Studies in Educational Evaluation 68 (2021) 100980

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Structural validity and invariance of the Feedback Perceptions Questionnaire

Jan-Willem Strijbos [a,c,*], Ron Pat-El [b,c], Susanne Narciss [d]

[a] Department of Educational Sciences, University of Groningen, Grote Rozenstraat 3, 9712 TG, Groningen, the Netherlands
[b] Department of Statistics, Open University of the Netherlands, P.O. Box 2960, 6401 DL, Heerlen, the Netherlands
[c] Institute of Education and Child Studies, Leiden University, P.O. Box 9555, 2300 RB, Leiden, the Netherlands
[d] Psychology of Learning and Instruction, Technische Universitaet Dresden, Zellescher Weg 17, D-01062, Dresden, Germany

ARTICLE INFO

Keywords:
Feedback
Feedback perceptions
Confirmatory factor analysis
Structural validity
Measurement invariance
Structural equation modelling

ABSTRACT

Despite a growing interest in instructional feedback, students' feedback perceptions received limited attention. We examined the structural validity and measurement invariance of the Feedback Perceptions Questionnaire (FPQ). The FPQ measures feedback perceptions in terms of perceived fairness, usefulness, acceptance, willingness to improve, and affect. Secondary school students (N = 1486) received a fictional scenario containing Concise General Feedback or Elaborated Specific Feedback by a fictional peer. Students rated their perceptions as if they had received the feedback themselves. Confirmatory Factor Analysis (CFA) supports the structural validity of the FPQ and its invariance for the two types of peer feedback, gender, four grade-levels and two tracks. Perceived fairness of peer feedback was a strong positive predictor of willingness to improve and affect, whereas perceived usefulness and acceptance of peer feedback showed a more complex pattern in predicting willingness to improve and affect.

1. Introduction

Feedback is one of the most powerful instructional techniques (e.g., Evans, 2013; Hattie & Clarke, 2018; Hattie & Timperley, 2007; Jönsson, 2013; Narciss, 2008; Van der Kleij, Feskens, & Eggen, 2015; Winstone, Nash, Parker, & Rowntree, 2017). Feedback can be defined as all post-response information which informs the learner on their actual state of learning and/or performance, in order to help detect if their current state corresponds to the learning aims (Narciss, 2008). Feedback provided by an external source can be designed and delivered in many ways and unfolds its impact depending on a complex interplay of individual and situational factors of a given instructional context.

Thus, despite the considerable volume of research on feedback, many aspects of how feedback operates are still poorly understood. The historical review and meta-analysis by Kluger and DeNisi (1996) on external evaluative feedback of task or process aspects of performance revealed, for example, that one-third of external feedback interventions actually reduced performance. They proposed the Feedback Intervention Theory (FIT) and added the metacognitive, motivational and affective planes to explain (non)processing of feedback by the recipient. Although considered by Kluger and DeNisi (1996), the review by Butler and Winne (1995) synthesized feedback and self-regulation explicitly. They highlight the interplay between internal feedback (within oneself) and external feedback (by others), as a result of which the recipient may (a) ignore the external feedback, (b) reject the external feedback, (c) judge the external feedback irrelevant, (d) consider the external and internal feedback as unrelated, (e) re-interpret external feedback to make it conform to the internal feedback, and (f) make superficial rather than fundamental changes. Building on both reviews, (Narciss (2006, 2008, 2013, 2017) developed the Interactive Tutoring-Feedback (ITF) model. The ITF-model adopts a multidimensional view on feedback by differentiating the core internal and external factors and processes that influence how the content of feedback (e.g., evaluative and/or informative remarks) provided by an external source (e.g., teacher, peer, computer-based learning environment) is perceived, processed and used for further learning. It provides a heuristic framework for examining how the interplay of the internal and external factors contributes to the effects of interactive feedback strategies on learners' cognition, metacognition and motivation. Although slightly different in their assumptions and external feedback components compared to Narciss, Hattie and Timperley (2007), Shute (2008) and Yang and Carless (2013) also advocate a multidimensional view on feedback.

* Corresponding author at: Department of Educational Sciences, University of Groningen, Grote Rozenstraat 3, 9712 TG, Groningen, the Netherlands.
E-mail address: j.w.strijbos@rug.nl (J.-W. Strijbos).

https://doi.org/10.1016/j.stueduc.2021.100980
Received 7 November 2019; Received in revised form 29 November 2020; Accepted 9 January 2021
Available online 23 January 2021

In line with the increased attention for the interplay of internal and external feedback and resultant (non)processing of external feedback, recent research emphasizes the agency and active role of the recipient in processing feedback (Boud & Molloy, 2013; Price, Handley, & Millar, 2011; Strijbos & Müller, 2014; Winstone et al., 2017). In fact, the emphasis on recipience processes stresses Mory's (2004) urge for research into feedback perceptions as an explanatory, moderator and/or mediator of feedback processing. Feedback perceptions refer to the outcomes of how recipients spontaneously experience the feedback content as provided by an external source—or the feedback process as a whole—in terms of cognitive, metacognitive, motivational, and/or affective reactions. Feedback perceptions may vary depending on various feedback content (e.g., evaluative and/or informative remarks) and/or various feedback sources (e.g., teacher, peer, computer-based learning environment), and thus unfold their moderating or mediating impact in various ways. Investigating issues related to feedback perceptions and their role for feedback processing requires reliable and valid instruments to measure feedback perceptions across various feedback processes and contexts (cf. Brown & Harris, 2018). In this paper we will first outline how feedback content and source may influence feedback perceptions in the processing phase. Secondly, we revisit how prior work has measured feedback perceptions in the processing phase. Finally, we will describe the Feedback Perceptions Questionnaire (FPQ) and investigate the structural validity and invariance of the FPQ.

2. Impact of feedback content and source on feedback perceptions and processing

The external feedback content, characteristics of the source (e.g., Leung, Su, & Morris, 2001) or a combination of external feedback content and source characteristics (Berndt, Strijbos, & Fischer, 2018; Raemdonck & Strijbos, 2013; Strijbos, Narciss, & Dünnebier, 2010) can influence feedback perceptions in the processing phase. However, there might be differences in perceived credibility of various feedback sources (e.g., teacher vs. peer; supervisor vs. co-worker) due to a perceived implicit and/or explicit power differential (Boud & Molloy, 2013; Carless, 2006; Leung et al., 2001; Raemdonck & Strijbos, 2013; Steelman, Levy, & Snell, 2004; Strijbos & Müller, 2014; Strijbos et al., 2010; Winstone et al., 2017; Yang, Badger, & Yu, 2006). Most studies in education predominantly focus on the teacher as the source of feedback, which is not surprising given that the teacher is (historically) considered a subject area expert. However, the increased interest in peer feedback signifies that the 'peer' is also a potential and relevant feedback source in education. Nevertheless, as students are not experts in a subject area, peer feedback is susceptible to variation in content (Lockhart & Ng, 1995). Moreover, students often doubt their own and their peers' knowledge within a subject area, and their own and their peers' evaluation and feedback skills (McConlogue, 2012; Rotsaert, Panadero, Estrada, & Schellens, 2017; Strijbos, Ochoa, Sluijsmans, Segers, & Tillema, 2009; Van Gennip, Segers, & Tillema, 2010). Additionally, peer feedback can vary due to relationships whereby students prefer not to assess their peer too harshly (Cheng & Warren, 1997; Panadero, Romero, & Strijbos, 2013). As such, feedback from a peer might differentially affect the recipients' feedback perceptions in the processing phase—possibly even more so than teacher feedback.

3. Measurement of feedback perceptions in the processing phase

Issues of how feedback content and feedback source exert a combined or separate influence on feedback perceptions require reliable and valid measurement. Most existing studies focus on how a recipient perceives external feedback content and the literature reflects a mix of studies that focused on feedback perceptions in the processing phase as their main phenomenon (e.g., Gamlem & Smith, 2013; Poulos & Mahony, 2008; Strijbos et al., 2010), and studies that include perceptions of feedback as one factor amongst others when evaluating an instructional intervention or setting (e.g., Brown, Peterson, & Yao, 2016; Duijnhouwer, Prins, & Stokking, 2012; Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013). The existing studies differ in their research methods, i.e. qualitative (e.g., interviews and focus groups; see Gamlem & Smith, 2013; Poulos & Mahony, 2008; Pokorny & Pickford, 2010), quantitative (e.g., questionnaire; see De Kleijn et al., 2013; Gibbs & Simpson, 2003; King, Schrodt, & Weisel, 2009; Linderbaum & Levy, 2010) or a mixed method approach (see Carless, 2006; Lizzio & Wilson, 2008; Weaver, 2006). Most studies adopting a quantitative approach measured feedback perceptions as a single dimension like 'perceived usefulness' (Duijnhouwer et al., 2012; Kwok, 2008; Rakoczy et al., 2013; Rotsaert, Panadero, Schellens, & Raes, 2018), 'perceived fairness' (Leung et al., 2001; Nesbit & Burton, 2006), 'perceived timeliness' (Bayerlein, 2014), or 'perceived helpfulness' (Brooks, Huang, Hattie, Carroll, & Burton, 2019).

4. Existing questionnaires that measure multiple dimensions of feedback perceptions in the processing phase

In line with a multidimensional view on feedback it is evident that measurement of feedback perceptions should be multidimensional. Thus, we compiled an overview of existing instruments that measure multiple dimensions of feedback perceptions in the processing phase. Since feedback can be provided by various sources, we adopted a broad perspective for our review and included also questionnaires developed for a workplace context. Studies measuring only a single dimension (e.g., perceived usefulness, fairness, timeliness, helpfulness, etc.) were excluded. Given our quantitative questionnaire orientation we also required that (a) psychometric information on reliability and/or validity was provided for all (sub)scales, (b) (sub)scales consisted of multiple items, and (c) the phrasing of all items was available in the main text, a table or an appendix. Our literature search identified twelve questionnaires of which eight met these three criteria and are summarized in Table 1.

First of all, the questionnaires are more or less well based on a definition of feedback perceptions. Four out of eight were developed in an iterative and bottom-up process, and the aspects of feedback perceptions measured by the (sub)scales has to be derived from the item-content (AEQ, AFQ, IFOS, and SCoF). In contrast, the FES is based on an a priori model of feedback environment factors, and the FOS is rooted in the construct of feedback orientation. Only the FS was based on a brief, yet explicit, definition of feedback perceptions: "Feedback perceptions are thus concerned with how a learner perceives the feedback, which is assumed to be influenced by the feedback message, characteristics of the feedback provider and the frame of reference of the feedback receiver" (De Kleijn et al., 2013, p. 1014) and items were based on Hattie and Timperley (2007) and Shute (2008). Finally, the development of the FPQ was guided by the ITF-model (Narciss, 2008), yet the definition of feedback perceptions was only implicitly provided through item-content. Second, the eight questionnaires differ in their main focus: five measure perceptions of the overall feedback practice and/or environment (AEQ, FES, IFOS, FOS, and SCoF), whereas three measure perceptions of specific feedback on a specific task (AFQ, FPQ and FS). This is reflected in phrasing of items as evidenced by the sample item per (sub)scale in Table 1. The questionnaires adopting an overall orientation contain items measuring feedback perceptions across a range of tasks and situations (i.e., trait-like), whereas questionnaires with a specific orientation contain items measuring specific perceptions of feedback by a specific source in relation to a specific task in a specific situation (i.e., state-like). This key difference between trait-like and state-like items and scales is well-known in questionnaire research on emotions (Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011), a topic that is gaining ground in research on feedback (Goetz, Lipnevich, Krannich, & Gogol, 2018). The distinction between trait-like and state-like is important because how feedback is perceived in general is much less (or not all) explanatory for how specific feedback on a specific task by a specific source in a

Table 1
Overview of existing questionnaires and psychometric indicators for quality of measurement.

	Subscales	#	α		Sample item	PCA/PAF, EFA/CFA and invariance testing
Assessment Experience Questionnaire (AEQ; Gibbs & Simpson, 2003)	Quantity and timing of feedback	6	.87		I would learn more if I received more feedback	Sample: 731 university students. Varimax rotated PCA (loadings > .40) yielded two feedback subscales: 'Quantity' and 'quality' combined into a single subscale, and 'Use' as a separate subscale. No EFA/CFA. No invariance testing.
	Quality of feedback	6	.77		I don't understand some of the feedback	
	Use of feedback	6	.74		I use the feedback to go back over what I have done in the assignment	
Feedback Environment Scale (FES; Steelman et al., 2004)	*Supervisor (S), Coworkers (C)*		*S*	*C*		Sample: 405 employees. Maximum likelihood CFA yielded seven subscales (better fit than 1 or 2-factor model) for both sources. Supervisor: $\chi^2(56) = 429.20$, $\chi^2/df = 1.81$, CFI = . 97, SRMR = .04. Co-worker: $\chi^2(56) = 429.20$, $\chi^2/df = 2.53$, CFI = . 93, SRMR = .05 (see article for test-rest reliability and predictive and concurrent validity). No invariance testing.
	Source credibility	5	.88	.85	I have confidence in the feedback my supervisor [coworkers] gives me	
	Feedback quality	5	.92	.92	The feedback I receive from my supervisor [coworkers] helps me do my job	
	Feedback delivery	5	.86	.84	My supervisor [coworkers] generally provides feedback in a thoughtless manner	
	Feedback favorability	4	.88	.86	My supervisor [coworkers] generally lets me know when I do a good job at work	
	Feedback unfavorability	4	.85	.85	On those occasions when I make a mistake at work, my supervisor [coworkers] tells me	
	Source availability *(C = 4 items)*	5	.82	.74	My supervisor [coworkers] is too busy to give me feedback	
	Promotes feedback-seeking	4	.84	.85	I feel comfortable asking my supervisor [coworkers] for feedback about my work performance	
Assessment Feedback Questionnaire (AFQ; Lizzio & Wilson, 2008)	Developmental feedback	7	.85		Comments helped me focus on areas I could improve	Sample: 277 university students. Oblique rotated PAF (loadings > .40; cross-loading items deleted) yielded three subscales. Low to moderate correlations between subscales ranging .26 to .49. No EFA/CFA. No invariance testing.
	Encouraging feedback	4	.82		Acknowledged my good points or ideas	
	Fair feedback	4	.66		Gave feedback that I couldn't understand	
Instructional Feedback Orientation Scale (IFOS; King et al., 2009)	Feedback utility	10	.85		I pay careful attention to instructional feedback	Study 1: 212 undergraduate students. Varimax rotated PCA (loadings ≥ .60; cross loadings ≤ .40) yielded four subscales. Maximum likelihood CFA in Study 2 (245 undergraduate students) confirmed the revised (alpha's) four subscales: $\chi^2(318) = 594.79$, NNFI = 0.94, CFI = 0.95, RMSEA = 0.06. No invariance testing.
	Feedback sensitivity	9	.86		Corrective feedback hurts my feelings	
	Feedback confidentiality	5	.74		I do not like for others to hear what feedback I am receiving	
	Feedback retention	3	.69		I can't remember what teachers want me to do when they provide feedback	
Feedback Perceptions Questionnaire (FPQ; Strijbos et al., 2010)	Fairness	3	.80		I would consider this feedback justified	Sample: 89 graduate students. Oblimin rotated PCA (loadings > .40) yielded four subscales (of which the affect subscales were combined in a single scale). Oblimin PCA on a sample of 173 secretarial employees yielded the same four subscales (Raemdonck & Strijbos, 2013). No EFA/CFA. No invariance testing.
	Usefulness	3	.82		I would consider this feedback useful	
	Acceptance	3	.69		I would accept this feedback	
	Willingness to Improve	3	.82		I would be willing to improve my performance	
	Affect	6	.81		I would feel [...] if I received this feedback on my revision	
	Positive Affect	3	.90		… satisfied / confident / successful	
	Negative affect	3	.83		… offended / angry / frustrated	
Feedback Orientation Scale (FOS; Linderbaum & Levy, 2010)	*Study 1 (S1), Study 2 (S2)*		*S1*	*S2*		Study 1: 300 undergraduate students. Oblique rotated EFA (cross loadings ≤ .30) yielded four subscales with moderate correlations between subscales ranging .32–.50 (see article for test-rest reliability and criterion, convergent and divergent validity). Study 2 (267 employees): maximum likelihood CFA revealed adequate fit for a second-order factor model, $\chi^2(166) = 429.20$, SRMR = .08, RMSEA = .08, CFI = .89, TLI = .87. No invariance testing.
	FOS (all subscales combined)	20	.86	.91		
	Utility	5	.86	.88	Feedback is critical for improving performance	
	Social awareness	5	.80	.73	Feedback lets me know how I am perceived by others	
	Feedback self-efficacy	5	.77	.85	I believe that I have the ability to deal with feedback effectively	
	Accountability	5	.74	.78	It is my responsibility to apply feedback to improve my performance	

(*continued on next page*)

Table 1 (*continued*)

	Subscales	#	α	Sample item	PCA/PAF, EFA/CFA and invariance testing
Feedback Scale (FS; De Kleijn et al., 2013)	Focus			The feedback of my supervisor concerns […]	Sample: 1016 graduate students. Oblique rotated EFA (loadings > .40) yielded six subscales with low to moderate correlations between subscales ranging .19–.62. Cronbach's alphas were not reported. No CFA. No invariance testing.
	Focus on task	4		…The structure of my written work	
	Focus on self-regulation	2		… My work ethos	
	Elaboration			The feedback of my supervisor concerns […]	
	Positive elaboration	2		… Why something is good enough	
	Negative elaboration	2		… What I do not do well yet	
	Goal-relatedness			The feedback of my supervisor indicates […]	
	Feed up	3		… How a good study is executed	
	Feed back-forward	5		… What I have to do to reach my goals	
Student Conceptions of Feedback (SCoF; Brown et al., 2016)	Active use of feedback/Enjoyment	7	.92	I actively use feedback to help me improve	Sample: 300 university students. MAP^2 indicated a five or six factor solution and the 2-item 'Enjoyment' scale was modeled as a subfactor of 'Active use of feedback'. Maximum likelihood EFA with oblique minimization rotation had sufficient fit, $\chi^2 = 793.45$, $df = 313$; $\chi^2/df = 2.54$, p = .11; CFI = .91; Gamma Hat = .91; RMSEA = .065; SRMR = .070. No CFA on SCoF items only. No invariance testing.
	Enjoy receiving feedback	2	.78	I look forward to feedback from the markers	
	Peer feedback helps	6	.89	Feedback from my classmates helps my learning	
	Ignore feedback	5	.77	I ignore comments the markers make about my work	
	Tells if expectations are met	4	.62	I know I have done well if the result is better than last time	
	Tutor/marker comments are helpful	3	.74	Feedback from my markers makes it clear how to improve	

Note. # = items; α = Cronbach's alpha; PCA = Principal Component Analysis; PAF = Principal Axis Factoring; EFA = Exploratory Factor Analysis; CFA = Confirmatory Factor Analysis; SRMR = Standardized Root Mean-square Residual; RMSEA = Root Mean-Square Error of Approximation; CFI = Comparative Fit Index; TLI/NNFI = Tucker-Lewis/Non-Normed Fit Index; MAP^2 = Velicer's Minimum Average Partial test.

specific situation is perceived, and, thus, might lead to biased inferences. Third, the questionnaires differ in the degree to which multiple sources are either included in and/or considered while developing the instrument (FES, FPQ, and SCoF), and whether it can be used in multiple contexts (FPQ and FOS).

With respect to psychometric quality of an instrument it is important to determine whether the instrument is adequate (i.e., the usability in terms of reliability and validity to answer research questions and detect a desired effect size, e.g. small to medium in an educational context). Table 1 shows that for all questionnaires at least an exploratory analysis of the assumed subscales structure was reported (i.e., PCA, PAF or EFA; typically in combination with Cronbach's alpha). For three questionnaires structural validity (i.e., the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured; see Mokkink et al., 2010) was determined through confirmatory factor analysis (i.e., CFA). However, psychometric quality also depends on whether the instrument is robust (i.e., whether the instrument performs the same in various contexts and populations) and any numerical comparisons require that the instrument is invariant between them. When invariant, items are interpreted in a conceptually similar manner and will be answered in the same way by different subgroups in the sample. Since students typically receive different feedback depending on their work, it is important to determine that a questionnaire is at least invariant for different types of feedback. Furthermore, there is some evidence that points to gender differences in the processing of feedback (Dweck, Davidson, Nelson, & Enna, 1978; Roberts & Nolen-Hoeksema, 1989; Schmidt, 1995; Turner & Gibbs, 2010). Finally, in order to enable cross-sectional and longitudinal studies of feedback perceptions it is important to ensure that the questionnaire is invariant for grade level (a proxy for age) and different tracks (a proxy for cognitive ability; dependent on the educational system of a specific country). If achieved, invariance signifies that (a) group differences in factor means are unbiased and (b) group differences in observed means are directly related to group differences in factor means and not contaminated by differential response bias (Gregorich, 2006). If subgroups have statistically similar characteristics, then it can be concluded they have been drawn from the same population and thus comparison of scale mean scores can proceed (Wu, Li, & Zumbo, 2007). In sum, invariance can provide a strong(er) foundation for claims based on subgroup comparisons which is a prime focus in educational research. Yet, none of the eight existing questionnaires were tested for measurement invariance.

Out of these eight, we will focus specifically on the FPQ. First of all, the FPQ distinguishes two broad dimensions: (a) perceptions that relate to the cognitive function of feedback and the degree to which its contents are perceived in terms of perceived fairness, usefulness and acceptance, and (b) perceptions that relate to the motivational function of feedback and the degree to which its contents motivate the recipient to improve their performance while weighing affective reactions prompted by the feedback. Second, the FPQ measures feedback perceptions as a state-like phenomenon (i.e., specific perceptions of feedback on a specific task in a specific situation) whereas most instruments measure feedback perceptions as a trait-like phenomenon (i.e., general perceptions of feedback across a range of tasks and situations). Since most studies investigate student perceptions in relation to specific tasks in a specific situation, trait-like measures are problematic. Third, the FPQ has been used in various studies investigating perceptions of peer feedback (Berndt et al., 2018; Dijks, Brummer, & Kostons, 2018; Huisman, Saab, Van Driel, & Van den Broek, 2018; Peters, Körndle, & Narciss, 2018; Strijbos et al., 2010) teacher feedback (Agricola, Van der Schaaf, Prins, & Van Tartwijk, 2020), both peer and teacher feedback (Prins, De Kleijn, & Van Tartwijk, 2017), and feedback on study progress (Fonteyne et al., 2018).

5. Research aims

The present study reports on the structural validity and measurement

invariance of the Feedback Perceptions Questionnaire (FPQ) developed by Strijbos et al. (2010). In light of the increased attention for the peer as a feedback source and its impact on feedback perceptions, the study was conducted in the context of peer feedback. Finally, as questionnaires are likely used for subgroup comparisons, measurement invariance is crucial as violations of invariance may preclude meaningful interpretation of compared data. Hence, we investigated three research questions: (a) Can peer feedback perceptions be measured adequately and robustly with the FPQ?, (b) Are perceived fairness, usefulness and acceptance predictive of willingness to improve and affect?, and (c) Is the FPQ invariant for two types of peer feedback, gender, four grade levels and two tracks?

6. Method

6.1. Sample

The initial sample consisted of 1535 secondary education students in the Netherlands from 132 schools. There were 817 female and 713 male students (five students did not reveal their gender). Their mean age was 15.75 ($SD = 1.19$). The data was collected in classrooms from four grade levels (9–12) and two tracks (Senior general, Academic). In the Dutch educational system, 'track' represents students' cognitive ability, and the curriculum which is tailored to the track. In each school data was collected in three classrooms, covering all disciplines ranging from arts to sciences, in which four students per classroom were randomly selected by their teacher to complete the questionnaire. Student participation was based on informed consent. In all, 422 schools were approached of which 132 (31.28 %) agreed to participate, 222 (52.60 %) declined, and 68 (16.11 %) did not respond. The classroom response rate was 98.40 % and the participating schools were spread across different regions in the Netherlands to avoid a bias towards urban areas.

6.2. Materials

As part of the large scale questionnaire study, secondary school students were presented with a scenario in which a fictional student received feedback by a fictional peer. The scenario was embedded in the task of 'writing a formal (business) letter', which is part of the Dutch language curriculum and an authentic task for all students. In addition to the peer feedback the students received the evaluation criteria for a (business) letter (main criteria: components, content, spelling and style) and a fictional 'letter assignment'. Two feedback scenarios were designed in line with Narciss' (2008) feedback classification. The feedback was either Concise General (CGF) or Elaborated Specific (ESF), which are very common in classroom feedback and they constitute two extremes in terms of feedback content. CGF contained only general remarks regarding the performance, whereas ESF provided the position and error type, as well as information on how to proceed. To enhance comparability, ESF was constructed as an elaboration of CGF. Appendix A shows CGF and ESF, and error types according the classification by Narciss (2006, 2008).

6.3. Measures

We used the Strijbos et al. (2010) multi-dimensional 18-item Feedback Perceptions Questionnaire (FPQ). The FPQ measures feedback perceptions in terms of the perceived Fairness (FA, 3 items), Usefulness (US, 3 items), Acceptance (AC, 3 items), Willingness to Improve (WI, 3 items) and Affect (AF, 6 items – three items measuring positive affect and three measuring negative affect). Items were measured on a 10 cm visual analogue scale from 0 (fully disagree) to 10 (fully agree). Negatively phrased items were recoded. All items were phrased using English, German and Dutch adjectives, and by translation and re-translation we ensured that all items addressed the same semantic aspects regardless of language. Appendix B displays the subscales and items of the FPQ.

6.4. Procedure

Participating schools were informed prior to the study and information about the study was provided in a letter for the parents. Secondary schools were visited by research assistants who distributed the questionnaire in classrooms. The purpose of the study was explained beforehand and students received written instructions on how to answer the questionnaire. Students created an anonymous code and were informed that participation was voluntary and that they could stop whenever they wanted. When presented with the scenario students were asked to consider the peer feedback as if they had received it themselves, and indicate how they perceived the peer feedback in terms of fairness, usefulness, acceptance, willingness to improve, and affect. The reading of the scenario and answering the FPQ took about 10 min.

6.5. Data-analysis

The structural validation was conducted via Confirmatory Factor Analyses (CFA) using Structural Equation Modelling (SEM) to determine (a) the adequacy and robustness of the factor structure by Strijbos et al. (2010), (b) whether feedback perceptions in terms of fairness, usefulness and acceptance adequately predict willingness to improve and affect, and (c) whether the factor structure is invariant. CFA and SEM was performed in R version 3.6.1 with the lavaan package (Rosseel, 2012). Measurement Invariance was tested in EQS version 6.1. We used the unbiased 'wishart' estimator in R to make the output as similar to EQS as possible (Rosseel, 2012).

6.5.1. Confirmatory factor analysis

To interpret a model's fit, it is customary to use strict criteria for excellent fit. But, as Browne and Cudeck (1992) have demonstrated, the sample size places a soft upper-bound on potential fit. We use the typical indicators of excellent fit (e.g., Byrne, 1998), and use the sample-size constrained soft-upper bound as an indicator of adequate fit. As such the following indicators were used: Standardized Root Mean-square Residual (SRMR) and Root Mean Square Error of Approximation (RMSEA) below 0.10 is considered adequate fit and below 0.05 an excellent fit, and Comparative Fit Index (CFI) scores above 0.90 indicate adequate fit and above 0.95 excellent fit. Moreover, we also used the Gamma Hat (Fan & Sivo, 2007) to determine model fit. Gamma Hat ($\widehat{\gamma}$) does not penalize small or simple models as RMSEA does and in the case of small models it is less sensitive to model misspecification compared to CFI (Fan & Sivo, 200). Since the $\chi 2$ statistic becomes increasingly unreliable in large sample sizes > 250, the $\chi 2$ statistic will not be used as a criterion for model fit (Byrne, 1998; Putnick & Bornstein, 2016).

Negatively worded items in scales can result in two-factor models, or unidimensional models with an underestimate of reliability due to statistical artifacts because of wording-effects. The fit and subsequent reliability of scales with negatively worded items can be improved by either correlating the errors of negatively or positively worded items, or by loading the negatively worded items on an (extra) negative wording-effect factor (Distefano & Motl, 2009). In this study we correlated the error terms of the two negatively worded items in the AC subscale and three positively worded items in the AF subscale.

6.5.2. Invariance tests

The FPQ was designed to compare populations that received different types of feedback (CGF or ESF), and it is to be expected that the FPQ will also be used to compare between gender, grade levels, and track. In order to ensure that such comparisons are valid, the invariance of the FPQ was tested for two types of peer feedback, gender, four grade levels and two tracks. Invariance relates to the requirement that quantitatively measured constructs have the same meaning across groups, and that group comparisons of sample estimates (e.g., means and variances) reflect true differences free from contamination from group-specific attributes that are unrelated to the constructs of interest. The CFA-

framework allows invariance to be tested in a sequence of increasingly strict invariance. The most basic form of invariance is dimensional (baseline) invariance, when the same number of factors can be found in all groups, regardless of the underlying configuration of items and factors. The most commonly tested form of invariance is configural invariance, in which it is tested whether the same factors are associated with the same items in all groups. Because factor-intercorrelations and factor-weights per item can differ across groups, configural invariance is insufficient to defend quantitative comparisons between groups (Gregorich, 2006). The next level of invariance is metric invariance (sometimes called weak-factorial invariance), which tests whether (common) factors have the same meanings across groups—expressed as equal factor loadings. The first level of invariance that is sufficient to defend quantitative group-comparisons is scalar invariance (sometimes called strong-factorial invariance), when factor-loadings and factor intercorrelations are equal across groups (meaning that common factors have invariant meaning across populations) and there is no differential response bias between populations (i.e., equal item and factor intercepts). Scalar invariance allows for meaningful comparisons between populations because (1) group differences in estimated factor means are unbiased and (2) group differences in observed means will be directly related to group differences in factor means and will not be contaminated by differential additive response bias. Finally, the most stringent form of invariance is strict factorial invariance (sometime called residual invariance), when item-residuals are equal across populations. This would allow researchers to not only meaningfully compare means between populations, but also their variance estimates (Gregorich, 2006). Whereas Cheung and Rensvold (2002) recommended ΔCFI to evaluate invariance as it outperformed $\Delta\chi^2$, we will also report ΔRMSEA and ΔSRMR in line with recent recommendations by Putnick and Bornstein (2016), and we will additionally add ΔGamma Hat. We will use the following critical values to assess invariance: -0.01 ΔCFI for metric and scalar invariance and .015 for strict invariance, 0.015 ΔRMSEA for metric and scalar invariance, 0.030 ΔSRMR for metric invariance and 0.015 ΔSRMR for scalar invariance (Putnick & Bornstein, 2016; Chen, 2007). Since there is not yet consensus of a critical value for change in Gamma Hat ($\hat{\gamma}$), we adopted the -0.01 used for ΔCFI also for ΔGamma Hat ($\hat{\gamma}$). Finally, it should be noted that—akin to using multiple fit-indices to determine model fit—none of the change indicators and the proposed critical value rules supreme (Putnick & Bornstein, 2016).

7. Results

7.1. Data-inspection

All variables were examined for data entry accuracy, missing values, and fit between their distributions. Variables had between 16–19 missing cases (between 1 % and 1.2 %). No variables had missing values over 5 %, and there was no pattern to the missing data (MCAR's $\chi^2(9) = 1.50$, $p = .997$). Missing values were replaced by EM-estimates (see Musil, Warner, Yobas, & Jones, 2002) based on all other variables in the dataset. No continuous variables deviated from the normal distribution. No univariate extreme cases ($z > |3.00|$) were found for US and AF. In the variables FA ($N = 2$), AC ($N = 5$) and WI ($N = 10$) extreme values ranged between -3.59 ≤ z ≤ 3.01. Forty-nine cases were identified via Mahalanobis distance as extreme multivariate outliers ($p < .001$) and about equally distributed for gender. These outliers were removed from subsequent analyses. The final sample consisted of 1486 students (796 female, 685 male, five missing; mean age = 15.74, $SD = 1.19$).

Table 2
Means, standard deviations and correlations between fairness, usefulness, acceptance, willingness to improve, and affect (p < .01).

	Mean	SD	FA	US	AC	WI	AF
Fairness (FA)	6.26	2.03	-				
Usefulness (US)	6.04	2.63	.79	-			
Acceptance (AC)	6.15	1.95	.70	.69	-		
Willingness to Improve (WI)	6.83	1.86	.52	.49	.46	-	
Affect (AF)	5.02	1.55	.41	.35	.33	.10	-

7.2. Descriptives

We computed scale means for the five subscales, which shows common variance to a considerable degree as reflected in moderate to high correlations (Table 2).

7.3. Confirmatory factor analysis

As a baseline test a single factor multilevel CFA with school as the clustering variable was run with all 18 items sharing one common factor. This yielded a very poor fit, $\chi^2(270) = 4431.013$, Gamma Hat = .763, CFI = 0.693, SRMR_within = .115, SRMR_between = .636, RMSEA = .102 [90 % CI = .099; 0.105]. A multilevel CFA on all 18 items with correlated latent factors and with school as the clustering variable yielded a poor fit, $\chi^2(250) = 2834.703$, Gamma Hat = 0.869, CFI = .809, SRMR_within = .101, SRMR_between = .447, RMSEA = .083[90 % CI = 0.081; 0.086]. Inspection of Lagrange-Multipliers suggested that item FA1 was more indicative of the US and the AC subscale, but loading FA1 on either US and/or AC modified RMSEA only trivially. Lagrange-Multipliers suggested a wording effect in the AC and AF subscales due to both negatively and positively worded items. Correlating errors for the positively worded AF items and negatively worded AC items yielded an adequate fit, $\chi^2(242) = 1736.497$, Gamma Hat = 0.899, CFI = 0.890, SRMR_within = .076, SRMR_between = .450, RMSEA = .064 [90 % CI = 0.062; 0.067].

Finally, to test the proposed theoretical relationship between FA, US, AC, WI and AF, a multilevel structural equation model was tested. Modelling FA, US and AC as correlated predictors of both WI and AF—with school as the cluster variable—yielded an adequate fit, $\chi^2(242) = 1755.963$, Gamma Hat = 0.898, CFI = 0.888, SRMR_within = .075, SRMR_between = .369, RMSEA = .065 [90 % CI = 0.062; 0.068]. Fig. 1 shows the regression estimates. Appendix B displays the subscales, items, and Cronbach's alpha as an internal consistency indicator per subscale. Due to the cross-sectional nature of the data, the causal model was reversed as a sensitivity analysis. The reversed model showed adequate fit—except for the SRMR, $\chi^2(244) = 1872.866$, Gamma Hat = 0.891,CFI = 0.880, SRMR_within = .093, SRMR_between = .499, RMSEA = .067 [90 % CI = .064; 0.070]. Even though the fit indices of the reversed model indicate a slightly worse fit than the theoretical model, we inspected the AIC to compare these two non-nested models. The AIC of the theoretical model was 113662.543, whereas the reversed model AIC was 113775.445. Based on which model has the lowest AIC, the theoretical model was the preferred model. Following the rule of thumb outlined in Burnham and Anderson (2004), the difference between the two models (AIC_theory – AIC_reverse = 112.902) is larger

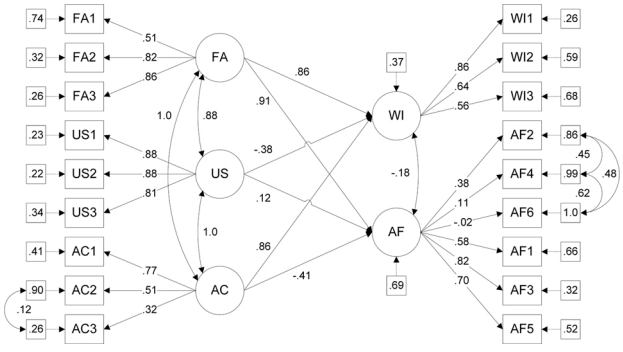

Fig. 1. Structural equation model of perceived fairness, usefulness and acceptance as predictors of willingness to improve and affect.

Table 3
Structural equivalence (baseline/configural) and invariance (metric/scalar/strict) model comparisons for Perceived Fairness (FA), Usefulness (US) and Acceptance (AC).

	Model	χ^2(df)	$\widehat{\gamma}$	CFI	SRMR	RMSEA (90 % CI)	Model comp.	Δχ2(Δdf)	$\Delta\widehat{\gamma}$	ΔCFI	ΔSRMR	ΔRMSEA	decision
Feedback (F)													
CGF ($N = 736$)	F1a	211.39 (22)	.944	.947	.048	.110 (.097; .123)	–	–	–	–	–	–	–
ESF ($N = 750$)	F1b	202.28 (22)	.949	.949	.036	.105 (.091; .118)	–	–	–	–	–	–	–
Configural	F2	413.68 (44)	.899	.948	.042	.107 (.098; .117)	–	–	–	–	–	–	–
Metric	F3	438.71 (51)	.895	.945	.054	.102 (.093; .111)	F2	25.03*** (7)	.004	.003	.012	.005	Accept
Scalar	F4	494.76 (54)	.893	.949	.061	.100 (.092; .109)	F3	56.05*** (3)	.002	.004	.007	.002	Accept
Strict factorial	F5	965.21 (68)	.831	.921	.119	.116 (.108; .123)	F4	47.45*** (14)	.062	.028	.058	.016	Reject
Gender (G)													
Male ($N = 796$)	G1a	17.14 (22)	.954	.960	.039	.099 (.085; .113)	–	–	–	–	–	–	–
Female ($N = 685$)	G1b	278.60 (22)	.933	.940	.043	.121 (.108; .134)	–	–	–	–	–	–	–
Configural	G2	448.74 (44)	.891	.949	.041	.112 (.102; .121)	–	–	–	–	–	–	–
Metric	G3	46.61 (51)	.891	.949	.047	.104 (.095; .113)	G2	11.87 (7)	<.001	<.001	.006	.008	Accept
Scalar	G4	477.04 (54)	.889	.949	.049	.102 (.093; .110)	G3	16.43*** (3)	.002	<.001	.002	.002	Accept
Strict factorial	G5	528.82 (68)	.878	.945	.053	.096 (.088; .104)	G4	51.78*** (14)	.011	.004	.004	.006	Accept
Grade level (L)													
Grade 9 ($N = 289$)	L1a	5.09 (22)	.979	.977	.032	.067 (.042; .091)	–	–	–	–	–	–	–
Grade 10 ($N = 533$)	L1b	189.66 (22)	.934	.943	.046	.120 (.104; .135)	–	–	–	–	–	–	–
Grade 11 ($N = 451$)	L1c	14.99 (22)	.944	.954	.038	.110 (.092; .127)	–	–	–	–	–	–	–
Grade 12 ($N = 211$)	L1d	101.56 (22)	.923	.934	.056	.131 (.106; .157)	–	–	–	–	–	–	–
Configural	L2	482.30 (88)	.809	.951	.044	.110 (.100; .120)	–	–	–	–	–	–	–
Metric	L3	518.90 (109)	.802	.949	.065	.101 (.092; .109)	L2	36.60* (21)	.007	.002	.021	.009	Accept
Scalar	L4	502.54 (92)	.805	.949	.052	.109 (.100; .118)	L3	16.36 (17)	.003	<.001	.013	.008	Accept
Strict factorial	L5	53.07 (106)	.791	.948	.051	.106 (.097; .115)	L4	27.53* (14)	.014	.001	.001	.003	Accept
Track (T)													
Senior general ($N = 605$)	T1a	144.72 (22)	.957	.958	.039	.096 (.081; .111)	–	–	–	–	–	–	–
Academic ($N = 876$)	T1b	281.07 (22)	.938	.949	.041	.116 (.104; .128)	–	–	–	–	–	–	–
Configural	T2	425.79 (44)	.898	.952	.040	.108 (.099; .118)	–	–	–	–	–	–	–
Metric	T3	428.79 (51)	.898	.953	.043	.100 (.091; .109)	T2	3.00 (7)	<.001	.001	.003	.008	Accept
Scalar	T4	432.54 (54)	.899	.953	.043	.097 (.088; .105)	T3	3.75 (3)	.001	<.001	<.001	.003	Accept
Strict factorial	T5	47.33 (68)	.891	.951	.045	.090 (.082; .098)	T4	37.79*** (14)	.008	.002	.002	.007	Accept

Note. * p < .05. ** p < .01. *** p < .001; $\widehat{\gamma}$ = Gamma Hat; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation; CI = confidence interval; Model comp. = Model comparison; Δ… = Change in fit index; decision = accept or reject measurement invariance.

Table 4
Structural equivalence (baseline/configural) and invariance (metric/scalar/strict) model comparisons for Willingness to Improve (WI) and Affect (AF).

	Model	χ^2(df)	$\widehat{\gamma}$	CFI	SRMR	RMSEA (90 % CI)	Model comp.	Δχ2(Δdf)	$\Delta\widehat{\gamma}$	ΔCFI	ΔSRMR	ΔRMSEA	decision
Feedback (F)													
CGF ($N = 736$)	F6a	126.80 (23)	.970	.944	.067	.078 (.065; .092)	–	–	–	–	–	–	–
ESF ($N = 750$)	F6b	157.50 (23)	.962	.928	.080	.088 (.075; .101)	–	–	–	–	–	–	–
Configural	F7	32.68 (53)	.925	.928	.079	.083 (.074; .091)	–	–	–	–	–	–	–
Metric	F8	321.35 (56)	.926	.928	.079	.080 (.071; .088)	F9	0.67 (3)	.001	<.001	<.001	.003	Accept
Scalar	F9	436.62 (63)	.916	.931	.082	.081 (.072; .089)	F8	115.27*** (7)	.010	.003	.003	.001	Accept
Strict factorial	F10	468.04 (72)	.911	.925	.083	.078 (.070; .086)	F9	31.42*** (9)	.005	.006	.001	.003	Accept
Gender (G)													
Male ($N = 796$)	G6a	142.31 (23)	.963	.928	.068	.087 (.073; .101)	–	–	–	–	–	–	–
Female ($N = 685$)	G6b	167.44 (23)	.961	.932	.078	.089 (.076; .102)	–	–	–	–	–	–	–
Configural	G7	311.84 (46)	.922	.925	.077	.091 (.081; .101)	–	–	–	–	–	–	–
Metric	G8	312.47 (56)	.926	.929	.076	.080 (.072; .089)	G7	0.63 (10)	.004	.004	.001	.011	Accept
Scalar	G9	361.94 (63)	.914	.927	.079	.082 (.073; .091)	G8	49.47*** (7)	.012	.002	.003	.002	Accept
Strict factorial	G10	42.67 (72)	.901	.914	.081	.083 (.075; .091)	G9	58.73*** (9)	.013	.013	.002	.001	Accept
Grade level (L)													
Grade 9 ($N = 289$)	L6a	73.19 (23)	.963	.916	.082	.087 (.065; .109)	–	–	–	–	–	–	–
Grade 10 ($N = 533$)	L6b	124.52 (23)	.959	.926	.078	.091 (.076; .107)	–	–	–	–	–	–	–
Grade 11 ($N = 451$)	L6c	87.45 (23)	.969	.946	.066	.079 (.062; .096)	–	–	–	–	–	–	–
Grade 12 ($N = 211$)	L6d	59.44 (23)	.963	.938	.078	.087 (.060; .114)	–	–	–	–	–	–	–
Configural	L7	345.00 (92)	.869	.933	.077	.086 (.076; .096)	–	–	–	–	–	–	–
Metric	L8	379.83 (122)	.865	.931	.082	.076 (.067; .084)	L7	35.219 (30)	.004	.002	.005	.010	Accept
Scalar	L9	367.42 (109)	.848	.931	.079	.086 (.077; .096)	L8	12.41 (13)	.017	<.001	.003	.010	Accept
Strict factorial	L10	379.16 (118)	.847	.930	.080	.083 (.074; .092)	L9	11.74 (9)	.001	.001	.001	.003	Accept
Track (T)													
Senior general ($N = 605$)	T6a	119.77 (23)	.966	.932	.073	.083 (.069; .098)	–	–	–	–	–	–	–
Academic ($N = 876$)	T6b	162.72 (23)	.966	.939	.070	.083 (.071; .095)	–	–	–	–	–	–	–
Configural	T7	282.49 (46)	.934	.937	.072	.083 (.074; .093)	–	–	–	–	–	–	–
Metric	T8	294.18 (56)	.933	.936	.075	.076 (.067; .084)	T7	11.69 (10)	.001	.001	.003	.007	Accept
Scalar	T9	315.37 (63)	.925	.936	.075	.076 (.067; .084)	T8	21.19** (7)	.008	<.001	<.001	<.001	Accept
Strict factorial	T10	327.10 (72)	.925	.936	.077	.071 (.063; .079)	T9	11.73 (9)	<.001	<.001	.002	.005	Accept

Note. * p < .05. ** p < .01. *** p < .001; $\widehat{\gamma}$ = Gamma Hat; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation; CI = confidence interval; Model comp. = Model comparison; Δ… = Change in fit index; decision = accept or reject measurement invariance.

than 10 which indicates essentially no support for the reversed model in comparison to the theoretical model.

7.4. Invariance tests

Tests for measurement invariance were conducted for FA + US + AC and WI + AF on (a) type of feedback, (b) gender, (c) grade level, and (d) track. No multilevel models that included all five latent factors converged. Removing school as a cluster variable still yielded no models that converged. For this reason measurement invariance was tested without a multilevel component for FA + US + AC and for WI + AF separately. Table 3 presents the findings for all tested models for FA + US + AC. Table 4 presents the findings for all tested models for WI + AF.

7.4.1. Invariance for type of peer feedback

Testing for the hypothesized baseline model for FA + US + AF yielded an excellent fit in both the CGF (model F1a) and ESF condition (model F1b). Incremental addition of constraints for equal factor loadings and covariances (model F3), and observed variable intercepts (model F4), indicated scalar invariance. Constraints on all estimated error terms did not fit satisfactory on all fit indices (Model F5). Testing for the hypothesized baseline model for WI + AF yielded an excellent fit in the CGF (model F6a) and ESF condition (model F6b). Incremental addition of constraints on factor loadings and covariances, observed variable intercepts and on all estimated error terms showed an adequate fit (Model F10), providing evidence for strict invariance on all fit indices.

7.4.2. Invariance for gender

Testing for the hypothesized baseline model for FA + US + AC yielded an excellent fit for male (model G1a) and female (model G1b) students. Incremental addition of constraints provided evidence for strict invariance (model G5). The only fit index rejecting strict factorial invariance was ΔGamma Hat (0.011) which was slightly above our cutoff of -0.01. All other indicators are far below the cutoff thresholds. Testing for the hypothesized baseline model for WI + AF yielded an excellent fit for male (model G6a) and female (model G6b) students. Incremental addition of constraints provided evidence for strict invariance (Model G10). The only fit index rejecting scalar and strict factorial invariance was ΔGamma Hat (-0.012 and -0.013 respectively) which were slightly above our cutoff of -0.01. All other indicators are far below the cutoff thresholds.

7.4.3. Invariance for grade level

Testing for the hypothesized baseline model for FA + US + AC yielded an excellent fit in Grade 9 (model L1a), Grade 10 (L1b), Grade 11 (L1c), and Grade 12 (L1d). Incremental addition of constraints provided evidence for strict invariance (L5). The only fit index rejecting strict factorial invariance was ΔGamma Hat (-0.014) which was above our cutoff of -0.01. All other indicators are far below the cutoff thresholds. Testing for the hypothesized baseline model for WI + AF yielded an excellent fit for Grade 9 (model L6a), Grade 10 (model L6b), Grade 11 (model L6c), and Grade 12 (model L6d). Incremental addition of constraints provided evidence for strict invariance (L10). The only fit index rejecting scalar factorial invariance was ΔGamma Hat (-0.017) which was well above our cutoff of -0.01. All other indicators are far below the cutoff thresholds, even the delta chi-square.

7.4.4. Invariance for track

Testing for the hypothesized baseline model for FA + US + AC yielded an excellent fit in both the Senior general (model T1a) and Academic track (model T1b). Incremental addition of constraints on factor loadings and covariances, observed variable intercepts and on all estimated error terms showed an adequate fit (Model T5), providing evidence for strict invariance on all fit indices, even all chi-square tests, save for strict invariance, $\chi^2(14) = 37.79$; $p < .001$. Testing for the hypothesized baseline model for WI + AF yielded an excellent fit in the Senior general (model T6a) and Academic track (model T6b). Incremental addition of constraints on factor loadings and covariances, observed variable intercepts and on all estimated error terms showed an adequate fit (Model T10), providing evidence for strict invariance on all fit indices, even all chi-square tests, save for scalar invariance, $\chi^2(7) = 21.9$; $p = .003$.

8. Discussion

In line with the increased interest in instructional feedback and the multidimensional view on feedback and feedback perceptions, we examined the psychometric quality of the Feedback Perceptions Questionnaire (FPQ; Strijbos et al., 2010). We focused specifically on the FPQ because it distinguishes two broad dimensions of perceptions that relate to the cognitive function of feedback (perceived fairness, usefulness and acceptance) and perceptions that relate to the motivational function of feedback while weighing affective reactions (willing to improve and affect), measures feedback perceptions as a state-like phenomenon, and has been used in various studies investigating peer and teacher feedback. More specifically, we investigated (a) the adequacy and robustness of the FPQ, (b) whether perceived fairness (FA), usefulness (US) and acceptance (AC) predict willingness to improve (WI) and affect (AF), and (c) whether the FPQ is invariant for two types of peer feedback, gender, four grade levels and two tracks.

Perceived fairness, usefulness, acceptance, willingness to improve and affect were found to be correlated—yet distinct—measures, comprised of two dimensions: the FA + US + AC part of the FPQ relates to the cognitive function, whereas as the WI + AF part of the FPQ relates to the motivational function. Perceived fairness, usefulness, and acceptance were confirmed as predictors of willingness to improve and affect, and was a far more likely model than the reverse. Perceived fairness is a strong positive predictor of willingness to improve and affect, whereas perceived usefulness and acceptance showed a more complex pattern. More specifically, a student perceiving the peer feedback as potentially useful will hold slightly more positive affect (0.12), but as it also signals that the performance was not yet good enough it reduces their willingness to improve (-0.38). Similarly, a student with an accepting perception of the peer feedback will be willing to improve their performance (0.86), however, doing so also implies acknowledging that the performance was not up to standard and leads to more negative affect (-0.41). The results clearly underscore the need and added value of a multidimensional view on feedback as well as feedback perceptions.

Furthermore, we found strict factorial invariance for gender, grade level and track for both the FA + US + AC and WI + AF part of the FPQ, and scalar invariance for type of peer feedback for FA + US + AC and WI + AF. Both the FA + US + AC and WI + AF parts are not only robust, but they are also similarly interpreted across peer feedback types, gender, grade levels and tracks. Moreover, the scalar level of invariance for types of peer feedback, gender, grade level and track implies for researchers in the social sciences that the comparisons of group means are meaningful. Scalar invariance is statistically important to meaningfully defend comparisons of factor and observed means. FA + US + AC was found to be strictly invariant for gender, grade level and track, which makes not only comparisons of group means defensible, but also comparisons of observed variances and covariances. WI + AF was strict invariant for types of peer feedback, gender, grade level and tracks. Although strict invariance is deemed desirable for quantitative comparisons, it is usually considered as excessively stringent (Byrne, 2006) and scalar invariance a more readily attainable goal (Gregorich, 2006), and in most cases adequate for comparative research. Moreover, to counter statistical artifacts due to wording-effects and an underestimate of reliability, we correlated the errors for two negatively worded items in the AC subscale and for three positively worded items in the AF scale (cf. Distefano & Motl, 2009). In sum, our results underline that students' peer feedback perceptions—in terms of FA, US, and AC, and WI and AF—can be adequately and robustly measured by the FPQ.

9. Limitations

It should be noted that the reported structural validity and invariance are strictly speaking limited to peer feedback and students in secondary education as the peer feedback recipient. However, as the items are not source specific, we consider it likely that a similar pattern in terms of estimates and invariance could be observed with the teacher as the feedback source or in other educational contexts; especially given the scalar invariance, and in 7 out of 8 cases even strict factorial invariance. Nevertheless, some of the factor loadings of individual items are somewhat lower than preferred and these signal areas for further improving the FPQ. Future research could examine the psychometric quality of the FPQ in the case of teacher feedback to (partially) replicate our findings. Likewise, the study could be repeated in higher education and, for example, compare responses by students from different disciplines.

We acknowledge that the use of scenarios might be considered artificial, however, scenarios have been shown to invoke almost identical reactions from persons, comparable to real situations (Robinson & Clore, 2001). In fact, a scenario's proximity to real-life settings and opportunity to (at least partially) control contextual variables, contributes to scenarios' high internal and external validity (Atzmüller and Steiner, 2010). Moreover, given the vast diversity in external feedback content, intrapersonal, interpersonal and situation factors, and associated diversity in feedback perceptions, scenarios including worked-out task exemplars with typical errors offer more control for investigating feedback perceptions, and feedback effects (Narciss, 2013). Finally, the present study examined only one scenario due to time constraints. Future research could include multiple scenarios (see e.g., Atzmüller and Steiner, 2010), for example, comparing peer and teacher feedback. Nevertheless, future research can reaffirm the FPQ's structural validity and measurement invariance when measuring feedback perceptions on a recipient's actual own performance; thus enhancing ecological validity (enhanced realism)—albeit to some extent at the expense of internal validity.

10. Implications for practice

Given the increased interest in (peer) feedback, students' feedback perceptions could be a crucial determinant of how they process the (peer) feedback and possibly help to uncover why elaborated (peer) feedback types are not always more efficient (see Berndt et al., 2018; Raemdonck & Strijbos, 2013; Strijbos et al., 2010). Moreover, due to the state-like measurement of feedback perceptions, the FPQ can inform teachers directly on how feedback by a specific peer on a specific task is perceived by a particular student or how feedback by a teacher on a particular student's task performance is perceived. Although the findings of the present study are limited to a setting with peer feedback, the established degree of invariance show promise for obtaining identical psychometric quality in a future study of the FPQ using teacher feedback. In sum, the FPQ offers researchers a structurally valid, reliable and invariant questionnaire to investigate relations between (peer) feedback perceptions, performance and feedback efficiency, and teachers a questionnaire to assess student responses to (peer) feedback and adjust instructional support accordingly.

Declaration of Competing Interest

None.

Appendix A. Concise General Feedback (CGF) and Elaborate Specific Feedback (ESF) components

Concise General Feedback (CGF)		Feedback component
Components	There are errors in technical components.	KR + KM (general)
Content	The content is sometimes too extensive.	KR + KM (general)
Spelling	There are spelling errors in the letter.	KR + KM (general)
Style	Sometimes there are style errors in the letter.	KR + KM (general)
Elaborated Specific Feedback (ESF)		**Feedback component**
Components	There are errors in technical components, like in the letter head.	KR + KM
	Also the address and date are not written correctly.	KR + KM
	The paragraphs are not neatly aligned, although this should be the case.	KR + KM + KH (implicit)
Content	Everything that should be in the letter is included, it is a bit extensive though.	KR + KM (general)
	Sometimes things are included that do not apply, such as the bank account number.	KR + KM + KH (implicit)
	Marieke does ask not for a financial refund, thus it is not necessary to write the bank account number.	KR + KM + KH
Spelling	There are spelling errors in the letter.	KR + KM (general)
	In the last sentence of the first paragraph for example, the word "voltooid" is written with a **t** and not with a **d**, even though the present perfect applies.	KR + KM + KH (implicit)
Style	Sometimes there are style errors in the letter.	KR + KM (general)
	In business letters, for example, you cannot use the "&" symbol.	KR + KM + KH (implicit)
	You can also not start a sentence with "And".	KR + KM + KH (implicit)
	It is better not to write "Again a disappointing telephone conversation".	KR + KM + KH

Note. KR (knowledge of result/ response), KM (knowledge of mistakes) and KH (knowledge on how to proceed). See (Narciss (2006, 2008) for more detail.

Appendix B. Feedback Perceptions Questionnaire (FPQ) subscales, items and Cronbach's alpha for the present study

Scale	α	Variable	Item
Fairness	.73	FA1	I would be satisfied with this feedback.
		FA2	I would consider this feedback fair.
		FA3	I would consider this feedback justified.
Usefulness	.90	US1	I would consider this feedback useful.
		US2	I would consider this feedback helpful.
		US3	This feedback would provide me a lot of support.
Acceptance	.56	AC1	I would accept this feedback.
		AC2	I would dispute this feedback.
		AC3	I would reject this feedback.
Willingness to Improve	.74	WI1	I would be willing to improve my performance.
		WI2	I would be willing to invest a lot of effort in my revision.
		WI3	I would be willing to work on further business letter assignments.
Affect	.69		I would feel … if I received this feedback on my revision.
Positive	.75	AF2	Satisfied
		AF4	Confident
		AF6	Successful
Negative	.75	AF1	Offended
		AF3	Angry
		AF5	Frustrated

Note. In addition to the passive phrasing suited for a scenarios study (i.e., "would"), actively phrased items (i.e., "will") for the measurement of feedback perceptions on one's own performance in a classroom setting can be obtained from the first author (it should be noted that these items are not yet validated).

References

Agricola, B. T., Van der Schaaf, M. F., Prins, F. J., & Van Tartwijk, J. (2020). Shifting patterns in co-regulation, feedback perception, and motivation during research supervision meeting. *Scandinavian Journal of Educational Research, 64*(7), 1030–1051. https://doi.org/10.1080/00313831.2019.1640283.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology, European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 128–138. https://doi.org/10.1027/1614-2241/a000014.

Bayerlein, L. (2014). Students' feedback preferences: How do students react to timely and automatically generated assessment feedback? *Assessment and Evaluation in Higher Education, 39*(8), 916–931. https://doi.org/10.1080/02602938.2013.870531.

Berndt, M., Strijbos, J. W., & Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *European Journal of Psychology of Education, 33*(1), 31–49. https://doi.org/10.1007/s10212-017-0343-z.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462.

Brooks, C., Huang, Y., Hattie, J., Carroll, A., & Burton, R. (2019). What is my next step? School students' perceptions of feedback. *Frontiers in Education: Assessment, Testing and Applied Measurement, 4*, 1–14. https://doi.org/10.3389/feduc.2019.00096.

Brown, G. T. L., & Harris, L. R. (2018). Methods in feedback research. In A. A. Lipnevich, & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 97–119). Cambridge, UK: Cambridge University Press.

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology, 86*(4), 606–629. https://doi.org/10.1111/bjep.12126.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245.

Byrne, B. M. (1998). *Structural equation modeling with lisrel, prelis, and simplis*. Mahwah, NJ: Erlbaum.

Byrne, B. M. (2006). *Structural equation modelling with EQS*. Mahwah, NJ: Erlbaum.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233. https://doi.org/10.1080/03075070600572132.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

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. https://doi.org/10.1080/03075079712331381064.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, (2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

De Kleijn, R. A. M., Mainhard, M. T., Meijer, P. C., Brekelmans, M., & Pilot, A. (2013). Master's thesis projects: Student perceptions of supervisor feedback. *Assessment and Evaluation in Higher Education, 38*(8), 1012–1026. https://doi.org/10.1080/02602938.2013.777690.

Dijks, M. A., Brummer, L., & Kostons, D. (2018). The anonymous reviewer: The relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assessment and Evaluation in Higher Education, 43*(8), 1258–1271. https://doi.org/10.1080/02602938.2018.1447645.

Distefano, C., & Motl, R. W. (2009). Methodological artifact or substance? Examinations of wording effects associated with negatively worded items. In T. Teo, & M. S. Khine (Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 59–77). Rotterdam: Sense.

Duijnhouwer, H., Prins, F. J., & Stokking, K. (2012). Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process and performance. *Learning and Instruction, 22*, 171–184. https://doi.org/10.1016/j.learninstruc.2011.10.003.

Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learning helplessness: II. The contingencies of evaluative feedback in the classroom and III. An experimental analysis. *Developmental Psychology, 14*(3), 268–276. https://doi.org/10.1037/0012-1649.14.3.268.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70–120. https://doi.org/10.3102/0034654312474350.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. https://doi.org/10.1080/00273170701382864.

Fonteyne, L., Eelbode, A., Lanszweert, I., Roels, E., Schelfhout, S., Duyck, W., et al. (2018). Career goal engagement following negative feedback: Influence of expectancy-value and perceived feedback accuracy. *International Journal for Educational and Vocational Guidance, 18*, 165–180. https://doi.org/10.1007/s10775-017-9353-2.

Gamlem, S. V., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education: Principles Policy and Practice, 20*(2), 150–169. https://doi.org/10.1080/0969594X.2012.749212.

Gibbs, G., & Simpson, C. (2003). Measuring the response of students to assessment: The assessment experience questionnaire. In *Paper Presented at the 11th Improving Student Learning Symposium*. Retrieved from https://pdfs.semanticscholar.org/5711/1701f9d713bde7780f6442e6bb4c1cf1b43b.pdf.

Goetz, T., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In A. A. Lipnevich, & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 554–574). Cambridge, UK: Cambridge University Press.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory analysis framework. *Medical Care, 44*(11 Suppl 3), 78–94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f.

Hattie, J., & Clarke, S. (2018). *Visible learning: Feedback*. New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.

Huisman, B., Saab, N., Van Driel, J., & Van den Broek, P. (2018). Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment and Evaluation in Higher Education, 43*(6), 955–968. https://doi.org/10.1080/02602938.2018.1424318.

Jönsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63–76. https://doi.org/10.1177/1469787412467125.

King, P. E., Schrodt, P., & Weisel, J. J. (2009). The instructional feedback orientation scale: Conceptualizing and validating a new measure for assessing perceptions of instructional feedback. *Communication Education, 58*(2), 235–261. https://doi.org/10.1080/03634520802515705.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, (2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254.

Kwok, L. (2008). Students' perception of peer evaluation and teachers' role in seminar discussions. *Electronic Journal of Foreign Language Teaching, 5*, 89–97.

Leung, K., Su, S., & Morris, M. W. (2001). When is criticism not constructive? The roles of fairness perceptions and dispositional attributions in employee acceptance of critical supervisory feedback. *Human Relations, 54*(9), 1155–1187. https://doi.org/10.1177/001872670154900.

Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the Feedback Orientation Scale (FOS). *Journal of Management, 36*(6), 1372–1405. https://doi.org/10.1177/0149206310373145.

Lizzio, A., & Wilson, K. (2008). Feedback on assessment: Students' perceptions of quality and effectiveness. *Assessment and Evaluation in Higher Education, 33*(3), 263–275. https://doi.org/10.1080/02602930701292548.

Lockhart, C., & Ng, P. (1995). Analyzing talk in ESL peer response groups: Stances, functions, and content. *Language Learning, 45*(4), 605–655. https://doi.org/10.1111/j.1467-1770.1995.tb00456.x.

McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. *Assessment and Evaluation in Higher Education, 37*(1), 113–123. https://doi.org/10.1080/02602938.2010.515010.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd ed., pp. 745–783). New York: Erlbaum.

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research, 24*(7), 815–829. https://doi.org/10.1177/019394502762477004.

Narciss, S. (2006). *Informatives Tutorielles Feedback [Informative tutorial feedback].* Münster, Germany: Waxmann.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. Van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–143). Mahwah, NJ: Erlbaum.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review, 23*, 7–26. Retrieved from http://revistes.ub.edu/index.php/der/article/view/11284.

Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 173–189). Singapore: Springer.

Nesbit, P. L., & Burton, S. (2006). Student justice perceptions following assignment feedback. *Assessment and Evaluation in Higher Education, 31*(6), 655–670. https://doi.org/10.1080/02602930600760868.

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. https://doi.org/10.1016/j.stueduc.2013.10.005.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002.

Peters, O., Körndle, H., & Narciss, S. (2018). Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education, 33*(1), 117–143. https://doi.org/10.1007/s10212-017-0344-y.

Pokorny, H., & Pickford, P. (2010). Complexity, cues and relationships: Student perceptions of feedback. *Active Learning in Higher Education, 11*(1), 21–30. https://doi.org/10.1177/1469787409355872.

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment and Evaluation in Higher Education, 33*(2), 143–154. https://doi.org/10.1080/02602930601127869.

Price, M., Handley, K., & Millar, J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education, 36*(8), 879–896. https://doi.org/10.1080/03075079.2010.483513.

Prins, F. J., De Kleijn, R., & Van Tartwijk, J. (2017). Students' use of a rubric for research theses. *Assessment and Evaluation in Higher Education, 42*(1), 128–150. https://doi.org/10.1080/02602938.2015.1085954.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

Raemdonck, I., & Strijbos, J. W. (2013). Feedback perceptions and attribution by secretarial employees: Effects of feedback-content and sender characteristics. *European Journal of Training and Development, 37*(1), 24–48. https://doi.org/10.1108/03090591311293275.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63–73. https://doi.org/10.1016/j.learninstruc.2013.03.002.

Roberts, T.-A., & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles, 21*(11/12), 725–747. https://doi.org/10.1007/BF00289805.

Robinson, M. D., & Clore, G. L. (2001). Simulation, scenarios, and emotional appraisal: Testing the convergence of real and imagined reactions to emotional stimuli. *Personality & Social Psychology Bulletin, 27*(11), 1520–1532. https://doi.org/10.1177/01461672012711012.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from: http://www.jstatsoft.org/v48/i02/.

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. https://doi.org/10.1016/j.stueduc.2017.02.003.

Rotsaert, T., Panadero, E., Schellens, T., & Raes, A. (2018). "Now you know what you're doing right and wrong!" peer feedback quality in synchronous peer assessment in secondary education. *European Journal of Psychology of Education, 33*(2), 255–275. https://doi.org/10.1007/s10212-017-0329-x.

Schmidt, C. F. (1995). Attributions of success, grade level, and gender as factors in choral students' perceptions of teacher feedback. *Journal of Research in Music Education, 43*(4), 313–329. https://doi.org/10.2307/3345730.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795.

Steelman, L. A., Levy, P. E., & Snell, A. F. (2004). The feedback environment scale: Construct definition, measurement, and validation. *Educational and Psychological Measurement, 64*(1), 165–184. https://doi.org/10.1177/0013164403258440.

Strijbos, J. W., & Müller, A. (2014). Personale Faktoren im Feedbackprozess. In H. Ditton, & A. Müller (Eds.), *Feedback und rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder [Feedback and evaluation: Theoretical foundations, empirical findings, practical implementation]* (pp. 87–134). Münster, Germany: Waxmann.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008.

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.

Turner, G., & Gibbs, G. (2010). Are assessment environments gendered? An analysis of the learning responses of male and female students to different assessment environments. *Assessment and Evaluation in Higher Education, 35*(6), 687–698. https://doi.org/10.1080/02602930902977723.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. (2015). Effects of feedback in a computer-based environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881.

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*, (4), 280–290. https://doi.org/10.1016/j.learninstruc.2009.08.010.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment and Evaluation in Higher Education, 31*(3), 379–394. https://doi.org/10.1080/02602930500353061.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), 1–26. Retrieved from https://pareonline.net/getvn.asp?v=12&n=3.

Yang, M., & Carless, D. (2013). The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education, 18*(3), 285–297. https://doi.org/10.1080/13562517.2012.719154.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200. https://doi.org/10.1016/j.jslw.2006.09.004.