Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

Encouraging professional learning communities to increase the shared consensus in writing assessments : the added value of comparative judgement

Reference:
Van Gasse Roos, Lesterhuis Marije, Verhavert San, Bouw er Renske, Vanhoof Jan, Van Petegem Peter, De Maeyer Sven.- Encouraging professional learning communities to increase the shared consensus in w riting assessments : the added value of comparative judgement
Journal of Professional Capital and Community - ISSN 2056-9548 - (2019), p. 1-18
Full text (Publisher's DOI): https://doi.org/10.1108/JPCC-08-2018-0021

uantwerpen.be
Institutional repository IRUA

Encouraging professional learning communities to increase the shared consensus in writing assessments: The added value of Comparative Judgement

Abstract

Purpose. The Flemish Examination Centre designed an intervention to establish a professional learning community on the topic of writing assessment. The aim of this study was to investigate the effects of this intervention and explain how this intervention succeeded in establishing a professional learning community.

Design/ methodology / approach. A mixed method design was used to answer the research questions. Quantitative analysis of comparative judgement data provided insight into the effects of the intervention. More specifically was analysed whether examiners judged more in line after the intervention. Qualitative analysis of the conversations within the intervention served to examine how interdependent examiners behaved in the professionalization exercises and to gain insight into how a professional learning community was established.

Findings. The analysis showed that the intervention of the Flemish Examination Centre facilitated the formation of a professional learning community. This was visible in the quantitative analysis. The qualitative analysis showed that highly interdependent activities were helpful in establishing the professional learning community.

Research and practical Implications. This study shows that interactions of high interdependence are beneficial to facilitate professional learning communities.

Originality/value. This study shows that assessment data can guide a well-thought out design of interventions to establish professional learning communities among assessors. Assessment data can be a guidance for supportive group constellations and discussions to improve assessment practices. The key in this regard lies in the level of interdependence that is created among participants.

Keywords:

writing assessment; professional learning communities; shared consensus; comparative judgement

Introduction

Assessing writing performances in a reliable and valid way is not an easy endeavour for teachers and examiners. A common educational practice in assessing writing is a division of work among different assessors. The different standards those assessors use to base their judgements of writing on, imply hazards for inter-rater differences in scoring (Myford & Wolfe, 2003). Or, in other words, large differences in severity may occur. Next to that,

individual assessors complete writing assessments from different perspectives. This might lead to valuing quite different aspects in texts (Messick, 1989; Olinghouse, Santangelo, & Wilson, 2012). And although different perspectives are important for the validity of writing assessments, in the end it is essential that assessors come to agreement on the quality of the products (Woehr & Huffcutt, 1994).

Given the importance of high quality judgements (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007), particularly in high stakes judgement settings, research has emphasized the importance of implementing ways to increase the consensus among assessors (Knoch, Read, & von Randow, 2007; Weigle, 1999; Woehr & Huffcutt, 1994). Such consensus contributes to higher reliability and increased validity of writing assessments. Therefore, the need for assessor training in writing assessment has been repeatedly indicated (Brown, Eckes, 2008; Glasswell, & Harland, 2004; Baird, Greatorex, & Bell, 2004). Nevertheless, even assessors who are trained have to balance between their own implicit criteria for judging writing and the criteria they were explicitly trained to look at. As a result, the effects of formal training prior to the assessment on assessors judgements have shown to be rather limited (Jølle, 2014; Weigle, 1998).

Training assessors for writing assessment is quite complex. A shared consensus is needed on the aspects to value in texts (e.g. grammar, spelling, content) *and* on the standards of evaluation (e.g. how the A2 level of English proficiency is reflected in texts) (Baird, et al., 2004). To reach such differentiated consensus, researchers suggested a bottom-up and community building approach instead of a formal (top-down) training (Shay, 2004; Skar & Jølle, 2017; Whitehouse, 2012). In this regard, different opinions between assessors are not to avoid but rather to embrace in a sense that they are necessary steps towards shared norms and standards (Colombini & McBride, 2012). Moreover, building consensus on text quality between assessors is not hindered by differences in what these assessors value in writing (Reid, 2007; Wyatt-Smith, Klenowski, & Gunn, 2010). Text quality is a complex and multidimensional construct, which implies that the construct transcends the different aspects that contribute to it. Therefore, considering that estimating text quality by means of identical perspectives is a prerequisite for high quality assessments is outdated. The interpretation of 'training' should not be to teach assessors 'where to look at'. The complexity of the construct 'text quality' requires the creation of optimal conditions to learn and reflect on what other assessors value in writing. In other words, to support them to become a professional learning community, a reflective and learning-oriented group of people, sharing and critically questioning their practice (Stoll, Bolam, McMahon, Wallace, & Thomas, 2006).

In order to construct mutual understandings on text quality, it is essential to identify different perspectives assessors rely on in their judgements. And because of the common division of work in writing assessments, overlap in assessors' work is limited. This implies that differing perspectives of assessors are often hard to detect. This is different when using the method of comparative judgement. Instead of assigning single writing products to different assessors, writing products are presented in random pairs. Those pairs are distributed over multiple assessors in such a way that each product is assessed by multiple assessors. They only have to indicate which of the two products is of higher quality. The Bradley-Terry-Luce model is then applied to rank products from lowest to highest quality (or the lowest to the highest probability to win in a future comparison). The multiple comparisons of assessors provide opportunities to statistically calculate how individuals' judgement patterns deviate from the group consensus (i.e., misfit statistics). Those data have been supposed to and, to some extent, found to indicate differences in the perspectives that assessors rely on in their judgements (AUTHOR, 2016; AUTHOR, 2018; Bramley, 2007;

Whitehouse & Pollitt, 2012). Thus, comparative judgement generates data that permits assessors to reflect and learn upon their assessment practices and, as such, serve as input to develop a professional learning community among assessors.

Despite the major opportunities of Comparative Judgement to inform assessor groups on differences regarding what they value in writing, no research has investigated in-depth whether and how this kind of data can be successfully used to identify and embrace the different perspectives of assessors of writing quality and how this, eventually, results in greater consensus in writing assessments. The present study aims to bridge this gap using a mixed-method case study design. We will investigate the effects of an intervention informed by Comparative Judgement data in the Flemish Examination Centre. An intervention that uses individual examiner data from a writing assessment was set up, aiming at achieving greater consensus in the writing assessments. Our analysis will focus on whether the consensus among assessors increased after the intervention in a sense that higher similarity in judgements was reached. To better understand these effects, also in terms of their validity, we will examine how the intervention contributed to shared understandings on assessing writing. In other words, we will gather insights into the establishment of a professional learning community through assessor interactions within the intervention. Therefore, the following research questions will guide this paper:

1) To what extent did the intervention of the Flemish Examination Centre result in increased alignment in the judgements of an informal writing assessment?
2) How did the interactions within the intervention established a professional learning community among assessors?

The case of the Flemish Examination Centre

The Flemish Examination Centre is an organization that provides opportunities for students to obtain their qualification for the first, second and third grade of secondary education without being registered in a school. The Flemish Examination Centre employs several internal examiners for each course, which are selected because of their (teaching and assessment) expertise in the discipline. Generally, the internal examiners are experienced teachers who changed direction in their career and ended their employment in the regular school system. They work together on a daily basis on, *inter alia*, the type of assessments, the development of assignments and tests and the definition of assessment criteria. In doing so, the internal examiners form a professional learning community on their own.

However, due to the high number of exam-takers each year and peeks in the number of exams in certain periods, the Flemish Examination Centre consults a pool of external examiners next to their internal examiners to judge the exams. External examiners are generally motivated teachers of all grades of regular education who engage to develop and judge exams outside of their teaching hours. Also retired teachers are included in the pool of external examiners. The pool of external examiners remains quite stable across the year. Consensus in judgements of internal and external examiners is one of the main concerns of the Flemish Examination Centre. Therefore, two professionalization days are organized each year to provide the external examiners with insights on what is expected of the judgement processes in the organization. Despite that external examiners share the vision and goals of the Flemish Examination Centre and that they function as a professional community together with the internal examiners, it remains unclear whether they judge students' work in similar ways. Related to that, the validity of judgements (i.e. 'Do we look at the same aspects in our judgements?') is a concern.

Together with the Flemish Examination Centre, we took different steps to investigate the consensus among internal and external assessors. First, a Comparative Judgement assessment was used to collect data on the judgements of assessors. Differences between internal and external examiners were analysed and interpreted. Subsequently, an intervention was set up in which internal and external examiners ran through some exercises in which important aspects of judging writing exercises were discussed. Finally, after a second Comparative judgement assessment changes in the judgements of the deviating external examiners were evaluated. The duration of this process was 7 months in total. In the next sections, more information will be provided about the theoretical background and methods used in the design of the process.

Theoretical background

The aim of this study is to investigate whether and how a professional learning intervention increases the consensus among assessors in their judgements. To this end, we will first provide more detailed information on the data comparative judgement generates to inform us about the consensus among internal and external examiners prior to and after the intervention . Subsequently, we will delineate our approach to study the interactions within the intervention to evaluate whether a professional learning community was established. This information on the intervention process will provide opportunities to get more grip on why changes in judgements did or did not occur.

Misfit statistics

The first research question aims to investigate changes in the alignment among internal and external examiners after an intervention of the Flemish Examination Centre. To identify differences in alignment statistically, the Flemish Examination Centre used the method of comparative judgement.

Comparative judgement generates misfit statistics. Such statistics identify examiners whose judgements deviate from the common sense (Pollitt, 2012a; 2012b; Whitehouse & Pollitt, 2012). Misfitting assessors prefer, for example, low-ranked texts over high-ranked texts in pairs. Such choices deviate from the consensus (the rank order). Making counter-logical choices may be due to diverging conceptualizations of the competence being assessed, in our case: writing (AUTHOR, 2016; Bramley, 2007; Whitehouse & Pollitt, 2012). Our own analysis confirmed this for the case of the Flemish Examination Centre. It showed that statistically deviating examiners also valued different aspects in writing than the other examiners (AUTHOR, 2017).

Professional learning communities

The second research question focuses on identifying interactions that contribute to reaching common understandings on what to value in writing. In other words, we aim to evaluate the interactions of assessors from the perspective of professional learning communities.

Central in the concept of professional learning communities is the notion of community. It means that a group of people are involved and that mutually supportive – professional - relations are established among these people (Stoll et al., 2006). In addition to that, the concepts' focus lies on learning. McLaughlin & Talbert (2001) point out that not all professional communities exhibit an improvement orientation. Yet, professional learning communities do, in a sense that they promote learning, both individually and collectively (Stoll et al., 2006).

The core dimensions of professional learning communities are shared beliefs and values, shared and supportive leadership, supportive structural and relational conditions, intentional collective learning and sharing practices with peers (Hord, 2009). Therefore, professional learning communities are not easy to establish (DuFour, 2007). And even if groups of professionals become at some point professional learning communities, the sustainability of professional learning communities cannot be taken for granted. Research on the effectiveness of professional learning communities has indicated that, for example, mutual responsibility and reflective professional inquiry are essential (Fullan, 2001; King & Newmann, 2001). The key in this regard lies in the type of interactions involved. Effective professional learning communities are characterized by feelings of interdependence among participants involved (Datnow, 2011; Stoll et al., 2006).

Because of this importance of interdependence, we use Little's (1990) framework to evaluate the professional learning community installed by the intervention. The framework is built upon different levels of interdependency in human interactions. Depending on these different levels of interdependency, different learning processes are established. Little (1990) distinguishes four types of interactions, ranging from limited to large interdependency involved: storytelling, helping, sharing and joint work.

In storytelling activities, people are nearly completely independent of one another. The activity comprises a quick exchange of information through daily conversations. Whether or not this information is used depends completely on individuals (Little, 1990). This implies that the use of this information will not derive from shared values or thinking. With regard to assessment practices, storytelling activities contain conversations about assessment situations that educators experienced. For example, language teachers may have conversations about the difficulty of assessing the content of texts with a lot of language errors.

In helping activities, people seek for help or advice and – subsequently – decide independently to follow or ignore the help or advice that is offered (Little, 1990). Helping is less open ended on the side of the help-seeker compared to storytelling activities because of the underlying purpose of help-seeking. This implies that some learning will be established at the level of the help-seeker, but that mutual learning is not necessarily constructed. In assessment contexts, helping can occur when educators experience problems with their judgement tasks. For example, assessment criteria can be unclear or vague which makes that an assessor seeks help.

Sharing implies the distribution of data, materials and methods, or the open exchange of ideas and opinions (Little, 1990). Teachers take initiatives to make aspects of their work accessible for others, and to expose their materials, choices and rationales. Sharing inherits a higher level of interdependence, although people are not bound to share strategies or materials with regard to how they shape their daily practice (Little, 1990). With regard to assessments, it is likely that examiners have strong opinions on appropriate judgement strategies or on aspects that are more or less important to assess. Those opinions or strategies can be the subject of sharing conversations.

The fourth activity in Little's (1990) framework is joint work. The idea in joint work is that people depend on joint work outcomes in their daily practice. As such, joint work implies higher levels of interdependency in terms of collective purposes and collective action, such as work groups and agreements (Little, 1990). In assessment contexts, joint work can be found when examiners make agreements on, for example, judgement strategies or what they do (not) value in (particular) assignments.

Based on the theory on professional learning communities, the different activities within the Little (1990) framework will have a different contribution to the establishment of a professional learning community. To establish such a community, higher interdependent activities (e.g. sharing or joint work) are assumed to be more fruitful (Stoll et al., 2006).

Method

The current study focuses on the effects of an intervention on the validity of a writing assessment and the community-building activities of examiners during the intervention. The triangulation mixed method design merges quantitative data on the validity of a writing assessment and qualitative data on examiners' interactions during the intervention to understand how the intervention contributed to the validity of writing assessment (Creswell & Plano Clark, 2007). In the next sections, we will first give an overview of the quantitative data and analysis included in this study. Afterwards, the qualitative data and analysis will be described.

Quantitative data and analysis: Identifying deviating examiners

Quantitative data to evaluate the validity of examiners' judgements were collected in pre- and post-assessments. The Flemish Examination Centre chose the competence of 'informal writing' for a Comparative Judgement assessment to identify deviating examiners in the pool of external examiners.

Background characteristics of the pre- and post-assessment. The level of English proficiency intended to evaluate in the assessment was A2 according to the European Reference Framework, which is the second lowest level of written proficiency in a foreign language (Council of Europe, 2001). This level implies that language users are capable of expressing themselves using simple wording. In total, 33 written assignments of diverse quality were selected for the writing assessments. These informal letters were original assignments of students and were already scored in previous years. The 33 assignments were randomly selected to achieve sufficient variation in written proficiency.

Design and analysis of the pre- and post-assessment. The pre-assessment consisted of two rounds. In the first round, 14 internal examiners conducted a comparative judgement assessment with an online tool, developed in the Development of a Platform for the Assessment of Competences project (d-pac.be). Next to judging pairs of informal letters, assessors were asked to provide feedback for each letter. The internal examiners were all experts in writing assessment. The assessment among internal examiners resulted in a reliable expert rank order of the 33 informal letters (Scale Separation Reliability of 0.86[1]). In the second round, a pool of 16 external examiners conducted the same assessment. Based on the individual comparative responses/judgements of the external examiners the deviation of each external examiner's judgements to the expert rank order was calculated.

For all examiners, internal and external, the standardized likelihood misfit (l_z) was calculated (Seo & Weiss, 2013). This measure indicates how far decisions of assessors deviate from what is expected from the consensus based on the Rasch model. To illustrate, if an assignment is placed high in the rank order (i.e. is a text of high quality), it is generally expected that assessors prefer this assignment over lower ranked assignments. If, however,

[1] The Scale Separation Reliability (SSR) is used because CJ assessments are analysed using Rasch modelling (Bramley, 2015). The measure reflects the certainty of an assignment's position on the rank order. High SSR values indicate smaller measurement error and more reliable separation of assignments on the final scale of the assessment (Andrich, 1982). SSR has been shown to be a measure of inter-rater reliability (AUTHOR, 2017).

an assessor generally prefers low ranked assignments over high ranked assignments, these judgements are inadequate to the group consensus. A considerable number of these inadequate decisions implies that an assessor will deviate significantly, which is represented by a lower l_z score. Values are identified as misfitting when the l_z is lower or equal to minus two (Seo & Weiss, 2013). In these cases, assessors made a considerable number of choices that go against the statistical model. The judgements of assessors with l_z scores higher than minus two do not consistently deviate from what is expected in the statistical model.

Within the internal examiner group, no deviating examiners could be identified based on their l_z value. Analysis in the external examiner group showed that six out of 16 external examiners deviated from the consensus in the expert rank order (i.e. $l_z < -2$; Table 1). Qualitative analysis revealed that this deviation indicated validity issues (AUTHOR, 2017).

The post-assessment was used to evaluate changes in the values of the l_z of the deviating assessors. Participants of the intervention were asked to participate in the post-assessment. This assessment was identical to the initial assessment (i.e. involved the same set of informal letters). For every strongly deviating examiner the judgements in the post test were again used to calculate the deviation of each examiner to the initial consensus among internal examiners by means of the l_z.

Qualitative data and analysis: professionalization exercises

Participants and group composition. In total, 18 examiners were selected to participate to the intervention of the Flemish Examination Center: 4 internal examiners, representing the values of the organization, and 14 external examiners which were heterogeneous in l_z scores. An overview of the participants (using pseudonyms for confidentiality purposes) is included in Table 1. Most participants were female, both in the group of internal examiners (3 out of 4) and in the group of external examiners (10 out of 14). Participants in both the groups were similar in terms of their prior education. In the group of internal examiners, all participants held a master degree in languages. The group of external examiners consisted both of participants holding a bachelor degree and of participants holding a master degree. Each of the external examiners held a degree in English. The age of participants ranged from 35 years to 67 years. Participants were on average older in the external examiner group (M = 59) compared to the internal examiner group (M = 48).

For the intervention, pairs and groups were composed. The aim was an evenly distribution of internal examiners among the external examiners and of misfitting examiners among non-misfitting examiners. As a result, in the final group composition, each group consisted of at least one internal examiner and one misfitting examiner (Table 1). The participants did not know their or other assessors' l_z score nor whether they or other assessors deviated from the expert consensus.

Intervention. Three exercises were executed in which interactions between internal and external examiners were facilitated to increase the validity of the assessment. In the first exercise, examiners were paired and asked to describe their judgement routines or strategies with each other (Table 1; 'Pair' column). To this end, each pair was provided with four assignments from the assessment. This provided assessors with a concrete context to discuss their judgement routines and strategies in. For the next exercise, the main criteria for assessors to look at in their judgement were previously analyzed from the feedback assessors provided during the assessment. Then, the pairs of examiners were challenged to discuss the importance they dedicated to these criteria. In doing so, they were asked to put the criteria in order of importance towards a bull's-eye. Examiners could also choose to

throw off criteria they did not consider as important. Subsequently, the same exercise was conducted in groups of four examiners (two pairs together – Table 1; 'Group' column). In this third exercise, examiners were explicitly asked to search agreement around the order of importance of the different criteria. However, the plenary feedback showed that this had not been an easy endeavor.

<< Insert Table 1 about here >>

Coding and analyzing process. After consent of the participants, the discussions with regard to the second and third exercise of 6 pairs and 3 groups were recorded and transcribed at verbatim. The data-loss of discussions in two pairs and one group (pair D and E and group EH) were due to participants not giving consent for recording and technical problems with the recorder respectively.

A four-step coding process took place, using Nvivo 11 software. In a first step, a researcher (researcher A) coded the transcriptions inductively, using open codes (Pandit, 1996). A second researcher (researcher B) checked the content of the open codes in terms of validity. This resulted in the need to concretize or rephrase certain codes. In a second step, researchers A and B discussed the conceptual characteristics of axial codes related to the Little (1990) framework (Table 2). Subsequently, the coding process took a deductive approach. Researchers A and B independently put the open codes under the axial codes (step 3). In a fourth step, the inter-rater reliability between researcher A and researcher B on the axial coding (headcodes) was calculated. The kappa value of 0.55 reflects moderate agreement. Because the kappa statistic is underestimated in some cases of high agreement, Scott's pi index was additionally calculated (Gwet, 2008). This index is similar to the kappa statistic, but slightly different in the calculation of expected agreement (Gwet, 2012). Scott's pi index reflected good inter-rater agreement with a value of 0.89.

<< Insert Table 2 about here >>

After the coding process, we analysed the qualitative data by means of binarization (Onwuegbuzie, 2003). If an axial code was present in a pair or group conversation, score 1 was given and if not, score 0. This way, we got insight into which interactions occurred in which pairs and groups. Eventually, the binarization provided evidence on the general occurrence of storytelling, helping, sharing and joint work across pairs and groups.

Results

In this result section, we will first describe the results with regard to the first research question, i.e. to what extent the validity of the post-assessment increased compared to the pre-assessment. Afterwards, we will examine how interdependent examiners behaved during the intervention for a broader understanding of the results in the post-assessment.

Results of the post-assessment

The post-assessment served to investigate whether or not the external examiners that deviated from the internal examiner group in the pre-assessment were better aligned after the intervention. In the pre-assessment, six assessors were identified as significantly

deviating from the consensus in their judgements of informal letters, which could be attributed to their conceptualization of 'informal writing' (i.e., Monica, Jannice, Matt, Barbara, Michael, Nancy). Five of those assessors participated in the post-assessment (i.e. Monica, Matt, Barbara, Michael, Nancy). Jannice dropped out the post-assessment. For Monica, Matt, Barbara, Michael and Nancy, the l_z was calculated again with the data of the post-assessment (see Table 3).

<< Insert Table 3 about here >>

The results show that all external examiners that were identified as deviating from the consensus among internal examiners improved in terms of the l_z. The l_z of one external examiner (i.e. Michael) remained below the value of -2. This means that Michaels judgements were still not aligned to the judgement of internal examiners. All other examiners could not be identified as deviating assessors from the post-assessment data. The l_z scores of Monica (l_z = -0.86), Matt (l_z = 0.93), Barbara (l_z = 0.51) and Nancy (l_z = -1.68) were above -2 in the post-assessment. This indicates that their judgements were better aligned to those of the internal examiners after the intervention.

Interactions during the intervention

The interactions during the intervention were examined by using the Little (1990) framework, which distinguishes storytelling, helping, sharing and joint work. Table 4 provides an overview of the occurrence of interactions during the intervention. As outlined in the method section, score 1 was given to a pair or group when a certain interaction occurred in the conversation and score 0 if not. The sum of these binary scores over pairs and groups is showed in Table 4. To illustrate, score 3 in the cell storytelling – pair exercise means that this type of interaction occurred in three (out of six) pairs.

Table 4 shows that all interactions occurred during the intervention. However, there are differences between the interactions. Storytelling and helping were generally less present in the discussions than sharing and joint work. In all pairs and groups, examiners shared opinions and aimed at reaching consensus in how informal letters should be judged. In the next sections, we will go into detail on the prevalence of storytelling, helping, sharing and joint work activities.

<< Insert Table 4 about here >>

Storytelling and helping. At the lowest level of interdependency, we distinguish storytelling activities. Analysing the content of the pair and group discussions, we find storytelling in half of the pair discussions (i.e. in three conversations) and in all group discussions (i.e. in six conversations). Storytelling generally occurs to illustrate the importance of (aspects of) judging the informal letters. Participants quoted situations out of their daily practice to illustrate what they think good judgements of informal letters contain. For example, one group discusses what can be understood by judging 'lay-out'. While a group member stated that layout include visual aspects, such as whitespaces or paragraphs, another one indicated it is more than that because she got an e-mail of a student

without a salutation; an element of writing that is essential to her but that she could not place under another dimension than 'lay-out'.

Thus, the storytelling fragments found in the content analysis of the pair and group exercises have an illustrative function. Those activities generally serve to support activities with higher levels of interdependency (e.g. sharing or joint work). This is exposed in a conversation between Jannice and Kelly, who discussed the importance of language errors in informal letters in English. In this citation, talking about what they run into in daily practice is used as an illustration why readability is considered important (i.e., sharing).

> *"… some letters have so many errors that the readability is subverted. And in these cases, language errors are important. But a moderate number of errors does not interfere with the readability to me. And then I don't mind so much. Readability is very important to me."*

Helping activities are barely found in the different conversations of the pairs and groups. Only in two groups, participants provided help to group members. In both cases, the level of English proficiency was the subject of these helping activities. To some participants, it was unclear what level of English to expect when judging the informal letters at A2 level. Subsequently, group members tried to name how the level of English should be approached (e.g. *"A2 is the level in which students can make themselves be understood"* – group Christy, Shana, William, Barbara, Lisa).

Sharing. The sharing of judgement strategies or ideas about (what is important) in judging informal letters in English incorporates a higher level of interdependency compared to storytelling and helping. The analysis shows the appearance of sharing in the content of all pair and group conversations. Therefore, compared to storytelling and helping, the professionalization exercises elicit a lot more sharing among participants.

Sharing varies content-wise in the conversations of participants. In the pair and group exercises, the majority of sharing fragments are about the importance of certain aspects of informal writing. In these conversations, participants share their ideas and opinions on why a certain aspect is important for the judgement of informal letters in English. For example, Marta and Monica differed in their opinion about the most important aspect in the judgement of informal letters in English. Monica was convinced that the correct use of English is important and pays a lot of attention to language errors. Marta did not agree with Monica and valued content more than language errors. In this regard, the following discussion starts, in which Marta clearly indicated why the content of the letter is so important to her for judging text quality:

> *"Marta: … No matter how it is written, I primarily find the message the most important. Because, a letter might be written so nice, if the message does not reach the receiver, the letter is worthless in my opinion.*
>
> *Monica: What bothers me the most are language errors to understand the message. But maybe that isn't. It might be personal.*
>
> *Marta: Yes, but if you keep the goal in mind, then I think that the message remains the most important. And therewith, a logical text structure. Because, content is dependent of logic in the text structure. If consistency is lacking, well, then the message does not reach the receiver either."*

Related to the importance of aspects of informal writing, participants discussed aspects they find important with regard to the particular assignment that was given to students. For

example, Christy and Shana discussed the importance of content in this regard. They argued that the content of the informal letters should meet the assignment's requirement to invite a friend to come over for a visit. Next to discussing important aspects of informal writing, a considerable number of sharing conversations are about the content of these aspects. Participants shared their opinions on how they understand certain aspects of informal writing (e.g. lay-out, language errors or rich language use) or how these aspects can be understood differently. For example, William and Barbara had a conversation on the importance of a rich use of the English language and agreed that it is related to the form requirements of the text. In another conversation, Jannice, Kelly, Matt and Julie discussed the nuance that is needed in the aspect 'language errors'. They argued about the importance of certain language errors and where the boundaries are of language errors they find more or less important than others.

Other sharing conversations are about judgement strategies. However, there is a considerably smaller number of conversation fragments in which this last type of content is included. The conversations about judgement strategies are generally discussions about when students should or should not pass for their informal letter. Most of the time, the discussions among participants make clear that this is strongly related to the aspects of informal letters they value the most. For example, in the group conversation of Jannice, Kelly, Matt and Julie, Matt clarified that he is a proponent for a mathematical sum of language errors because this is easy to justify towards students.

Joint work. Joint work among participants is on the highest level of interdependency. At this level, participants come to agreements about aspects of the judgement of informal letters. In the content of all the pair and group conversations, we find indications for joint work. This means that the exercises challenge participants to not only share aspects of their judgement practice, but also to come to agreements with regard to the judgement of informal letters in English.

Generally, joint work is found in conversations about the importance of certain aspects in the judgement of informal letters (e.g. language errors, content or layout). Participants sought agreement in the value they dedicate to certain aspects. In a lot of conversations, examiners valued the same aspects of informal letters and were in line with each other. In these cases, limited discussion was needed about the most important aspects in the judgement of informal letters in English. Often, participants expressed themselves positively about their (easy) reach of consensus. The following fragment from the group conversation of Marta, Monica and Lydia illustrates how discussions among group members lead to reaching agreement:

> *"Marta: … But we can agree that lay-out is the least important?*
>
> *Monica and Lydia: For this assignment!*
>
> *Marta: And the most important is that the communicative purpose is reached. What can be next?*
>
> *Monica: You also need a language aspect.*
>
> *Lydia: Proper language or rich language for this level of proficiency.*
>
> *Monica: Yes.*
>
> *Marta: Yes, but it cannot be quite proper or rich at this level."*

Other joint work conversations are related to the content of certain aspects of writing informal letters (e.g. language errors, lay-out). In some cases, the content-wise discussion of aspects of informal writing results in consensus among examiners about what can be understood under the different aspects. This leads to some examiners broadening their understanding of the aspects. This is illustrated by the discussion between John and Lydia. Both examiners started from a different understanding of the aspect 'language inaccuracy' and discuss it. In the end, they reached a common understanding.

> *"John: Inaccuracy and language errors. There is some overlap there. A language error is an inaccuracy. I don't see a difference there.*
>
> *Lydia: I saw some difference, though.*
>
> *John: Yes?*
>
> *Lydia: Because, for me, accuracy is also not needing 100 words to tell something. Finding the right words to articulate something.*
>
> *John: I see that different. Inaccuracy is just incorrect use of words. For example, tuning Dutch word use to English.*
>
> *Lydia: Those are language errors.*
>
> *John: Yes, or writing table where you mean chair?*
>
> *Lydia: Ok, yes. That's inaccuracy."*

Discussion and conclusion

The method of comparative judgement provides data to inform assessors on the different perspectives they rely on when judging texts. Nevertheless, up until now, to our knowledge no studies have been made available on how these data, misfit statistics, are useful to guide the establishment of professional learning communities among assessors and, eventually, increase the validity of writing assessments of these assessor groups. In order to contribute to the current research base, this study used a mixed method case study design to address (1) to what extent the consensus within a writing assessment increased after the intervention, and (2) how interdependent examiners behaved during the intervention. In the next sections, we will provide an overview of the most important results, discuss their implications and highlight some of the limitations of the study.

This study showed that the data-based intervention of the Flemish Examination Centre was effective to increase consensus in their writing assessment. Via a pre-assessment using the method of comparative judgement, six external examiners who possessed different ideas of quality regarding 'informal writing' were identified by means of l_z statistics (AUTHOR, 2017). Analysis of the post-assessment showed that all examiners judged more in line with the consensus of the expert examiners after the intervention. One external examiner could, despite the increase of his l_z value, still be identified as deviating from the consensus. Given that deviating examiners at first strongly differed in their conception of quality in 'informal writing' (AUTHOR, 2017), the increase of the l_z values indicates that this conception might have changed during the intervention and became more in line with their colleagues' conception of quality in 'informal writing'. The fact that the post-assessment was carried out on short term after the intervention (i.e., examiners finished it less than one month after the intervention), the exercises can explain at least partly the changes in examiners'

judgements. As such, the intervention proved valuable to affect examiners' learning and the overall validity of writing assessment within the Flemish Examination Centre.

In order to explain the effects of the data-based intervention in this case study, we analysed the interactions of examiners within the intervention exercises. These interactions were classified on the basis of the level of interdependence (i.e., storytelling, helping, sharing and joint work). Storytelling and helping are interactions incorporating limited interdependence. They serve examiners with certain information, but do not make examiners responsible for changes in their judgement practice. Sharing and joint work are activities that inherit higher levels of interdependency. The fact that examiners are provided with insights into why other judgement approaches are taken (i.e., sharing) or that they make agreements on future judgement approaches (i.e., joint work) makes them responsible to reflect on how their judgements are going and how this can better be aligned (Little, 1990).

Our analysis revealed that the conversations of internal and external examiners generally involved sharing and joint work; activities that involve higher interdependence. Examiners used examples out of their daily practice (i.e., storytelling) to illustrate which aspects of informal writing they considered important for the judgement of informal letters (i.e., sharing) and to come to agreements on which aspects were more important than others for the judgement of informal letters (i.e., joint work). Despite that storytelling and helping activities were found to a limited extent, the emphasis in examiners' conversations lied on sharing and agreements on the conception of informal writing. This may for a large part explain why judgements of significantly misfitting assessors showed more consistency with the internal assessors.

The great amount of interdependence in examiners' interactions can be explained by the intervention setting. By instructing examiners to work in pairs and groups, a certain level of interdependency was created. Examiners were asked to discuss the role of certain dimensions (e.g. language errors, lay-out) in their judgement of informal letters, which supports the sharing of opinions on which aspects are important. Additionally, examiners were explicitly asked to reach a consensus on the importance of the different dimensions in the group, which implies that examiners were challenged to come to agreements (i.e. joint work). This might have led to higher levels of interdependency among examiners than were found in earlier studies that used Little's (1990) framework (e.g. AUTHOR, 2016; Katz & Earl, 2010; Meirink et al., 2009). The high levels of interdependence indicate that the intervention established a professional learning community. As such, this study confirms the need for high interdependent activities for such communities to be effective (Stoll et al., 2006). Although examiners found it quite hard at times to agree on the important aspects to judge in writing, the concrete context of the assignment made it helpful for them. As such, discussions were padded by concrete examples and disagreements remained task-oriented instead of becoming examiner-oriented.

Furthermore, it is notable that the changed conceptions via such a community were directly visible in examiners' judgements, which also might have been the result of building the intervention around a specific assignment. Thus, this study confirms that professional learning communities are an effective way for assessors to increase the validity of writing assessments (Shay, 2004; Skar & Jølle, 2017; Whitehouse, 2012). The fact that the deviating examiners behaved differently after the intervention indicates that their interactions during the intervention led to changes in their perspective on text quality. The discussions among assessors led to differences in the frame of reference the deviating assessors used in their

judgement. This might be more effective than asking assessors to balance between formal criteria and their own perspective while judging.

Although the case of the Flemish Examination Centre is a promising example in light of the potential of using comparative judgement data on the level of assessors to improve assessment practices, there remain some limitations. The first is that possible changes in internal examiners' conceptions of informal writing were not addressed in the research design. Given that internal examiners were involved during the intervention to share their expertise with external examiners, not only the judgements of external examiners might have been affected, but also those of internal examiners. This might have led to a (slight) shift in the internal consensus or a more shared consensus among internal and external examiners. Future research has to address this issue and take into account that also differences in internal examiners' judgements can occur. Additionally, in the study design the assignments that were judged within the assessment were the same in the pre-assessment and the post-assessment. This implies that, although the time span between the pre-assessment and post-assessment was long (i.e., five months), some recognition of the assignments may also explain the greater alignment among assessors to some extent.

A second limitation of this study is that it only includes data that were collected over a rather limited time span. First, the post-assessment took place in short term (i.e. less than one month) after the professionalization exercises. This was needed to capture the impact of the intervention on examiner's judgements as good as possible. However, we do not know whether the effect would sustain after a longer time period. The current study design did not provide insights into how examiners' judgements evolved after a longer time period and thus how sustainable the professionalization exercises were. Second, the results showed that the intervention was successful in encouraging a professional learning community. Nevertheless, we did not gain information in how the internal and external examiners behaved and interacted in the period after the intervention. To evaluate whether the Flemish Examination Centre succeeded in establishing a sustained professional community, there is need for additional data. Therefore, it is recommended that future research invests in both a short term and a longer term evaluation of learning results and examiner behavior and interactions.

A last limitation lies in the design of the intervention, which was task-specific. Examiners only discussed their judgements of one specific assignment (i.e. an informal letter). Therefore, this study only provides insights into how examiners' judgements are better in line in the context of assessing this (type of) assignment only and not in whether or not some general judgement competences of examiners have improved. Future research could address this issue by investigating whether improvements in examiners' judgement strategies also transfer to other assignments.

Despite these limitations, this study provides a rich description of how a data-based intervention can contribute to better assessment practices in educational settings. The method of Comparative Judgement was particularly useful to generate insights into the reliability and validity of the writing assessment. The inclusion of multiple assessors provides opportunities for an efficient calculation of statistics that are informative about assessment quality (Pollitt, 2012; AUTHOR, 2016). Collecting similar data in common assessment methods (e.g. rubrics or criteria) is time-consuming (Stemler, 2004). Compared to analytic methods, comparative judgement provides many examples in a short time span, which makes the method extremely feasible to involve multiple judges in the same assessment (AUTHOR, 2018). As such, Comparative Judgement, and the data it generates, can be a powerful

method to generate data to investigate assessment quality and to design more focused interventions, as was illustrated by the case of the Flemish Examination Center.

To conclude, this study shows that assessment data can guide a well-thought out design of interventions to establish professional learning communities among assessors. Although the dedication and eagerness of assessors to collectively strive towards more reliable and valid writing assessment is essential, assessment data can be a guidance for supportive group constellations and discussions to improve assessment practices. The key in this regard lies in the level of interdependence that characterizes interactions among participants.

References

AUTHOR (2016)

AUTHOR (2016)

AUTHOR (2017)

AUTHOR (2018)

AUTHOR (2018)

Andrews, D., & Lewis, M. (2007). Transforming practice from within: The power of the professional learning community. In L. Stoll & K. S. Louis (eds) *Professional learning communities: Divergence, depth and dilemmas*. Maidenhead: Open University Press.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Evaluation assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114-129.

Baird, J., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331-348. doi: 10.1080/0969594042000304627

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. A. Baird, H. Goldsteing, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246-300). London: QCA.

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

Colombini, C. B., & McBride, M. (2012). "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing, 17*, 191-207.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K.: Press Syndicate of the University of Cambridge.

Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks: SAGE Publications, pp. 62-79.

Datnow, A. (2011). Collaboration and contrived collegiality: Revisiting Hargreaves in the age of accountability. *Journal of Educational Change, 12*(2), 147-158.

DuFour, R. (2007). Professional learning communities: A bandwagon, an idea worth considering, or our best hope for high levels of learning? *Middle School Journal, 39*(1), 4-8. DOI: 10.1080/00940771.2007.11461607.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25* (2), 155-185.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.) New York and London: Teachers College Press and Routledge Falmer.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29-48.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC; Gaithersburg, 279 pp.

Hord. S. M. (2009). Professional learning communities. *Journal of Staff Development, 30*(1), 40-43.

Jølle, L. (2015). Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice, 22*(4), 458-474. doi: 10.1080/0969594X.2015.1034087

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*, 151 -177. doi:10.1007/s10763-013-9497-6

Katz, S., & Earl, L. (2010). Learning about networked learning communities. *School Effectiveness and School Improvement, 21*(1), 27-51. doi: 10.1080/09243450903569718

King, M. B. & Newmann, F. M. (2001). Building school capacity through professional development: Conceptual and empirical considerations. *International Journal of Educational Management, 15*(2), 86-93.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing writing, 12*(1), 26-43.

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record, 91*(4).

Mclaughlin, M. W. & Talbert, J. E. (2001). *Professional communities and the work of high school teaching*. Chicago: University of Chicago Press.

Meirink, J. A., Meijer, P. C., Verloop, N., & Bergen, T. C. M. (2009a). How do teachers learn in the workplace? An examination of teacher learning activities. *European Journal of Teacher Education, 32*(3), 209-224. doi: 10.1080/02619760802624096

Messick, S. (1989). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. Research report RR-94-45. Princeton, New York: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386- 422.

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices (Vol. 27)*. Leiden, The Netherlands: Brill.

Onwuegbuzie, A. J. (2003). Effect sizes in qualitative research: A prolegomenon. *Quality & Quantity, 37*, 393-403.

Pandit, N. R. (1996). The creation of theory: A recent application of the grounded theory method. *The Qualitative Report, 22*(4), 1-15.

Pollitt, A. (2012a). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*, 281 –300. doi:10.1080/0969594X.2012.665354

Pollitt, A. (2012b). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*, 157–170. doi:10.1007/s10798-011 -9189-x

Reid, L. (2007). Teachers talking about writing assessment: Valuable professional learning? *Improving Schools, 10*, 132-149.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-19.

Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: a review of the literature. *Journal of Educational Change, 7*(4), 221-258.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. doi: 10.1037/h0070288

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester: AQA Centre for Education Research and Policy.

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. Manchester: AQA Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/research-library/usingadaptive-comparative-judgement-obtain-highly-reliable-rank-order-summative-assessment

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*, 59-75.