 'reciprocity effects' is typically used to refer to bias in PA caused by interpersonal processes (Strijbos et al., 2009). Specific indicators for reciprocity when conducting PA within groups are collusive marking (high ratings to fellow group members), decibel marking (high ratings to dominant group members) and parasite marking (profiting from the efforts invested by fellow group members) (Cheng & Warren, 1997; Harris & Brown, 2013; Pond & Ul-Haq, 1997). Magin (2001) is one of the rare studies that explicitly investigated reciprocity effects, operationalised as follows: if person A rates person B higher than expected, person B also rates Person A higher than expected. Reciprocity effects were found to be minuscule (1% explained variance), however, the operationalisation only considered directionality and excluded other aspects interpersonal processes such as friendship.

Although friendship has been acknowledged as a potential bias in PA, empirical studies that specifically investigate the effect of friendship are scarce. Friendship has been acknowledged as a potential bias in peer assessment and could lead to over-scoring (Pond & Ul-Haq, 1995; Strijbos et al., 2009). Students indeed typically prefer not to assess their friends too harshly (Cheng & Warren, 1997) and/ or fear that "(…) other students did not take the exercise seriously and might have been cheating by favouring friends" (Smith, Cooper, & Lancaster, 2002, p. 74). A higher degree of friendship could lead to over-scoring (Pond & Ul-Haq, 1995) and negatively affect the reliability, validity, and perceived fairness of assessment (Sambell, McDowell, & Brown, 1997). Moreover, teachers consider friendship as a negative factor influencing fairness and thus reducing the reliability and validity of PA (Karaca, 2009). For example, Papinczak, Young and Groves (2007, p. 180) observed that "(…) a strong reaction to peer assessment was the widespread perception that this process could be corrupted by bias

due to friendship marking or lack of honesty" (p. 180). Similarly, Harris and Brown (2013) conducted interviews with K-12 teachers and found that one of them (as well as her students) specifically reported the experience that friendship(s) resulted in biased PA.

In more recent PA literature specific interpersonal variables have been identified such as psychological safety and trust, as well as the impact of structural features (i.e., type of peer interaction and group constellation; Van Gennip et al., 2009). Trust and psychological safety are particularly relevant processes in relation to friendship. First of all, trust during PA positively influences students' conceptions about PA (Van Gennip et al., 2009) and could foster a critical analysis of a peers' performance (Jehn & Shah, 1997; Tolmie et al., 2010). Secondly, friendship can foster a psychologically safer atmosphere for criticising peers (MacDonald & Miell, 2000), which in turn fosters interpersonal risk-taking in a group: "(…) a sense of confidence that the team will not embarrass, reject, or punish someone for speaking up" (Edmonson, 1999, p. 354).

A high degree of trust and psychological safety is likely to result in perceived comfort with PA on the part of the students, irrespective of the degree of friendship. In fact, several researchers have emphasised that stress and (dis)comfort should be considered when determining the impact of PA on students' emotions (Hanrahan & Isaacs, 2001; Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006; Pope, 2001, 2005). Comfort with the procedure of assessing peers and being assessed by peers might enhance the reliability and validity of PA. Finally, specificity and a common understanding of the criteria (e.g., by a rubric) might also increase students' comfort with PA.

Application of rubrics to assess concept map performance

Concept mapping is a learning strategy that increases students' performance and is an effective technique to evaluate students' domain knowledge (Nesbit & Adesope, 2006) and especially the structure of declarative knowledge (Shavelson, Ruiz-Primo, & Wiley, 2005). A concept map is defined by Shavelson et al. (2005) as "(…) a graph in which the nodes represent concepts, the lines represent relations, and the labels on the lines represent the nature of the relation between concepts. A pair of nodes and the labeled line connecting them is defined as a proposition." (p. 417). Designing a concept map is cognitively demanding, yet leads to better learning and is therefore an interesting strategy to train among higher education students (Berry & Chew, 2008; Jacobs-Lawson & Hershey, 2002). The use of concept maps has been studied previously in combination with metacognitive activities (e.g., Hilbert & Renkl, 2009) and with the use of rubrics to enhance learning and the reliability and validity of students' self- and peer assessment (Besterfield-Sacre, Gerchak, Lyons, Shuman, & Wolfe, 2004; Moni, Beswick, & Moni, 2005; Moni & Moni, 2008; Panadero, Alonso-Tapia, & Huertas, 2014; Toth, Suthers, & Lesgold, 2002). Therefore, in this study using a rubric and peer assessment will add empirical evidence to the application of rubrics for assessment of concept maps and the extent to which rubrics promote learning gains.

Research questions and hypotheses

The current study investigates the potential moderating effect of the application of a rubric and degree of friendship between the assessor and assessee on the reliability and construct validity of PA, perceived comfort and fairness during PA, and quality of performance (concept map). This quasi-experimental study investigates two variations: (a) rubric vs. non-rubric and (b) low vs. medium vs. high level of friendship between the assessor and assessee. The research questions and hypotheses are as follows:

- RQ 1. What is the effect of a rubric on the reliability and construct validity of PA of a concept map? It is expected that use of rubrics results in a higher construct validity from the student perspective (i.e., smaller deviation between the peer score and the teacher score) compared to no use of rubrics (Hypothesis 1).
- RQ 2. What is the effect of a rubric and the level of friendship on the construct validity from the student perspective of PA of a concept map, and is there an interaction between the use of a rubric and the level of friendship? It is expected that rubrics reduce the bias of friendship (Hypothesis 2a). It is expected that students with a low level of friendship evaluate their peers more validly compared to students with a high level of friendship, who will show a tendency to favour their peers by over-scoring (Hypothesis 2b).
- RQ 3. What is the effect of a rubric on students' perceived comfort and fairness while conducting PA? It is expected that rubrics increase participants' comfort while performing PA (Hypothesis 3a) and increase perceived fairness of PA (Hypothesis 3b).
- RQ4. What is the effect of a rubric on the quality of the concept map in terms of the peer score and expert score? It is expected that a rubric will lead to a higher quality concept map in terms of the peer score (Hypothesis 4a), and to a higher quality concept map in terms of the teacher score (Hypothesis 4b).

Method

Participants

Two-hundred and nine third year pre-service bachelor student teachers at a public university in Spain participated in this study of which 182 were female (87.08%)

and 27 male (12.92%), and with a mean age of 22.17 (*SD* = 3.92). The high presence of females is representative for pre-service teacher programs.

Design

The students were enrolled in the "Learning and development II" course and recruited from four classrooms. Each classroom was taught by a different teacher. The students read an excerpt from an instructional text and subsequently constructed a concept map identifying the main concepts and relations between them. Students from two classrooms were randomly selected for the rubric condition (*N* = 104) and the other two classrooms assigned to the non-rubric condition (*N* = 105). Each student was randomly assigned to assess the concept map by one of his or her peers. The activity counted for students' course credit but not their actual performance in terms of the quality of the concept map.

Instruments and measures

Task. Students read an excerpt from the text "Estrategias docentes para un aprendizaje significativo" [Teachers' strategies for a meaningful learning] by Diaz Barriga (2002). The text is a mandatory reading from the official curriculum. They subsequently constructed a concept map of the text. This concept mapping task was selected due to their widespread application as a learning strategy (Nesbit & Adesope, 2006) and when combined with a rubric resulted in strategy enhancement and better performance (Panadero et al., 2014; Toth et al., 2002).

The students were required to extract instructional concepts that were related to declarative, procedural and attitudinal knowledge. Then, organize them in a hierarchical order and specify relationships between the concepts. Figure 1 provides an illustration of a concept map by an expert in which the expected components are indicated.

****

Insert Figure 1 about here

****

Concept map assessment. A concept map was assessed on five criteria: (1) concepts, (2) hierarchy, (3) relationships among concepts in different hierarchical levels, (4) relationships among concepts from different columns, and (5) simplicity and easiness of understanding.

Assessment by condition. Students and experts in the rubric condition used the same rubric to assess the concept maps. Students in the non-rubric condition rated each of the criteria on a 4-point Likert scale (1 = *poor performance* to 4 = *best performance*), whereas the experts used the rubric to ensure comparability of conditions to determine construct validity.

Rubric. The rubric was created using expert models of concept maps in other studies (Panadero, Alonso-Tapia, & Huertas, 2012). Each criterion was further specified into four levels ranging from the poorest (1) to the best (4) performance (see Appendix A for the rubric used). Inter-rater reliability for the experts was determined by three independent experts scoring 63 concept maps (31 from the rubric condition, 32 from the non-rubric condition) which were selected at random and represented 30% of all maps (> 10% minimum threshold; see Neuendorf, 2002). Consensus agreement was calculated for each criterion scored by three independent experts with Krippendorff's alpha (Hayes & Krippendorff, 2007), at the ordinal scale level: (1) concepts [.92, 95% CI: .88-.95], (2) hierarchy [.96, 95% CI: .93-.99], (3) relationships among concepts in different hierarchical levels [.90, 95% CI: .85-.95], (4) relationships among concepts from different columns [.94, 95% CI: .87-.99], and (5) simplicity and easiness of understanding [.97, 95% CI: .93-1.00]. Consensus agreement was calculated for the sum-scores by the three independent experts with Krippendorff's alpha (ordinal scale)

and proved to be excellent: .96, [95% CI: .94-.98]. Consistency agreement of the sum-scores for the maps rated was excellent as well (Cronbach's $\alpha = .99$). Once the scoring system was reliable, one of the experts scored the remaining concept maps.

Concept map score. The score for each assessment of a concept map – whether by a peer or the expert, and irrespective of condition – consists of the sum-score of all criteria and ranges between a minimum of 5 to a maximum of 20 points.

Friendship. The degree of friendship was determined with a single 9 point Likert-scale item: "My level of friendship with the peer whose concept map I have to evaluate is …" (1 = *enmity* to 9 = *close friends*). The item was based on Hays (1984) who used a 7-point Likert-scale. In line with three levels discerned by Hays (1984) and Rose and Serafica (1986) we applied a 9 point Likert-scale which could be easily grouped in three levels of friendship: low (1 – 3), medium (4 – 6) and high (7 – 9). Friendship levels were reported across the rubric and non-rubric condition as follows: low ($N = 23$; with 15 rubric and 8 non-rubric), medium ($N = 172$; with 78 rubric and 94 non-rubric) and high ($N = 14$; with 11 rubric and 3 non-rubric). The research conditions were equivalent in their mean level of friendship, $F(1, 207) = 0.60$, $p = .438$, $M_{\text{rubric}} = 4.90$ ($SD = 1.26$), $M_{\text{non-rubric}} = 4.79$ ($SD = 0.81$) .

Peer assessment construct validity. In line with Brown et al. (2004), Cho et al. (2006) and Wang and Imbrie (2010), construct validity from the teacher perspective (TP) was determined by a Pearson correlation between the peer and expert rating, with the expert rating as unbiased. Construct validity from the student perspective (SPv) was calculated as the deviation between the peer and expert score for each individual (SPv2; see the section on construct validity) instead of the average deviation to the expert (SPv1, see Cho et al., 2006). More specifically, SPv was calculated as the peer rating minus the expert rating. Therefore, a deviation value of zero represents that peer rating

does not deviate from expert rating. A positive deviation value represents over-scoring compared to the expert (i.e., peer rating was higher than the expert rating). A negative deviation value represents under-scoring compared to the expert (i.e., peer rating was lower than the expert rating).

Perceived comfort. Perceived comfort was determined with a single 7 point Likert-scale item specifically developed for this study: "My level of comfort when scoring the concept map from a peer is …" (1 = *none* to 7 = *high*).

Perceived Fairness. Perceived fairness was determined with a single 5 point Likert-scale item specifically developed for this study: "Do you believe that your peer will conduct a fair assessment of your concept map?" (1 = *no* to 5 = *yes*).

Procedure

Students were informed one week in advance that they would be asked to conduct a task during their seminar for which they were required to read a designated text-excerpt and that the activity would count for their course credit (but not their actual performance in terms of the quality of the concept map). Subsequently, the students were instructed to design a concept map from the text-excerpt they had read. The rubric was handed out to the corresponding condition and its application was explained. The rubric condition was asked to score each rubric's criterion independently and then add them into a sum-score. In the non-rubric condition the five assessment criteria were stated aloud (see criteria column of the Rubric): "When assessing a concept map an expert would consider the following features: all relevant concepts have to be included, the hierarchy has to be clearly defined". Next, the students in the non-rubric condition assessed their peers' concept map by awarding 1 to 4 points for each criterion with higher score reflecting a better performance, resulting in a possible sum-score of 5 to 20 points (similar to the rubric condition). All students then worked for 30 minutes on their

concept map. Afterwards they received the concept map by a peer with 15 minutes to assess it. In a follow-up session, students received their scores (peer- and expert assessment).

Results

Data inspection

Prior to the analyses distribution assumptions were checked. The standardised skewness and kurtosis were within the +3 and -3 criterion (Tabachnik & Fidell, 2001) for all variables, except for peer assessment SPv ($z_{skewness}$ = 3.5 (.589/.168); $z_{kurtosis}$ = 3.43 (1.149/.335) and friendship ($z_{kurtosis}$ = 3.81(1.242/.335). Since the standardised skewness for peer assessment SPv and friendship were not extremely outside the criterion range a normal distribution was assumed.

Checks for equivalence of conditions over classrooms

Firstly, prior to the intervention students' experience with concept map design was checked. A sample of concept maps (N = 62) previously created by students, evenly distributed over the rubric and non-rubric condition, revealed no significant differences (p = .652). Although the intervention was conducted by one of the researchers, analyses were performed on all depending variables comparing the four classrooms as they were taught by four different teachers. The following differences were found. *Expert rating*: rubric classrooms differ, $F(1, 102) = 9.19$, $p = .003$, $\eta^2 = .08$, $M_{group1} = 15.18$ (SD = 2.14), $M_{group2} = 13.94$ (SD = 1.99). *Peer rating*: non-rubric classrooms differ, $F(1, 103) = 4.25$, $p = .042$, $\eta^2 = .04$. $M_{group3} = 13.65$ (SD = 1.74), $M_{group4} = 14.44$ (SD = 2.18). *Peer assessment SPv*: non-rubric classrooms differ, $F(1, 103) = 4.53$, $p = .036$, $\eta^2 = .04$. $M_{group3} = 1.96$ (SD = 1.79), $M_{group4} = 2.77$ (SD = 2.09). *Peer assessment SPv*: rubric classrooms differ, $F(1, 102) = 5.27$, $p = .024$, $\eta^2 = .05$. $M_{group1} = 0.37$ (SD = 2.21), $M_{group2} = 1.40$ (SD = 2.38). Given the observed differences the instructional setting

within the courses was further examined. The four teachers used the same pedagogical method in a highly structure program, sharing their activities, lectures and assessment tasks, aimed at creating the same instructional setting for the four groups. Hence, the observed differences can be regarded as natural teacher variance and do not affect the internal validity of the study.

Reliability and construct validity of peer assessment from the teacher perspective

Each concept-map was assessed by one peer and one expert which is a special case where reliability and construct validity from the teacher perspective (TPv) are expressed by the same consistency indicator, that is, a Pearson correlation between the experts' rating and peers' rating. A moderate correlation was found ($r = .47$, $p < .001$; Cohen, 1988: .10 (small), .30 (medium), .80 (high), which reveals that peer assessment scores were overall fairly reliable and valid, but still reflect large individual variations. When split by condition the reliability and construct validity were moderate for students in the rubric ($r = .34$, $p < .001$) and non-rubric ($r = .38$, $p < .001$) condition.

Effect of rubrics and friendship level on peer assessment construct validity from the student perspective (SPv)

Since each concept-map was assessed by one peer, inter-rater reliability of peer ratings (consensus and consistency) could not be determined, since each concept-map in that case should have been assessed by a minimum of two raters. The peer assessment construct validity from the student perspective (SPv) was calculated as the peer rating minus the expert rating.

A two-way ANOVA was conducted to investigate main and interaction effects of rubrics and friendship levels on SPv of peer assessment. There was a main effect for the rubric versus non-rubric condition on SPv of peer assessment, $F(1, 203) = 5.66$, $p = .018$, $\eta^2 = .03$ (small effect; Cohen, 1988: .01 < .06 = small, .06 < .14 = medium, > .14

= large). Students in the rubric condition (M = 0.84, SD = 2.33) showed less deviation from the expert score compared to students in the non-rubric condition (M = 2.33, SD = 1.96). The lower deviation from the expert (i.e., a score closer to zero) reflects that students in the rubric condition were more construct valid than the non-rubric condition. Both groups tended to over score their peers as reflected by the positive values. No main effect was observed for friendship level on SPv, $F(1, 203) = 1.66$, $p = .193$, $\eta^2 = .016$ (small effect), $M_{low} = 2.18$ (SD = 0.46), $M_{medium} = 1.43$ (SD = .16), $M_{high} = 2.23$ (SD = 0.69).

The analysis revealed that the RUBRIC × FRIENDSHIP interaction for SPv of peer assessment was not significant, $F(2, 203) = 2.06$, $p = .131$, $\eta^2 = .018$ (small effect). The means and standard deviations by friendship level (low, medium, high) signalled a potential difference between friendship level in the rubric condition: *Low*: $M_{rubric} = 0.73$ (SD = 2.25); $M_{non\text{-}rubric} = 3.63$ (SD = 1.59), *Medium*: $M_{rubric} = 0.63$ (SD = 2.40); $M_{non\text{-}rubric} = 2.23$ (SD = 1.97), and *High*: $M_{rubric} = 2.45$ (SD = 1.29); $M_{non\text{-}rubric} = 2.00$ (SD = 2.00). Follow-up one-way ANOVA for each condition revealed no significant difference between friendship levels for the non-rubric condition, $F(2, 102) = 1.93$, $p = .151$. However, for the rubric condition there was a significant difference between the friendship levels, $F(2,101) = 3.08$, $p = .050$ and a contrast analysis comparing the low and medium levels combined to the high level of friendship was also significant, $t(101) = 2.33$, $p = .022$, $d = 0.98$ (large effect: > 0.80, see Cohen, 1988). There is more over-scoring in the rubric condition by students with a high level of friendship compared to the combined over-scoring by students with a low or medium level of friendship.

Perceived comfort and fairness of peer assessment

A one-way ANOVA revealed no significant difference for perceived comfort, $F(1, 207) = 0.08$, $p = .77$, between the rubric (M = 4.59, SD = 0.81) and non-rubric (M =

4.62, *SD* = 0.83) condition. Likewise, a one-way ANOVA revealed no significant difference for perceived fairness, $F(1, 207) = 0.00$, $p = .99$, between the rubric (*M* = 3.88, *SD* = 0.82) and non-rubric (*M* = 3.89, *SD* = 0.74) condition.

We explored the relations between perceived comfort, perceived fairness, PA rating, and the student perspective of PA validity (SPv) within each research condition. In the non-rubric condition comfort and fairness are weak but positively correlated ($r = .21$, $p = .030$). In the rubric condition a weak positive correlation was observed between fairness and the peer rating ($r = .23$, $p = .020$), and a moderate positive correlation between comfort and student perspective of validity ($r = .30$, $p = .002$).

Effect of a rubric on performance quality in terms of peer score and expert score

A one-way ANOVA revealed a significant difference for the quality in terms of the peer score, $F(1, 207) = 28.82$, $p < .001$, $\eta^2 = .12$ (medium effect), between the rubric (*M* = 15.45, *SD* = 1.90) and non-rubric (*M* = 14.01, *SD* = 1.98) condition.

A one-way ANOVA revealed a significant difference for the quality in terms of the expert score, $F(1, 207) = 133.93$, $p < .001$, $\eta^2 = .39$ (large effect), between the rubric (*M* = 14.62, *SD* = 2.16) and non-rubric (*M* = 11.68, *SD* = 1.45) condition.

Discussion and conclusion

In line with the need for more (quasi) experimental studies on PA (see Strijbos & Sluijsmans, 2010), a quasi-experimental study was conducted to investigate the impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance. We will first present a summary and interpretation of the results organised around the research questions.

Effect of a rubric on PA reliability and construct validity

Correlation of the expert and peer ratings revealed that peer ratings were reliable (consistency) and construct valid from the teacher perspective (TPv) reflected by their

overall moderate correlation with the experts' rating, as well as the moderate correlation between peer- and expert ratings within each research condition. Although lower than the .60 to .80 range reported by Falchikov and Goldfinch (2000) and Cho et al. (2006), these correlations also contain a degree of systematic bias. Nevertheless, the observed correlation warrants closer scrutiny of expert and peer ratings. Construct validity from the student perspective (SPv) revealed that the students in the rubric condition were also more valid (i.e., a smaller deviation from the expert as reflected by less over-scoring) in comparison to students in the non-rubric condition (hypothesis 1 was confirmed). This is in line with previous research on rubrics which showed that rubrics increase peer assessment construct validity (e.g., Jonsson & Svingby, 2007; Sadler & Good, 2006). It is also noteworthy that the deviation between students' rating with a rubric and the expert rating was small ($M = 0.84$), i.e. within one point on a scale of 0 to 20 points. Hence, it can be concluded that the use of a rubric has a strong potential to increase construct validity of PA. These results emphasise the importance of rubrics as a scaffold for implementing peer assessment in the classroom (Jonsson & Svingby, 2007), because rubrics contain the assessment criteria and facilitate more reliable and valid PA (Andrade, 2010; Panadero & Jonsson, 2013).

Effect of a rubric and friendship level on PA construct validity

No main effect was observed for friendship levels on construct validity from the student perspective (SPv). The 'rubric × friendship' interaction was not significant, however, the descriptives for SPv by friendship level hinted at differences within the rubric and non-rubric conditions. Univariate analyses revealed no significant difference between friendship levels for the non-rubric condition, however, in the rubric condition the students with a high level of friendship significantly over-scored the medium and low friendship levels. Thus, rubrics do not appear to reduce friendship bias (hypothesis

2a was rejected) and only students with low and medium friendship in the rubric condition appear to be more valid than students with a high level of friendship (hypothesis 2b was partially confirmed). The use of a rubric leads to more valid assessment on the one hand, but also seems to amplify – or make more visible – potential friendship bias. Finally, irrespective of friendship level all students over-scored their peers' concept map. These results are supported by previous research on friendship marking (i.e., students over-scoring friends) (Cheng & Warren, 1997; Pong & Ul-Haq, 1997), although it should be noted that in previous studies no explicit differentiation was made between friendship levels as was the case in the present study. It seems that students in the rubric condition with low and medium levels of friendship are less "afraid" to provide a realistic rating because they have the rubric to justify their decision, whereas students with a high level of friendship did not want to "confront" a close friend with their real performance or overlooked mistakes as they were more lenient towards a friend, however, these assertions should be examined in future studies – for example, by "assessing the assessor" (Kali & Ronen, 2008).

In sum, the findings are promising as friendship has not been systematically investigated within the area of peer assessment, however, it also highlights the need for a follow-up study containing a more even distribution of friendship levels because the operationalization of friendship following Hays (1984) (three levels derived from one item) resulted in relatively low variance. A more fine-grained operationalization of friendship – treating friendship as a multi-layered and compound construct – could be more accurate. A multi-layer operationalization of friendship might consider (a) whether a student collaborates for their school work with specific students, (b) whether they meet with other students outside of class for social activities, (c) whether they were

already acquainted before enrolling a programme, and (d) what their relationship is to specific class mates (inside and outside of class).

Effect of a rubric on students' perceived comfort and fairness during PA

The results revealed no difference with respect to perceived comfort and fairness between the rubric and non-rubric condition (Hypotheses 3a and 3b were rejected). Nevertheless, perceived comfort and fairness are correlated in the non-rubric condition (small effect), and in the rubric condition perceived fairness is correlated with peer rating (small effect) and perceived comfort with student perspective of validity (moderate effect). Apparently a higher degree of perceived fairness is related to a higher peer rating of the concept map, and a higher degree of perceived comfort is related to a lower degree of under-estimation and a higher degree of over-estimation in the rubric condition. Nevertheless, future research could adopt longer interventions to test whether perceived comfort and fairness increase over time and enhance a positive view of peer assessment.

Effect of a rubric on performance (i.e., quality of a concept map)

Students with a rubric clearly outperformed students without a rubric regarding the quality of their concept maps – irrespective of whether peer scores or expert scores are considered. Rubrics not only appear to facilitate a more valid peer assessment, but the rubric and specifically its clear specification of assessment criteria also appears to act as feed-forward information for students by specifying clear goals to be attained.

Limitations

A quasi-experimental study is challenging in a naturalistic context, in which full equivalence of research conditions is very hard to guarantee. The random assignment of condition was achieved at the class-level to prevent any confusion for the teacher and students, as well as unintended mixing of conditions. Checks on all dependent variables

revealed that the classes were sufficiently equivalent – despite some limited, but non-systematic differences. Future studies might separate the students more systematically within classes according to research condition, which might increase internal validity, but possibly at the expense of ecological validity.

Practical implications

There are two important implications for practice. First of all, when teachers want to increase the reliability and construct validity of peer assessment, rubrics should be provided to the students. Secondly, since friendships are a persistent aspect of classrooms it is important to consider the role of friendship between the assessee and assessor and its effects on peer assessment reliability and construct validity. Rubrics appear to have the potential to enhance construct validity from the student perspective for low and medium friendship students when the assessee is known, but they appear to amplify over-scoring due to a high level of friendship – the latter might be countered by assessing the assessor for the quality of his/her peer assessment. Although anonymity is often advocated as a solution, and with online voting and debating it revealed promising results (Ainsworth, Gelmini-Hornsby, Threapleton, Crook, O'Malley, & Buda, 2011), use of anonymity in combination with assessment is still an open issue. Whether anonymity counters the impact of friendship or decreases authenticity in terms of future work contexts (where assessment is usually not anonymous) needs to be determined. Nevertheless, rubrics have a positive effect on the reliability and construct validity of peer assessment through clear assessment criteria and a structured format and it might even reduce potential friendship bias, but it is up to future research to further elaborate on our findings.

References

Ainsworth, S., Gelmini-Hornsby, G., Threapleton, K., Crook, C., O'Malley, C., & Buda, M. (2011). Anonimity in classroom voting and debating. *Learning and Instruction, 21*, 365-378.

Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. J. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90-105). New York: Routledge.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19.

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

Barron, B. (2003). When smart groups fail. *Journal of the Learning Sciences, 12*(3), 307-359. doi: 10.1207/s15327809jls1203_1

Berry, J. W., & Chew, S. L. (2008). Improving learning through interventions of student-generated questions and concept maps. *Teaching of Psychology, 35*(4), 305-312. doi: 10.1080/00986280802373841

Besterfield-Sacre, M., Gerchak, J., Lyons, M., Shuman, L. J., & Wolfe, H. (2004). Scoring concept maps: An integrated rubric for assessing engineering education. *Journal of Engineering Education, 93*(2), 105-115.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation In Higher Education, 24*(4), 413-426. doi: 10.1080/0260293990240405

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105-121. doi: 10.1016/j.asw.2004.07.001

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, *22*(2), 233-239. doi: 10.1080/03075079712331381064

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891-901. doi: 10.1037/0022-0663.98.4.891

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Diaz Barriga, F. (2002). Estrategias docentes para un aprendizaje significativo: Una interpretación constructivista [Teachers' strategies for a meaningful learning: A constructivist interpretation]. Mexico, DF: McGraw-Hill.

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.

Edmonson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*, 350-385.

Falchikov, N. (2003). Involving students in assessment. *Psychology Learning and Teaching*, *3*(2), 102-108.

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

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.

Hafner, O. C., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, *25*(12), 1509-1528. doi: 10.1080/0950069022000038268

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53-70. doi: 10.1080/07294360123776

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: http://dx.doi.org/10.1016/j.tate.2013.07.008

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77-89.

Hays, R. B. (1984). The development and maintenance of friendship. *Journal of Social and Personal Relationships*, *1*(1), 75-98. doi: 10.1177/0265407584011005

Hilbert, T. S., & Renkl, A. (2009). Learning how to use a computer-based concept-mapping tool: Self-explaining examples helps. *Computers in Human Behavior*, *25*(2), 267-274. doi: 10.1016/j.chb.2008.12.006

Jacobs-Lawson, J. M., & Hershey, D. A. (2002). Concept maps as an assessment tool in psychology courses. *Teaching of Psychology*, *29*(1), 25-29. doi: 10.1207/s15328023top2901_06

Jehn, K. A., & Shah, P. P. (1997). Interpersonal relationships and task performance: An examination of mediation processes in friendship and acquaintance groups.

Journal of Personality and Social Psychology, *72*(4), 775-790. doi: 10.1037/0022-3514.72.4.775

Jones, A., & Issroff, K. (2005). Learning technologies: Affective and social issues in computer-supported collaborative learning. *Computers & Education*, *44*(4), 395-408. doi: http://dx.doi.org/10.1016/j.compedu.2004.04.004

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130-144.

Kali, Y., Ronen, M. (2008). Assessing the assessors: Added value in web-based multi-cycle peer assessment in higher education. *Research and Practice in Technology Enhanced Learning*, *3*, 3-32.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Karaca, E. (2009). An evaluation of teacher trainees' opinions of the peer assessment in terms of some variables. *World Applied Sciences Journal*, *6*(1), 123-128.

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

MacDonald, R. A. R., & Miell, D. (2000). Creativity and music education: The impact of social variables. *International Journal of Music Education*, *os-36*(1), 58-68. doi: 10.1177/025576140003600107

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, *26*(1), 53-63. doi: 10.1080/03075070020030715

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry in score meaning. *American Psychologist, 50*(9), 741-749.

Moni, R. W., Beswick, E., & Moni, K. B. (2005). Using student feedback to construct an assessment rubric for a concept map in physiology. *Advances in Physiology Education, 29*, 197-203.

Moni, R. W., & Moni, K. B. (2008). Student perceptions and use of an assessment rubric for a group concept map in physiology. *Advances in Physiology Education, 32*, 47-54.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*(3), 413-448. doi: 10.3102/00346543076003413

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

O'Brien, C. E., Franks, A. M., & Stowe, C. D. (2008). Multiple rubric-based assessments of student case presentations. *American Journal of Pharmaceutical Education, 72*(3), Article 58.

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), 806-813. doi: 10.1016/j.lindif.2012.04.007

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2014). Rubrics vs. self-assessment scripts: Effects on first year university students' self-regulation and performance. *Infancia y Aprendizaje, 37*(1).

Panadero, E., Alonso-Tapia, J., & Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies in Educational Evaluation*. doi: 10.1016/j.stueduc.2013.04.001

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9,* 129-144. doi: http://dx.doi.org/10.1016/j.edurev.2013.01.002

Papinczak, T., Young, L., & Groves, M. (2007). Peer assessment in problem-based learning: A qualitative study. *Advances in Health Sciences Education, 12*(2), 169-186. doi: 10.1007/s10459-005-5046-6

Pond, K., & Ul-Haq, R. (1997). Learning to assess students using peer review. *Studies In Educational Evaluation, 23*(4), 331-348.

Pope, N. K. L. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26*(3), 235-246. doi: 10.1080/02602930120052396

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*(1), 51-63. doi: 10.1080/0260293042003243896

Rose, S., & Serafica, F. C. (1986). Keeping and ending casual, close and best friendships. *Journal of Social and Personal Relationships, 3*(3), 275-288. doi: 10.1177/0265407586033002

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31.

Sambell, K., McDowell, L., & Brown, S. (1997). “But is it fair”? An explorative study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*(4), 349-371.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education, 49*(4), 413-430. doi: 10.1007/s10734-004-9448-9

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education and Teaching International, 39*(1), 71-81. doi: 10.1080/13558000110102904

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology, 36*(2), 102-107. doi: 10.1080/00986280902739776

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., Segers, M. S. R., & Tillema, H. H. (2009). Fostering interactivity through formative peer assessment in (web-based) collaborative learning environments. In C. Mourlas, N. Tsianos, & P. Germanakos (Eds.), *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 375-395). Hershey, PA: IGI Global.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265-269. doi: 10.1016/j.learninstruc.2009.08.002

Tabachnik, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Tolmie, A. K., Topping, K. J., Christie, D., Donaldson, C., Howe, C., Jessiman, E., Livingston, K., & Thurston, A. (2010). Social effects of collaborative learning in primary schools. *Learning and Instruction*, *20*(3), 177-191. doi: http://dx.doi.org/10.1016/j.learninstruc.2009.01.005

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

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht, the Netherlands: Springer.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, *86*(2), 264-286.

Tsai, C.-C., & Liang, J.-C. (2009). The development of science activities via on-line peer assessment: The role of scientific epistemological views. *Instructional Science*, *37*, 293-310.

Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role for of the peer feedback; A study of high school computer course. *Computers and Education*, *49*, 1161-1174.

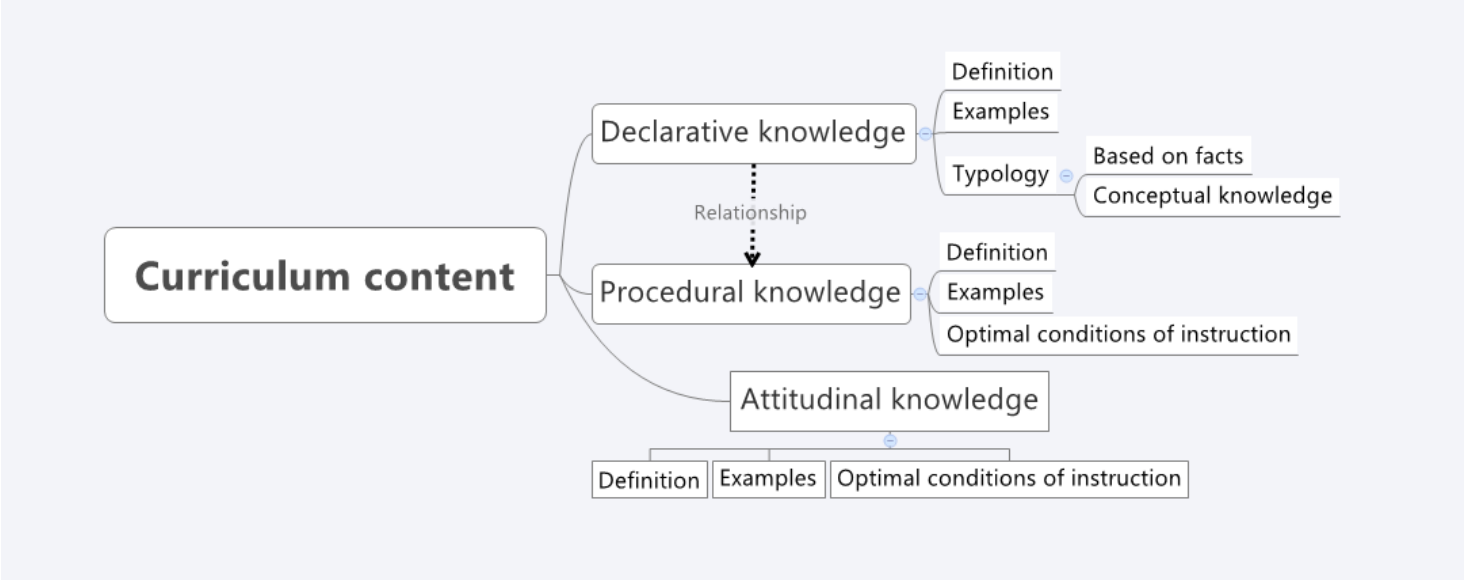
Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review*, *4*(1), 41-54. doi: 10.1016/j.edurev.2008.11.002

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270-279. doi: 10.1016/j.learninstruc.2009.08.004

Vass, E. (2002). Friendship and collaborative creative writing in the primary classroom. *Journal of Computer Assisted Learning*, *18*(1), 102-110. doi: 10.1046/j.0266-4909.2001.00216.x

Wang, J., & Imbrie, P. K. (2010, June). *"Students" peer evaluation calibration through the administration of vignettes*. Paper presented at the 2010 American Society for Engineering Education Annual Conference & Exposition, Louisville, KY, USA.

Figure 1. Illustration of a concept map by an expert (Note: the concept map has been simplified to visualise the task).



Appendix A: Rubric used for expert and peer assessment of concept maps

Criteria / **Score**	**4**	**3**	**2**	**1**
Concepts	All the important and secondary concepts are included.	Contains the important and some secondary concepts but not all.	The important concepts are included but not the secondary ones.	Some key concepts are lacking.
Hierarchy	The organization is complete and correct, and reflected by the map.	The organization is correct but incomplete: some levels or elements are lacking.	The organization is complete but incorrect: there are concepts in the wrong places.	The organization is incomplete and incorrect.
Relationships among concepts in different hierarchical levels	RELATIONSHIPS: They are correct making connections among the correct concepts. LINKS: They are explicit and help to better understand the relationships among concepts.	RELATIONSHIPS: They are correct but incomplete: some connections are lacking. LINKS: They are incomplete. Only some are explicit but they are correct.	RELATIONSHIPS: Some are incorrect making connections among concepts without relationship. LINKS: Only some are explicit, but some are incorrect.	RELATIONSHIPS: The majority are incorrect or there are only a few. LINKS: They are incomplete and incorrect.
Relationships among concepts from different columns	There are several connections making relevant relationships.	There is only one.	None	None
Simplicity and easiness of understanding	Its design is simple and easily understandable.	There are examples. Some relationships are difficult to understand.	Contains a few examples. There is an excessive number of connections.	There are no examples. Neither the relationships nor the hierarchy are understandable.