Feedback opportunities of Comparative Judgement: An overview of possible features and acceptance at different user levels.

Roos Van Gasse, Anneleen Mortier, Maarten Goossens, Jan Vanhoof, Peter Van Petegem, Peter Vlerick & Sven De Maeyer

Abstract

Given the increasing criticism on common assessment practices (e.g. assessments using rubrics), the method of C*omparative Judgement* (CJ) in assessments is on the rise due to its opportunities for reliable and valid competence assessment. However, up to now the emphasis in digital tools making use of CJ has lied primarily on efficient algorithms for CJ rather than on providing valuable feedback. *Digital Platform for the Assessment of Competences* (D-PAC) investigates the opportunities and constraints of CJ-based feedback and aims to examine the potential of CJ-based feedback for learning. Reporting on design based research, this paper describes the features of D-PAC feedback available at different user levels: the user being assessed (assesse), the user assessing others (assessor) and the user who coordinates the assessment (*Performance Assessment Manager* (PAM)). Interviews conducted with different users in diverse organizations show that both the characteristics of D-PAC feedback and the acceptance at user level is promising for future use of D-PAC. Despite that further investigations are needed with regard to the contribution of D-PAC feedback for user learning, the characteristics and user acceptance of D-PAC feedback are promising to enlarge the summative scope of CJ to formative assessment and professionalization.

Introduction

In current diverse assessment practices, one assessor is responsible for grading assessees' performances one after another solely. This grading is generally done by means

adfa, p. 1, 2011.

of rubrics[1], which are often lists of several criteria to grade competences. Then, the scores of the criteria are combined and a final mark is provided. Although this analytic method is widely spread among practitioners, the method has several issues (Pollitt, 2012). First, rubrics divide the competence in several artificial dimensions, assuming strict boundaries between all dimensions. However, competences cannot be divided into several dimensions, since these dimensions can overlap. Moreover, the dimensions combined cannot contain the entire competence (Sadler, 2009). Second, non-competence related aspects can influence the judgment, such as mood and time of the day (Bloxham, 2009). Even when assessors are trained, such issues still affect the assessment (Crisp, 2007). Last, even though assessees are assumed to be assessed independently, they are still compared with each other during grading. Therefore, absolute judgment is an illusion (Laming, 2004).

These biases have recently resulted in an increased popularity of an alternative assessment practice; Comparative Judgment (CJ). In this method performance is assessed by comparing representations[2] of the same competence (Pollitt, 2004), since comparisons are easier than grading (Thurstone, 1927). The intent is that, in contrast to rubrics, various assessors compare representations and decide which of them is better with regard to a certain competence. By means of statistical modeling, multiple comparisons of multiple assessors result in a rank order in which representations are scaled relatively from worst to best performance (Bramley, 2007). Major strengths of this method are the potential to result in a high level of reliability since CJ depends on direct comparisons (Kimbell et al., 2009; Pollitt, 2012), ruling out assessors' personal standards, limiting non-task relevant influences, and feeling more natural for assessors compared to other methods (Laming, 2004).

Up to now, a small number of digital tools has been developed in which CJ is the central assessment method (e.g. NoMoreMarking, e-scape). However, the majority of these tools are focused on assessing rather than learning. The emphasis in these tools lies on the (refinement of the) statistical algorithm behind CJ in order to obtain a reliable rank order more efficiently. Currently, there is lack of digital tools using CJ as a method that serves users of the tool with valuable learning possibilities. Therefore, the project developing a *Digital Platform for the Assessment of Competences* (D-PAC) investigates the opportunities and constraints of CJ-based feedback and aims to examine the potential of CJ-based feedback for learning. The D-PAC platform provides feedback at three levels: assessee level, assessor level and *Performance Assessment Manager* (PAM) level. Assessee level is defined as the level of assessed users, for example students in schools. Here, the main goal of feedback is to enhance students' performance.

---

[1]In education, a rubric is a scoring aid to evaluate the quality of responses. Rubrics Rubrics usually contain evaluative criteria, quality definitions for those criteria at particular levels of achievement, and a scoring strategy (Popham, 199).

[2] a representation is a medium that reflects a certain competence, for example an essay for writing skills, a video for acting skills, a drawing for artistic skills,…

Assessor level is defined as the level of assessing users, for example teachers in schools. At this level, the aim of feedback is to inform assessors of their performance in assessing. The PAM level is the level on which the whole assessment is coordinated, for example the head teacher or school board. The coordination role of the PAM implies that PAMs are also in charge to provide feedback towards assessees and assessors. D-PAC is built in a way that it is possible to adjust every little detail of feedback. Thus, PAMs can decide feedback features that are necessary for assessees and assessors and adjust the feedback reports towards both groups.

In this paper, we aim to present the possible feedback features of D-PAC at assessee, assessor and PAM level. Since the feedback research in D-PAC is ongoing, we will discuss some preliminary findings about user experiences for each feedback feature. The following research questions will guide our investigation:

1. Which feedback features are possible using CJ assessment?
2. How do feedback features of D-PAC contribute to users' acceptance of feedback at all levels?

Theoretical framework

Feedback is considered as a powerful instrument to enhance performances (Hattie & Timperley, 2007). In feedback literature can be distinguished between feedback characteristics (i.e. its focus, content and presentation) and feedback acceptance (i.e. its relevance, validity and reliability and accessibility and user friendliness) (Hattie & Timperley, 2007). Both are important to initiate learning opportunities.

Feedback characteristics

According to Hattie and Timperley (2007), feedback can be focused at different layers of performance. At the first layer ("self"), feedback is provided to the user as a person and is not necessarily linked to the performance itself or task goals for future learning. Next, indications can be provided for the evaluation of the task (e.g. what went good and what went wrong). Furthermore, feedback on the process focuses on which steps were followed to complete the task. Lastly, feedback on the regulation is related to processes within the feedback user (e.g. self-assessment) (Hattie & Timperley, 2007).

Regardless of the focus of feedback, effective feedback should answer three questions: (1) where am I going? (2) how am I going? and (3) where to next? The first question ("where am I going?") relates to the task or performance goals. The second question ("how am I going?") should provide information on one's relative standing

towards a certain standard. The last question ("where to next?") provides opportunities for a greater chance of learning. Additionally, feedback should reflect a certain amount of time invested in the process (van der Hulst, 2014). Differences in feedback content can occur depending on the source of feedback (e.g. peers or teachers) (Nicol & MacFarlane-Dick, 2006).

A last feedback characteristic of great importance is the presentation of feedback. Feedback should be readable, neat, comprehensible, individualized and provided within an appropriate time span to users (Nicol, 2009; Shute, 2008; Xu, 2010). Digital feedback has the potential to present feedback in a useful way for users and as such to boost its quality (van den Berg et al., 2014).

Feedback acceptance

Feedback can be a powerful tool to initiate learning at different levels in organizations as well (Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2010a). However, the contribution of feedback to users' learning can only be reached if feedback is accepted (Anseel & Lievens, 2009; Rossi, Lipsey & Freeman, 2004). A prerequisite of feedback acceptance is that feedback is perceived as useful and fair (Schildkamp & Teddlie, 2008; Van Petegem & Vanhoof, 2007; Visscher, 2002). This indicates that it is crucial to concern certain characteristics in feedback design to enhance the probability of feedback acceptance and thus future feedback use and user learning (see Figure 1). In what is next, we will describe the need of feedback to be relevant, reliable and accessible for users.

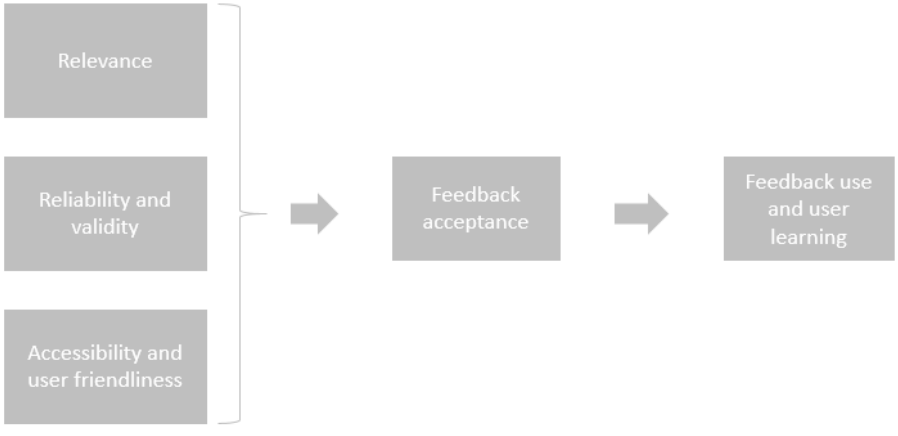

Figure 1. Model of feedback acceptance

Research has indicated that the perceived relevance of feedback is important for the way people will use it (Verhaeghe et al., 2010a). The importance of feedback relevance is reflected in the fact that people do not tend to use data which is not perceived as necessary in serving individual or organizational purposes (Vanhoof & Mahieu, 2011; Vanhoof et al., 2009).

Furthermore, the actual or perceived reliability and validity of the information concerned also influences whether or not feedback is used at different user levels (Kowalski & Lasley, 2009; Pierce & Chick, 2011; Schildkamp & Teddlie, 2008). A certain type of information is only used if users have the feeling that this information is correct, covers the right things and represents them accurately.

Lastly, literature has indicated that it is crucial that the feedback system is built in a user friendly and accessible way so that information becomes easily available (Schildkamp & Teddlie, 2008). Additionally, feedback should be ready for use (i.e. users should not have to further manipulate data). Good feedback systems are built in a way that reduce extra work load and enhance efficiency (Young, 2006).

Method

In order to investigate our present research questions, we started from a design research lens. This means that implementation in organizations was crucial for the (further) development of D-PAC feedback (Collins, Joseph, & Bielaczyc, 2004; Hevner, 2007).

In a first phase, the team[3] investigated which feedback features were possible with regard to CJ assessments. Then, with the team, several brainstorm sessions were held to decide on interesting features to include in D-PAC feedback. The time frame was set up for when these features needed to be implemented.

Next, the team discussed the features mentioned in the brainstorm sessions with a group of nine practitioners (education and HR context). These discussions were used to decide which features would be included in D-PAC feedback.

Subsequently, a draft feedback report was developed for the different feedback levels (assessee, assessor, PAM). The draft reports were discussed within the team and again presented to a group of practitioners. Remarks mentioned by the group of practitioners were taken into account for the further development of the feedback report.

---

[3] D-PAC is funded by the Flemish Agency of Innovation and Entrepreneurship and is a partnership between researchers of three Flemish universities (Universiteit Antwerpen, Universiteit Gent and Vrije Universiteit Brussel (iMinds))..

Then, several assessments via D-PAC were set up in various organizations in education and HR contexts (try-outs). Together with the organization in which the assessment was unrolled, decisions on which feedback would be provided were made for the three levels (assessee, assessor, PAM). After feedback was delivered, feedback users (see Table 1) were questioned (interview or focus group) on their understanding of the different features in the feedback report and their use of the feedback report. The individual and focus group interviews were transcribed ad verbatim. Next, summaries of the transcriptions were synthesized, compared and discussed by several members of the D-PAC team. Subsequently, conclusions were drawn about the acceptance of feedback at different user levels and decisions were made on the feedback features in future try-outs.

Thus, the try-outs were used to increase our understanding of how D-PAC feedback was accepted by its users. In the upcoming years, our findings of the try-outs will be used to further develop D-PAC feedback. Table 1 provides an overview of the try-outs in which feedback was provided. Different competences were assessed (e.g. writing, moodboards, …) in education and (education) selection domain. Assessees varied from high school students to candidate principals. Assessors and PAMs were mostly teachers (in training) or lecturers.

Table 1. Overview of try-outs in D-PAC in which feedback was provided.

	Competence	**Domain**	**Assessees**	**Assessors**	**PAM**	**Interviewed**
1	Argumentative writing	Education	136 high school students	68 teachers and teachers in training	10 head teachers	36 assessees and 6 PAMs
2	Writing formal letters	Education	12 bachelor students	11 teachers in training	1 lecturer	11 assessors
3	Visual representation	Education	11 bachelor students	13 teachers	1 lecturer	1 PAM
4	Analysing skills	Education	84 master students	4 professors	1 professor	1 PAM
5	Moodboards	Education	60 bachelor students	60 bachelor students 4 lecturers	1 lecturer	60 assessees 4 assessors 1 PAM
6	Self-reflective skills	Education	22 master students	9 lecturers	1 professor	9 assessors 1 PAM
7	Leadership skills	Selection	20 candidate school leaders	6 professionals (jury)	1 chair jury	6 assessors 1 PAM

Results

Table 2 provides information on the features that are available in D-PAC feedback Since PAMs manage the assessments, they can also decide which features are available for certain roles (assessee, assessor or PAM). For example, a PAM can decide to (not) insert a rank order in the feedback to assessees. Another aspect of PAM feedback is that this type of feedback is available during the assessment as well as at the end of the assessment. This way, PAMs can monitor aspects of the assessment in order to decide about the progress of the assessment (e.g. reliability). In this section, we will describe the different features of D-PAC feedback, following the structure of Table 2.

Table 2. Feedback features per level.

Feedback feature	Assessee	Assessor	PAM
General statistics			
Reliability	Y	Y	X
Mean comparisons per assessor	Y	Y	X
Mean comparisons per representation	Y	Y	X
Total time spent on assessment (overall)	Y	Y	X
Mean time per comparison (overall)	Y	Y	X
Total time spent on assessment (personal)	Y	Y	X
Misfit statistics			
Misfit assessor (personal)		Y	X
Misfit assessor (overall)			X
Misfit representations (overall)		Y	X
Rank order	Y	Y	X
Specific feedback	Y One assesse / all assessees	Y One assesse / all assessees	Overall

X= standard

Y= PAM can decide to switch the feature on or off

General statistics

Prior to the feedback features described below, users are provided with general information regarding the (finished) assessment. Therefore, an overview of the assessment is given in which the number of assessors, comparisons and representations of the assessment is included.

Reliability of the assessment

The advantage of CJ compared to rubric assessing is that the work of multiple assessors can be evaluated in terms of reliability (Kimbell et al., 2009; Pollitt, 2012). Therefore, the first feature of D-PAC feedback is a reliability measure for the assessment. Because the Rasch model is used to analyse the CJ data, the Rasch separation reliability or Scale Separation Reliability (SSR; Bramley, 2015) can be calculated. The measure represents the amount of spread in the results that is not due to measurement error (McMahon & Jones, 2014). According to Anshel, Kang, & Jubenville (2013) the SSR is an indication for how separable the representations are on the final scale of the assessment. So, a small measurement error implies that the relative position of the items on the scale is quite fixed (Andrich, 1982)

Like a Cronbach's Alpha value, the SSR is a value between 0 and 1. The higher the value, the higher the chance that the same rank order would be reached in a new assessment with the same set of assessors. Thresholds of SSR are similar to those of a Cronbach's Alpha value (i.e. 0.70 for a reasonable reliability and 0.80 for a good reliability).

Up to now, we have limited information about the value dedicated to the reliability in general. Current experiences with assessors and PAMs indicate that the reliability measure is evaluated in both groups. However, in these experiences reliability measures were each time sufficient (i.e. at least 0.70) so that no information is gained on implications of a failing reliability for how the assessment (and CJ) is evaluated in both groups. Thus, we do not (yet) have insight into how the reliability measure affects users' acceptance of feedback if the reliability is not sufficient.

First experiences with the reporting the reliability measure in assessee feedback showed mixed results. Some assessees found it interesting to know the reliability of the assessment, others did not care about it and were only interested in their own performance and the specific feedback they received.

Number of comparisons per assessor and per representation

In order to obtain a reliable assessment, choices need to be made about the number of comparisons that will be carried out by each assessor. Therefore, D-PAC feedback provides information on the number of comparisons in the assessment.

At assessee level, the number of comparisons of assessees' representation and the number of assessors who judged their representation can be included in feedback. Although some assessees perceived these measures as informative, others perceived this type of feedback as unnecessary.

At assessor level, the total number of comparisons, the mean number of comparisons per assessor, and one's own number of comparisons is given. Until now, it remains unclear whether users find this information relevant and whether or not this feature contributes to the acceptance of feedback among assessors.

For PAMs, two types of information are available on the number of comparisons. First, information on the number of comparisons is generated at assessor level. This means that the PAM is provided with insights into the number of comparisons made per assessor. Together with information on the time investment of each assessor, this type of information can help PAMs to evaluate assessors' workload. Second, information on the number of comparisons is generated at representation level. This means that PAMs are provided with information on how many times a representation is compared to another one. The number of comparisons per representation can influence the reliability of the assessment. Therefore, this statistic can be taken into account when the reliability is low. Up to now, little information is available on how the number of comparisons is used by PAMs. However, after one particular try-out, the PAM asked us to examine whether the reliability kept on increasing after a certain stage of the assessment. This information was important for the PAM in order to evaluate the assessment in terms of efficiency in order to improve the reliability costs balance in future use of CJ assessments.

Time investment

D-PAC provides indications of the time invested in the assessment. This comprises calculations of the duration of all comparisons. Several descriptive statistics on the time investment are generated, such as minimum and maximum duration of a comparison, and median and mean duration across comparisons.

At the level of the assessor, D-PAC provides an overview is provided of the invested time of an assessor within and across comparisons. These statistics are indicative for one's efficiency by comparing the mean time per assessor spend per comparison with one's own mean time spend per comparison. Also the total time spend on the assessment can be informative as the subjective idea of the total time is not always accurate. Up to now, assessors have pointed some of the time statistics as informative but not necessary.

At PAM level, an overview is generated of the total time investment in the assessment within and across assessors. The time investment statistic was included in PAM feedback so that PAMs have insight into the efficiency of the assessment. The statistic can have the potential to serve PAMs with information on *inter alia* what type of assignments are suitable for CJ, less efficient assessors in CJ or the efficiency of a CJ

assessment compared to other assessment methods used. Up to now, PAMs have indicated that the time investment statistic is interesting when running an assessment, but have not yet provided us with insights into how the statistic would be used during and after an assessment.

Misfit statistics

Comparative judgement assessments provide the opportunity to include misfit statistics in feedback. Misfit statistics are based on Rasch modelling (1960; 1980), which uses chi-squared goodness of fit statistics to quantify how far judgements are different to what the model predicts. Aggregating these estimates can produce a measure of how much judges deviate from the group consensus or how ambiguous a representation is for the assessor group. There are two types of misfit statistics: infit and outfit. In D-PAC the infit was included given it being less prone to occasional mistakes (Linacre & Wright, 1994). In short, the misfit statistic can be described as a quantification of an unexpected choice of an assessor in a specific comparison given the model expectation (Lesterhuis, Verhavert, Coertjens, Donche & De Maeyer, 2016). A misfit is identified if the infit estimate lies minimum two standard deviations above the mean (Pollitt, 2012) (see Figure 2).

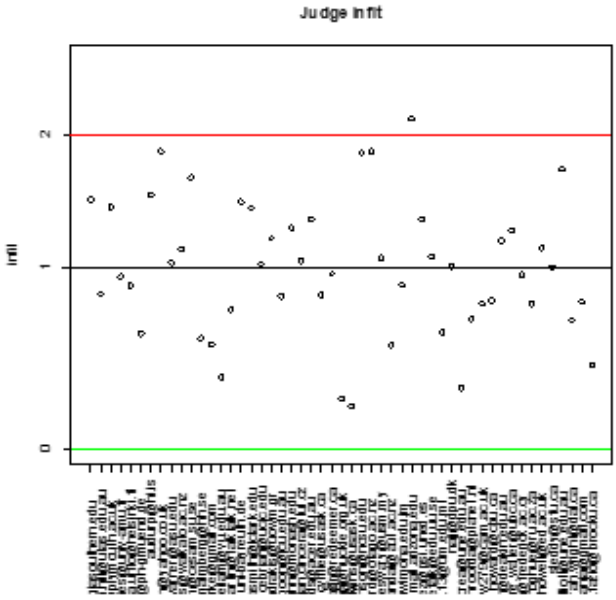

Figure 2. Judge infit.

A misfitting assessor implies that this assessor generally makes judgements that differ from the consensus. This could be explained by differences in conceptualization (Bramley, 2007) or finding certain aspects of a competence more important than other

assessors (Lesterhuis et al., 2016). Excluding misfitting assessors can lead to an increased reliability of the rank order (Pollitt, 2004).

A representation misfit implies that assessors make inconsistent judgements regarding this representation (Pollitt, 2004). Unusual content of representations regarding the performance can lead to representations being more difficult to judge (Bramley, 2007; Pollitt, 2004).

Misfit statistics are included in D-PAC feedback for assessors and PAMs, and can initiate valuable processes of professionalization. When a PAM notices a misfitting assessor, discussions between assessors on how certain competences are conceptualized can be organized in order to align this conceptualization among them. The same rationale goes for misfitting representations, which can be discussed in order to reach a consensus about these representations. This again can contribute to the assessment practice among assessors.

We do find that assessors and PAMs perceive misfit statistics as an interesting feature of D-PAC feedback. Assessors have indicated that they are often looking for data which tell them 'how they are doing' or 'if they judge in line with colleagues' during assessments. PAMs value misfit statistics because they provide them with a quality check.

Up to now, we have gained limited information about the use of misfit statistics by means of the try-outs. In a single try-out, a PAM asked to exclude a misfitting assessor in order to increase the reliability of the assessment. Currently, a try-out is ongoing in which misfit statistics are used for professionalization of assessors.

Rank order

In CJ assessments, a rank order is generated that represents the quality of representations in the light of the assessed competence (Bramley, 2007). The rank order is derived from the multiple comparisons assessors made. Contrary to common assessment methods, the rank order is relative instead of absolute. Based on these, the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) transforms the comparisons into the odds of winning a comparison with a random other representation (Pollitt, 2012). Next, these odds are translated into a rank order. Confidence intervals can be represented on the figure, to indicate the 95% confidence intervals of every performance. The rank order can be included at all feedback levels.

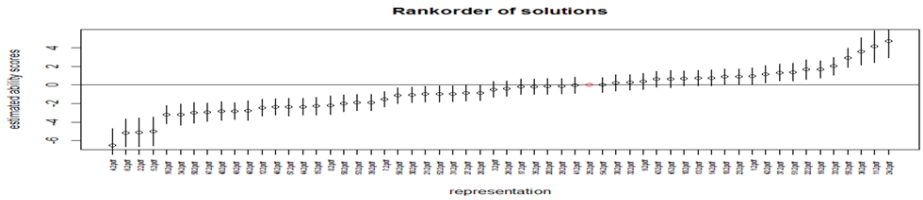

Figure 3. Rank order.

For assessees, providing the rank order gave mixed results. Some assessees perceived this measure as informative, since they could compare their own results with their peers' result. However, others stated that this is very relative, and it did not give an indication of their score or mark.

For assessors, the rank order is informative since it obtains the direct result of their work during the assessment. Assessors who already used D-PAC have generally indicated that comparing representations makes them curious about the rank order of the representations. They like to check whether the representations they assessed as good also get a good ranking. However, assessors have also mentioned that they would like to get immediate results of their judgements, for example by receiving their personal rank order.

For PAMs, the rank order is the central outcome of a CJ assessment. Depending on the assessment purposes (formative or summative), how PAMs deal with the rank order can differ. Up to now, rank orders generated by D-PAC have been used in various ways by PAMs. For formative purposes, the rank order for example has served as a guide to illustrate qualitative and less qualitative assignments. For example, in the moodboards try-out, a lecturer in higher education discussed aspects of moodboards that were higher and lower positioned on the rank order. For summative purposes, PAMs have for example used the rank order to score assignments or to select candidates in a selection procedure. For example, in this selection procedure, the first 14 candidates were selected for a training for future principles. In the try-out on debriefing notes, the first and the last representation on the rank order was discussed and scored among the assessors. Subsequently, they gradually graded each representation following the rank order.

Specific feedback

In D-PAC, there are two ways to provide specific feedback for assessees. First, it is possible that PAMs ask assessors to provide general feedback for each representation in a comparison (e.g. "*Task A is good on this competence, however it still needs more …")*. Second, PAMs can decide to include a feedback frame for assessors in which positive and negative aspects of each representation within a comparison are explicated (e.g. "*Task A is good because...; Task A is not good because…"*).

The aim of both feedback features is to provide assessees with insights into the strengths and weaknesses of their task. Because the specific feedback demands a greater time investment for assessors during the assessment, we investigated the necessity for specific feedback for assessees.

Our results indicate that D-PAC feedback is perceived more relevant, reliable and fair when specific feedback was included, independent of the type of feedback (i.e. receiving general feedback on the task or specified positive and negative aspects). Therefore, the specific feedback towards assessees strongly contributes to assessees' acceptance of D-PAC feedback. Additionally, assessees believe that specific feedback contributed to their learning[4]. Compared to the classic method of providing feedback (e.g. via comments in Word), specific feedback out of D-PAC is preferred when it comes to improving competences for future assignments. Providing both positive and negative feedback via D-PAC has been perceived as valuable, since in the classic feedback method (e.g. via comments in word), assessors are not inclined to insert positive comments. Content analysis confirmed assessees' perceptions, showing a significantly higher amount of motivational comments and learner-oriented feedback via D-PAC.

Discussion and conclusion

Comparative Judgement (CJ) has the potential to serve feedback opportunities that provide learning chances for different types of users. Nevertheless, digital tools for CJ assessment focus primarily on how to create a reliable rank order more efficiently and pay limited attention to the creation of CJ based feedback. The project *Developing a digital Platform for the Assessment of Competences* (D-PAC) addresses this shortcoming by investigating CJ based feedback opportunities. In the current paper, an overview has been given of feedback opportunities for users being judged (assessees), users who judge (assessors) and users who coordinate the assessment (Performance Assessment Managers: PAMs). Additionally, D-PAC users' acceptance of feedback has been described.

Considering how D-PAC feedback is constructed at different levels (i.e. assesse, assessor and PAM), the feedback can provide users at all levels with features which have the potential to affect user learning. Firstly, several embedded feedback characteristics may create valuable learning opportunities for users. At all user levels (i.e. assesse, assessor and PAM), feedback is focused at the task- and process level, which are important for user learning (Hattie & Timperley, 2007). At assessee level, the rank order can be used to gain insight in the quality of performance for a certain task compared to peers. Furthermore, assessees can be provided with insights of what was good and what needs to be improved with regard to the task (i.e. specific feedback). Strategies can be proposed on how to improve the current task or to complete a new, related task in a

[4] For more information, see Mortier et al., 2016.

better way. At assessor and PAM level, misfit statistics provide indications on the quality of the judgements made in the assessment (i.e. the task of assessors).

Secondly, content-wise all ingredients are available in D-PAC to provide users with answers on (1) "where am I going?, (2) "how am I going?" and (3) "where to next?" (Hattie & Timperley, 2007). Performance goals, process evaluation and improvement actions can be the result of a self-regulating process by means of D-PAC feedback. Assessees can for example use the rank order to look at good examples of peers to set objectives for future tasks and to improve their performances. At the level of assessors, the same rank order can be used to think about how the competences of assessees can be improved (in case the assessor is the actual tutor or lecturer of assessees) or one's misfit score can be a means of reflection (e.g. "What makes my judgements different to those of other assessors and what can I do to contribute to a better alignment in the judgement process?"). At PAM level, a trade-off between time investment and reliability can be made to set goals for future assessment in terms of assessment efficiency or misfit statistics can be used to introduce professionalization trajectories for assessors.

Lastly, several presentation issues are overcome. At assessee level, issues with readability (e.g. difficult handwriting) are overcome and feedback is individualized, which would increase assessees' engagement towards feedback (Xu, 2010). The overall feedback that needs to be provided in D-PAC results in higher-order feedback (Bouwer et al., 2016). Also at assessor level, feedback is individualized (e.g. one's own misfit score is available) in a comprehensible way. At all levels, feedback is provided within a reasonable time span. Although assessees need to wait for feedback until the assessment has ended, assessors are provided with feedback timely after they finished their comparisons and PAMs are able to request feedback during the assessment as well. A footnote to be made in this regard is that the PAM coordinates feedback for assessees and assessors. Thus, timely feedback depends on whether or not the PAM decides to provide it timely.

In short, feedback characteristics that are found to be crucial for user learning are embedded in D-PAC feedback.

Next to feedback characteristics, users' acceptance of feedback has implications for the feedback's potential for learning. General acceptance of D-PAC feedback is reached among the different users (assessees, assessors, PAMs), which is a first crucial step for future use of D-PAC feedback (Anseel & Lievens, 2009; Verhaeghe et al., 2010a; Schildkamp & Teddlie, 2008; Pierce & Chick, 2011). At all user levels, D-PAC feedback incorporates features that are perceived as relevant, reliable and valid. Moreover, the features are easily accessible and perceived as user friendly. At the level of assessees, in particular the specific feedback is perceived as valuable. The way feedback needs to be formulated in D-PAC (holistically and higher order) provides assessees with insights in how they need to develop their competences in order to better accomplish future assignments. At the level of the assessor and the PAM, the positive reactions on the misfit statistics are promising. For assessors, in most assessments, information on how they accomplish their assessment practice is rare. In D-PAC, misfit statistics provide assessors with data to evaluate 'how they are doing'. At PAM level, misfit statistics inherit an assessor quality check for PAMs. When a misfitting assessor

is identified by a PAM, attempts can be made to improve the assessing process and introduce discussions regarding how the judging task is approached among assessors.

Up to now, the D-PAC tool has developed promising feedback features for assessees, assessors and PAMs. However, limitations still remain in the opportunities for examining the impact of the different feedback features. Given that until now, try-outs were single occasions to use the D-PAC tool, feedback was given after the assessment at all three levels. Each time, the feedback was discussed afterwards and feedback acceptance was analyzed, but no follow up of feedback with a view to future assessments was possible in the organizations. Therefore, we were able to gain insight into the user satisfaction of feedback, but not in what users actually do with feedback and in how feedback drives learning. Thus, knowledge on how useful CJ feedback in general and D-PAC feedback in particular can be for users in individual and organizational learning processes is still lacking.

The current lack of knowledge on how D-PAC feedback is used and the contribution of D-PAC feedback to individual and organizational learning makes this a research area that should be addressed in future research. At the level of assessees, more research is needed in how the current feedback features, and in particular specific feedback, can initiate learning. It is necessary to gain insight into if and how this specific feedback is valuable for assessees in order to improve their competence under assessment. At the level of assessors, questions remain regarding the use of different feedback features, in particular misfit statistics, and its learning results. An interesting option would be to examine how assessors use misfit statistics and how the use of this data contributes to their personal development regarding the assessing task. At PAM level, knowledge on how PAMs use D-PAC feedback at different stages of the assessment would be useful. Therefore, it is needed that PAMs can be followed in assessments that are unrolled in D-PAC. This would provide information on how PAMs monitor and use the different feedback features. Also, future research should make attempts to address the contribution to organizational learning that PAM feedback can have. It would be valuable to obtain knowledge on how the different features of feedback can improve assessment strategies in organizations. Thus, several options remain in the further exploration of CJ based feedback in general and D-PAC feedback in particular.

The interesting options for feedback at assessee, assessor and PAM level reframe the use of CJ for assessments in schools and organizations. Whereas CJ has been seen as a valuable method to obtain a reliable method in an efficient way (Kimbell et al., 2009; Pollitt, 2012), insights into the potential feedback features can broaden this lens. Instead of a useful method for summative assessment, in which one needs to select the best candidates in a reliable way, D-PAC feedback can also be seen as convenient for formative assessment. The CJ method has a lot of potential with regard to feedback features that potentially initiate learning at different user levels in organizations. Despite the limited research opportunities of feedback via D-PAC so far, promising feedback features are available in D-PAC to enlarge the summative scope of CJ to formative assessment and professionalization.

Acknowledgement

This work was supported by Flanders Innovation & Entrepreneurship and the Research Foundation – Flanders (grant number 130043).

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9*(1), 95-104.

Anshel, M. H., Kang, M., & Jubenville, C. (2013). Sources of acute sport stress scale for sports officials: Rasch calibration. *Psychology of Sport and Exercise, 14*(3), 362-370.

Anseel, F., & Lievens, F. (2009). The Mediating Role of Feedback Acceptance in the Relationship between Feedback and Attitudinal and Performance Outcomes. *International Journal of Selection and Assessment, 17* (4), 362 376. DOI: 10.1111/j.1468-2389.2009.00479.x

Balzer, W.K., Doherty, M.E., & O'Conner, R.O. (1989). Effects of cognitive feedback in performance. *Psychological Bulletin, 106*, 410 – 433

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education, 34*(2), 209-220.

Bradley, R.A., & Terry, M.E. (1952). Rank analysis of incomplete block designs, 1. The method of paired comparisons. *Biometrika, 39*. DOI: 10.2307/2334029

Bramley, T. (2007). Paired comparisons methods. In Newton, P., Baird, J-A., Goldstein, H., Patrick, H., & Tymms, P. (Eds). *Techniques for monitoring the comparability of examination standards* (246 294). London: Qualification and Authority

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78*, 210–216.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design Research: Theoretical and Methodological Issues. *Journal of the Learning Sciences, 13*(1), 15-42.

Crisp, V. (2007). *Do assessors pay attention to appropriate features of student work when making assessment judgments?* Paper presented at the International Association for Educational Assessment, Annual Conference, Baku, Azerbaijan.

Fedor, D. B., Davis, W. D., Maslyn, J. M., & Mathieson, K. (2001). Performance improvement efforts in response to negative feedback: The roles of source power and recipient self-esteem. *Journal of Management, 27*(1), 79–97.

Hattie, J. and Timperley.H. (2007). The Power of feedback. *Review of Educational Research, 77*, 81-112.

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. Scandinavian *Journal of Information Systems, 19*(2), 87-92.

Hubbard, L., Datnow, A., & Pruyn, L. (2014). Multiple initiatives, multiple challenges: The promise and pitfalls of implementing data. *Studies in Educational Evaluation, 42*, 54-62.

Hubers, M. D., Poortman, C. L., Schildkamp, K., Pieters, J.M., & Handelzats, A. (2016). Opening the black box: knowledge creation in data teams. *Journal of Professional Capital and Community, 1*(1), 41-68.

Laming, D. (2004). *Human Judgment: The eye of the beholder*. London: Thomson Learning.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative judgement as a promising alternative. In Cano, E., Ion, G. (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (in press). Hershey, Pennsylvania: IGI Global.

Linacre, J., & Wright, B. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350.

Luce, R.D; (1959). *Individual choice behaviors. A theoretical analysis*. New York: J. Wiley.

McMahon, S., & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*(ahead-of-print), 1-22.

Michaud, R. (2016). The Nature of Teacher Learning in Collaborative Data Teams. *The Qualitative Report, 21*(3), 529-545.

Nicol, D. (2009). Good design of written feedback for students. In: *McKeachy. Teaching Tips: Strategies, research and theory for college and university teachers*. 13th edition, Houghton Miffin, New York. Pp. 108-124.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulating learning: A model of seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Pollitt, A., & Elliott, G. (2003). *Monitoring and Investigating comparability: a proper role for human judgement*. Paper presented at the Invited paper, QCA comparability seminar, Newport Pagnall. Qualifications and curriculum authority, London. Available at: http://www. camexam.co.uk.

Pollitt, A. 2004. *Let's stop marking exams*. Paper presented at the annual conference of the International Association of Educational Assessment, June 13–18, in Philadelphia, USA.

Pollitt, A. (2012). The method of Adaptive Comparative Judgment. *Assessment in Education: Principles, Policy & Practice. 19*(3), 1-20. DOI: 10.1080/0969594X.2012.665354

Popham, J. (1997). "What's Wrong - and What's Right - with Rubrics". *Educational Leadership, 55* (2): 72–75.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*, 444–450

Rasch, G. (198011960). *Probabilistic Models for Some Intelligence and Attainment Tests* (expanded edition). Chicago: University of Chicago Press (original work published in 1960)

Rossi, P.H., & Freeman, H.E. (2004). *Evaluation: A systematic approach* (7th ed.). London, UK: Sage.

Sadler, D.R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. *Assessment, learning, and judgment in higher education, 1*(19).

Shute, V.J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78* (1), 153 – 189.

Song, S. H., & Keller, J. M. (2001). Effectiveness of motivationally adaptive computer-assisted instruction on the dynamic aspects of motivation. *Educational Technology Research and Development, 49*(2), 5–22

Thurstone, L.L. (1927). The law of comparative judgment. *Psychological Review, 34* (4), 273 286. DOI: http://dx.doi.org/10.1037/h0070288

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010a). Effecten van ondersteuning bij schoolfeedbackgebruik. *Pedagogische Studiën, 88*, 90-106.

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010b). Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement, 21* (2), 167-188.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053-1098). Greenwich, CT: Information Age Publishing.

van den Berg, I., Mehra, S., van Boxel, P., van der Hulst, J., Beijer, J., Riteco, A., & Gratama van Andel, S. (2014) *Onderzoeksrapportage SURF-project: SCALA- Scaffolding Assessment for Learning.*

van der Hulst J., van Boxel, P. & Meeder, S. (2014). Digitalizing Feedback: Reducing Teachers' Time Investment While Maintaining Feedback Quality. In R. Ørngreen and K. Tweddell Levinsen (Eds.), *Proceedings of the 13th European Conference on e-Learning, ECEL-2014*, pp. 243-250, Copenhagen, Denmark.

Xu, Y. (2010). Examining the Effects of Digital Feedback on Student Engagement and Achievement. *Journal of Educational Computing Research, 43* (3), 275 – 292.