Validity of comparative judgement to assess academic writing : examining implications of its holistic character and building on a shared consensus

Validity of Comparative Judgement to assess academic writing: examining implications of its holistic character and building on a shared Consensus

Tine van Daal[*1]

Marije Lesterhuis[*1]

Liesje Coertjens[2,1]

Vincent Donche[1]

Sven De Maeyer[1]

[*]Marije Lesterhuis and Tine van Daal contributed equally to this article as first author

[1] Department of Training and Education Sciences, Faculty of Social Sciences, University of Antwerp, Belgium

[2] Psychological Sciences Research Institute, Université catholique de Louvain, Belgium

Tine van Daal: tine.vandaal@uantwerpen.be, Sint-Jacobstraat 2, B-2000 Antwerp, +32 3 265 48 25

Marije Lesterhuis: marije.lesterhuis@uantwerpen.be, Gratiekapelstraat 10, B-2000 Antwerp, +32 3 265 46 34

Liesje Coertjens: liesje.coertjens@uclouvain.be, Place de l'Université, B-1348 Louvain-la-Neuve, +32 1 047 25 18

Sven De Maeyer: sven.demaeyer@uantwerpen.be, Sint-Jacobstraat 2, B-2000 Antwerp, +32 3 265 49 32

Vincent Donche: vincent.donche@uantwerpen.be, Sint-Jacobstraat 2, B-2000 Antwerp, +32 3 265 49 85

Biographical notes

Tine van Daal is a PhD student at the department Training and Education Sciences (University of Antwerp, Belgium). Her research interests include the assessment of competences using comparative judgement, the complexity of comparative judgement and its validity.

Marije Lesterhuis is a PhD student at the department Training and Education Sciences (University of Antwerp, Belgium). Her research concerns the validity of comparative judgement and takes an assessors' perspective. Next to that, she contributes to research and valorisation objectives of the D-PAC project (www.d-pac.be).

Liesje Coertjens is a Professor on Assessment for Learning at Université catholique de Louvain. Next to this, she works as a postdoctoral researcher at University of Antwerp in the department Training and Education Sciences. Her research interest includes student learning and (peer-assessment of) performance assessment. More specifically, she investigates the efficiency and reliability of rubrics rating and comparative judgement.

Vincent Donche is an Associate Professor of Research Methods in Education at the department Training and Education Sciences of the University of Antwerp (Belgium).

Sven De Maeyer is a Full Professor at the department Training and Education Sciences at the University of Antwerp (Belgium). His major research interests are methodological issues and measurement problems in educational sciences. Next to that, he is the project leader of the D-PAC project.

Validity of Comparative Judgement to assess academic writing: examining implications of its holistic character and building on a shared Consensus

Recently, comparative judgement has been introduced as an alternative method for scoring essays. Although this method is promising in terms of obtaining reliable scores, empirical evidence concerning its validity is lacking. The current study examines implications resulting from two critical assumptions underpinning the use of comparative judgement, namely its holistic characteristic and how the final rank-order reflects the shared consensus on what makes for a good essay. Judges' justifications that underpin their decisions are qualitatively analysed to obtain insight into the dimensions of academic writing they take into account. The results show that most arguments are directly related to the competence description. However, judges also use their expertise in order to judge the quality of essays. Additionally, judges differ in terms of how they conceptualize writing quality, and regarding the extent to which they tap into their own expertise. Finally, this study explores divergence conceptualisation of misfitting judges.

Keywords: comparative judgement, assessment, validity, academic writing

Introduction

Teaching academic writing is about informing students how to write essays and journal articles. Consequently, the best way of assessing students' writing ability is by letting them write such essays and articles. Due to the openness of writing assignments, students vary greatly in their responses. This makes scoring of essays a complex process (Moss, 1994). In terms of writing assessment, the discussion on which scoring method is most suitable is still ongoing (Azarfam, 2012). Most commonly, absolute analytic scoring using criteria or rubrics is relied upon. In this scoring method, judges make an absolute decision regarding predefined criteria in order to grade a single essay on its quality. However, concerns have been raised regarding the validity of these scores.

These validity concerns comprise three different problems related to scoring essays using rubrics. Firstly, research shows that judges differ in how they

conceptualize good writing (Bloxham, 2009). They interpret criteria and essays differently, which can lead to disagreement about the scores. Consequently, this raises problems regarding intra- and inter-rater reliability. Secondly, although criteria can guide the attention of judges to several predefined dimensions (Azarfam, 2012), it has been argued that discerning all relevant criteria in advance harms validity: not every essay may match the criteria (Jones & Alcock, 2014; Sadler, 2009) and criteria may show overlap. Subsequently, judges may adjust scores on sub-criteria to make the final score reflects how they, personally, define the quality of the essay. Finally, although scoring essays using rubrics is depicted as a process of absolute scoring, research shows that judges often make relative judgements: an essay is compared to other previously scored essays or to judges' internal abstract representation of a 'good' essay (Bloxham, 2009; Crisp, 2013). Hence, scores depend on the order in which essays are evaluated or on judges' internal representations of a good essay, instead of being rooted in the quality of the essay itself. To counter these validity problems, an alternative can be found in the recent work of Pollitt and others on comparative judgement (CJ) (e.g., Jones, Swan, & Pollitt, 2015; Pollitt, 2012a, 2012b).

Comparative Judgement: a promising alternative

CJ has its roots in the work of Louis Thurstone (1928) on the measurement of opinions and attitudes. Thurstone's 'Law of Comparative Judgement' states that people are more reliable in comparing two stimuli, for example essays, than in assigning scores to individual stimuli (Thurstone, 1927). More recently, this claim has been repeated by Laming who amplifies the initial thoughts of Thurstone by stating "all judgements are comparisons of one thing with another" (2003, pp. 7-8). The CJ approach uses the strength of human judgement, by asking judges to compare two pieces of student work, and then decide which of the two is better in terms of the assessed competence

(Greatorex, 2007). Consequently, judgement is, in contrast to absolute scoring methods, always relative. Based on multiple comparative judgements a rank-order of essays is generated according to the quality of their writing. This rank-order is based on all decisions made across judges, and results in reliability estimates ranging from .73 (Jones & Alcock, 2014) up to .98 (Heldsinger & Humphry, 2010) as described in Bramley (2015). CJ has been applied to the assessment of a wide range of competences (e.g., mathematics, Jones, Inglis, Gilmore, & Hodgen, 2013; geography, Whitehouse & Pollitt, 2012) including writing (Heldsinger & Humphry, 2010, 2013; Humphry & McGrane, 2015; Pollitt, 2012a). With regard to CJ, validity is rooted in its holistic character and in the shared consensus across judges on what constitutes good writing (Jones et al., 2015; Whitehouse & Pollitt, 2012). Research into the validity of CJ is scarce.

Putting the validity of CJ to the test

In this study, these two critical assumptions underpinning the validity of CJ will be investigated. To challenge these assumptions, the implications arising from them will first be identified and their critical aspects will then be discussed (Kane, 2013).

Holistic character of CJ

A critical assumption underpinning CJ is its reliance on holistic judgement. Judges do not receive criteria to guide their judgement process, but only a general description regarding the writing competence to be assessed. Although this description steers judges in their evaluation of the writing construct, judges are not tied to predefined criteria. Consequently, they can decide individually how different aspects of good writing are valued. This implies that CJ allows judges to vary in the way they conceptualize good writing. The literature on scoring using rubrics indicates that judges differ in their

conceptualization in both focus and broadness. Some judges develop a specific focus regarding the various dimensions of writing they pay attention to (Eckes, 2008; Sakyi, 2000) and regarding the weight given to these dimensions. Moreover, judges also range from having a broad view, focussing on several elements at once, to a narrow view, focussing only on a limited set of elements. Variation between judges in terms of focus and broadness is allowed and even valued within CJ. However, up to now, it is unclear whether and to what extent CJ allows judges to differ in conceptualization. A second implication of the holistic character claimed by its proponents is its building on expertise (Jones et al., 2015). By freeing judges from predefined criteria, CJ is said to enable judges to tap into their expertise, which is assumed to enhance the validity of scoring essays using CJ. Notwithstanding the critical and questionable status of this claim, it is, again, not backed up with empirical evidence.

Another consequence of this reliance on the expertise of judges is its trust in judges' ability to discern between essays on relevant features of the competence (Bramley, 2007). The latter is important because judges should base their decisions on construct-relevant features of the essays while avoiding letting irrelevant features influence their decision (Messick, 1989). To the best of our knowledge, Whitehouse (2012) conducted the only study that looked into what features of (geographical) essays judges took into account while deciding on the best essay. She asked judges after each comparison to provide comments about their reasons for a specific judgement. Examination of these comments indicated that most arguments were in line with the assessed competence, which is a precondition to claim validity in terms of the rank-order. Whitehouse (2012) attributed these results to the existence of a community of practice on assessment originating from examination training (for national testing programmes), published mark schemes, and general statements with regard to the

assessed competence. She questions the extent to which judges base their judgement on construct-relevant features can be realised in settings where such a community is absent.

Shared consensus

The use of CJ is also built on the claim that rooting the final rank-order in the shared consensus across judges adds to its validity. This claim is justified by pointing to the fact that each essay is looked at by multiple judges, which makes the final score a reflection of the judges' collective expertise (Jones et al., 2015; Pollitt, 2012a). This implies that the shared conceptualization of writing quality and judges' collective expertise defines the final rank-order. However, since judges are allowed to vary in their conceptualization of what constitutes good writing, it is unclear whether or not the final rank-order covers every dimension of good writing. For example, imagine every judge basing his/her decisions solely on the structure of essays, than it is arguable if the final rank-order represents a scale in writing. As Messick (1989) emphasized, full construct-representation is a necessity to claim validity (Messick, 1989). The extent to which the shared consensus within CJ realises full construct representation cannot be taken for granted and needs empirical underpinning.

Furthermore, in previous studies of CJ, judges are identified as taking decisions that deviate from the shared consensus (Pollitt, 2012a). These judges are labelled as 'misfit' and statistically identified by examining the extent to which judges make unexpected decisions (Pollitt, 2012a. See Pollitt 2012a, 2012b for an extended explanation of the identification of misfitting judges). A misfitting judge can either be judging too inconsistently compared to the average judge by taking too many unexpected decisions or judging too consistently compared to the average judge. However, within CJ, misfit statistics have to be interpreted relatively. Misfit statistics quantify the extent to which judges are misfit compared to the whole of judges (Pollitt,

2012a). It is suggested that judges with the highest misfit statistic may have a diverging conceptualization of the competence to be assessed (Bramley, 2007; Whitehouse & Pollitt, 2012). Empirical evidence underpinning this suggestion is lacking. Furthermore, it is unclear whether this divergence in conceptualization refers to the focus or the broadness of the judgements, or to both. Misfit judges may be focusing on construct-irrelevant features, or take into account too few or too many arguments when making decisions.

This study

This study examines the implications of the holding of two critical assumptions underpinning the use of a comparative judgement (CJ) for the assessment of academic writing: specifically, its reliance on holistic judgement and its rooting of the final rank-order in the shared consensus across judges.

In this study, judges' arguments as to why they choose one essay above another are examined to investigate whether or not the shared consensus across judges covers the whole of argumentative writing (RQ1), because full construct-representation is required to underpin the claim of validity (Messick, 1989). Following this, the degree to which the holistic character of CJ enables the tapping of judges' expertise is tested (Jones et al., 2015) (RQ2). Furthermore, the validity of CJ is said to be enhanced by allowing judges to vary in what they conceptualize as 'good academic writing'. Since empirical evidence underpinning this variation in conceptualization within CJ is lacking, this study will examine to what extent judges differ in focus and broadness in terms of their conceptualization regarding academic writing. Also variation in judges' use of expertise will be examined (RQ3). Finally, previous studies suggest that misfitting judges (i.e., judges statistically deviating from the shared consensus) may show a conceptualization that diverges from the shared conceptualisation across judges

(Bramley, 2007; Whitehouse & Pollitt, 2012). This is explored by comparing differences in focus and broadness of arguments provided by non-misfitting and misfitting judges (RQ4).

Methods

Context of the study: Course and Competence

The explorative study is carried out in a course that is part of a pre-Master's programme at a Flemish (Dutch-speaking part of Belgium) university. Within this course, students' skills in academic writing are assessed. Academic writing is thereby operationalized as the extent to which students can write an argumentative essay that synthetizes scientific literature and applies general guidelines with regard to academic writing (i.e., the structure of a scientific article, the APA-rules). Students have to write a review article of 700 words that integrates the insights of three academic research articles on one specific topic. For this study, 41 essays of one examination period are selected (June 2013). The essays covered one of the following topics: teacher learning, feedback seeking, effectiveness of school inspection, and the relationship between self-determination and students' involvement. The essays were anonymously presented to the judges.

Judges

As the papers are targeted at an academic audience, professors and researchers acted as judges. Eleven researchers were asked to participate in this study. Although they are experienced in reading and writing academic articles, these judges do not have a shared experience or a common training in evaluating essays. Judges varied in expertise with regard to the assessment of academic writing, ranging from almost no experience of

assessing essays, to highly experienced. All judges worked at the department where the study took place, indicating that they were familiar with the topic. However, none of the judges was involved in the academic writing course.

Therefore, judges were provided with information on the course before they started their judgements: the description of the competence "academic writing", as well as the assignment that was given to the students. Judges did not receive a rubric list with which to judge the essays, nor training, but were explicitly instructed to assess the academic writing of students by comparing two essays and deciding which was better. Judges were told to keep the provided description of academic writing in mind during the judgement process in order to enhance alignment with the task. The competence description is given below (translated from Dutch):

The student knows the most important principles underpinning academic writing and is able to apply these principles. More specifically:

- The student has insight into the most important principles underpinning academic writing.
- The student is able to write an argumentative text and to synthesize in a scientific manner.
- The student can apply the APA-rules.
- The student has insight into the structure of a scientific article (including abstract, problem statement, conclusion and discussion).

Data gathering

Pairwise comparison data were collected in ten judgement rounds following the approach described by Pollitt (2012a, 2012b). In the whole judgement exercise, 236 comparisons of all possible comparisons (236/820 comparisons = 28.8%) were made.

Each essay was assessed by 4 to 9 judges and compared to at least 8 and, at most, 16 other essays. The pairwise comparison data were modelled according to the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) using the R-package BradleyTerry 2 (Turner & Firth, 2012). Rank-ordering the 41 essays from worst to best resulted in an 'academic writing' scale with acceptable reliability (Scale Separation Reliability (SSR) = .84).

To answer the research questions, qualitative data were gathered. After each comparison, judges were asked to explain why they made their decision using the following prompt: "Briefly explain why you took this decision". It was decided to not compel judges to answer this question, because this might result in less intuitive answers, and could potentially distort the judgement process (Whitehouse, 2012). Nevertheless, judges provided an explanation for 229 of the 236 decisions (97%). As some judges provided multiple reasons for one decision, answers were divided into segments. Segments are the most basic part of the raw data that can be assessed in a meaningful way regarding the phenomenon (Braun & Clarke, 2009). This means that, in the current study, every (part of) a sentence informing as to why a judge made a certain decision, is a separate argument. These segments are the unit of research.

This segmentation resulted in 996 arguments. So, on average, 4.35 arguments were given per decision. Subsequently, these arguments were screened with regard to their information. Arguments were considered to be not useful if they did not refer to why a judge had chosen one of two papers. These arguments concerned mainly the difficulty of the decision (e.g., 'difficult decision' (judge A, round 9)). Omitting these arguments resulted in 967 arguments that were retained for analysis.

An initial coding scheme was developed based on the competence description, and informed by an existing coding scheme developed by Cumming, Kantor, and

Powers (2002). This coding scheme focused on the content of the arguments provided, and their relationship with the competence description provided to the judges. To test the initial coding scheme, two researchers independently coded the arguments belonging to the first two judgement rounds. Unclear arguments were discussed, and the initial coding scheme was slightly adjusted. Again, the same two researchers independently coded arguments of two other judgement rounds. This resulted in substantial inter-rater agreement (round 9 kappa = .6; round 10 kappa = .74) (Stemler, 2001). One researcher coded the arguments of the remaining judgement rounds. The researchers discussed every unclear argument.

The final coding scheme discerns three categories of argument: arguments related to the assessed competence and to the competence description (category 1), arguments related to the assessed competence but unrelated to the competence description (category 2) and arguments referring to the general quality of the essays (category 3). Each category consists of several types of argument that further specify the general category (e.g., APA and structure for category 1). The coding scheme, with examples of quotes, can be found in Table 1. The qualitative data were quantified by counting how often each type of argument was used. These counts were transformed to the relative frequency with which every type of argument was used.

Analyses

To warrant full construct-representation (RQ1), every aspect of academic writing that is defined within the competence description should be backed by arguments. As the arguments within category 1 are directly related to the competence description, an overview of the relative frequencies of the different types of argument within this category will be given to answer RQ1. However, judges might have provided other construct-relevant reasons not directly related to the competence description (category

2) for picking a certain essay. This category of arguments is assumed to reflect judges' use of expertise. As such, examining to what extent judges used this category in general, and the different types of arguments within this category, specifically provide evidence regarding the extent to which CJ enables judges to tap into their expertise (RQ2).

To explore differences between judges in terms of the conceptualization of academic writing (RQ3), their variation in expertise, focus and broadness, reflected in the arguments they provided, will be examined. Firstly, since arguments only referring to the competence and not the competence description (category 2) are assumed to reflect judges' use of expertise, the variation between judges in the relative use of the three main categories of argument will be examined. The significance of differences will be tested using the *chi²* statistic. *w* will be reported as effect size. This effect size is based upon Cramer's V, and takes into account the number of rows and columns of the matrix. To interpret *w*, the following guidelines will be used: .10 - .29 is a small effect, .30 - .49 is a medium effect and ≥ .50 is a strong effect (Volker, 2006). Secondly, using the same approach, judges' differences in focus will be investigated by examining differences in judges' relative share of the various types of argument used. This will be done for the arguments within category 1 and category 2 separately. Finally, to check whether judges vary in broadness, an ANOVA will be performed to test whether judges differ in the average number of arguments provided per comparison. η^2 indicates effect size, using Cohen's rules (1988; > .14 = strong effect, > .06 = medium effect, > .01 = small effect).

To understand the need for a shared conceptualization from a misfit perspective (RQ4), misfit judges' conceptualization will be compared to that of the average non-misfitting judge. Misfitting judges will be identified by their information-weighted mean square value (infit). Infit measures indicate the extent to which judges take

unexpected decisions and are preferred over outfit measures, because the first are the most robust (Pollitt, 2012a; See Pollitt, 2012a, 2012b for an explanation and formula for infit and outfit). Following the stringent approach suggested by Whitehouse and Pollitt (2012), judges will be labelled as misfit if their infit value is 1 standard deviation above or below the mean infit value. To investigate difference in focus and broadness between misfitting and non-misfitting judges, a similar approach as for RQ3 will be used. Due to the small sample size, differences in focus can only be examined for types of argument across all categories. Because the data could not fulfil the assumption regarding normal distribution, a Mann-Whitney test will be performed to test for significance of differences in focus found instead of a chi^2-test. The Mann-Whitney test takes the rank-order of the most used arguments of the misfitting judge and compares it to the rank-order of most used arguments of the sum of all non-misfitting judges. For calculating the effect size, Z is divided by the squared total number of samples, interpreting $r > .5$ as a large, $r > .3$ as a medium and $r > .1$ as a small effect size (Field, Miles, & Field, 2012). To test for the significance of differences in broadness, a t-test will be performed on the average number of arguments provided per comparison by misfitting and non-misfitting judges. Cohen's d indicates the strength of the effect with $d > .8$ representing a strong effect, $d > .5$ a medium effect and $d > .2$ a small effect (Cohen, 1988). All analyses are performed using R.

Results

Full construct-representation of academic writing and tapping of Judges' Expertise

To obtain an insight into the extent to which the final rank-order fully represents academic writing, we have investigated whether or not every dimension of academic

writing, as defined by the competence description, is backed by arguments provided by the judges. As shown in Table 1, almost 70% of the arguments provided refer directly to the competence description (category 1). Within this category of argument, every dimension of academic writing related to the competence description (e.g., argumentation, APA) is backed by judges' arguments. However, the relative share of the different dimensions varies. The arguments mostly used refer to the structure of the essay (20.2%), the degree of analysis and synthesis (11%), the problem definition (10.8%) and to the application of the APA norms (8.1%). Other types of arguments are named less frequently. Arguments related to the discussion (3.4%), to the argumentation students' used (3.9%), to the abstract or title of the essay (4.6%) and to the conclusion (5.7%) have a relatively low share.

[Table 1 near here]

Furthermore, two other categories of argument are identified: arguments relating to the general quality of the essays (category 3 in Table 1) and arguments related to the competence assessed, but not referring to the competence description (category 2 in Table 1). Arguments falling under category 2 can be seen as evidencing the ability of CJ to tap judges' expertise. About a quarter of all arguments (25.6%) can be categorized under this heading. These arguments mainly point at the style of the paper (15.4%), its content elements (6.4%) and language errors (3.6%). This is an indication of CJ's ability to tap into the expertise of judges. Finally, the general arguments (category 3) concern only 6.6% of all the arguments provided. These arguments refer to the general quality of the papers (6%) or the extent to which students fulfilled the assignment (0.6%). As it is unclear which specific features of the essays the judges took into account, their relationship to the competence assessed cannot be set.

Variety among Judges: Expertise, Focus and Broadness

With respect to variations in the use of expertise, Table 2 shows that judges do differ in their use of the three main categories of argument. Judges B, D, F, H and J give about two-thirds of arguments directly related to the competence description (range: 66.8% - 74.4%). The other third of the arguments are competence-related, but do not refer to the competence description (range: 22.9% - 33.4%) indicating their use of expertise. Their share of general arguments (< 3%) is negligible. Judges A and I also follow this pattern, although providing a little more general arguments (range: 5.5%-6.8%). For judges E, G and K, the share of competence description-related arguments ranges from 64.3% to 79.1%. These judges provide fewer arguments indicating their use of expertise (range: 15.3% - 22.7%). Furthermore, judges E (13.6%) and G (13%) provide slightly more general arguments. Finally, only half of the arguments (51.2%) given by judge C are directly related to the competence description. This judge also provides the highest share of general arguments (34%) and the lowest share of arguments indicating his/her use of expertise (14.9%). The chi^2-statistic points at the statistical significance of these differences ($chi^2(20) = 11.89$, $p > .001$, $w = 0.346$). w points to a medium effect. Judges do differ in their use of expertise.

Regarding variations between judges in terms of focus, examination of the differences in the use of the types of argument within category 1 and 2 evidences that judges vary in focus. Table 2 shows, for example, that within category 1 (arguments related to the competence description) most judges mainly focus at the structure of the essays. However, the second type of argument most commonly provided differs among judges. Judges A, B and K look, for example, most at analysis/synthesis, whereas judges E, F, G and I are more focused on the problem definition. Judges H and J have another focus. Judge H focuses mostly on analysis/synthesis (54.3%) and judge J on the

APA–rules (21.5%). The differences in judges' use of the various types of argument within category 1 is statistically significant ($\chi^2(70) = 564.89$, $p < .001$) and w points at a strong effect ($w = .717$). The same applies to judges' variations in the use of arguments only related to competence (category 2). The chi^2-statistic is statistically significant ($chi^2(30) = 323.34$, $p < 0.001$) with w indicating a strong effect ($w = .542$). From Table 2, it is clear that all judges mainly focus on style. However, the second aspect that judges focus on, varies. Some judges (C, D, G, H, J and K) focus on language errors, while others focus on content aspects (judges A, B, E, F and I). For judge H, again, his/her strong focus on one kind of argument stands out (style, 20%).

[Table 2 near here]

Finally, judges are found to differ in terms of the broadness of their judgements. Some judges take more or fewer aspects into account to justify their decisions. As Levene's test returned statistically significant outcomes ($F(10,225) = 2.639$, $p = .005$), an ANOVA not assuming equal variances is performed. The differences between judges in terms of the number of arguments per comparison is statistically significant ($F(10,225) = 18.89$, $p < .001$, $\eta^2 = .456$) and points to a large effect.

Misfitting Judges

To investigate the differences in conceptualization between judges who performed averagely and judges who were shown to behave relatively different from the other judges (i.e., misfitting judges), the latter were identified by examination of the misfit statistics. Figure 1 plots the infit values of all judges ($M = .87$, $SD = .19$). Applying the more stringent rule for misfit detection (mean +/- 1 *SD*; Whitehouse & Pollitt, 2012) results in the identification of 2 misfitting judges (triangles in Figure 1). Judge G falls above the acceptable range (infit = 1.36), while Judge H falls below the acceptable range (infit = .59).

[Figure 1 near here]

With regard to judge G, Mann-Whitney tests do not confirm the hypotheses that this judge differs in terms of the types of argument used across all categories of argument ($U = 84.5$, $z = -1.18$, $p = .598$) compared to the average non-misfitting judge. However judge G ($n = 22$, $M = 5.05$, $SD = 2.92$) provides, on average, more arguments per comparison than the average non-misfitting judge ($n = 130$, $M = 3.79$, $SD = 2.65$). This difference is statistically significant ($t(150) = -2.02$, $p = .045$) and points at a medium effect ($d = .452$). In contrast, judge H does significantly differ from the average non-misfitting judge in his/her use of the different types of argument ($U = 39$, $z = 3.02$, $p < .001$). This points to a small effect ($r = .28$). Table 2 shows that judge H has focused strongly on analysis/synthesis and on style compared to most other judges. The difference in broadness and in average arguments per comparison, did not render statistically significant different results ($t(150) = -.632$, $p = .528$, $d = .09$) between judge H ($n = 84$, $M = 4.01$, $SD = 2.19$) and the average non-misfitting judge ($n = 130$, $M = 3.79$, $SD = 2.65$). Based on these results, it can be concluded that judge G only differs in terms of average arguments per comparison, and judge H only in focus.

Discussion

Recently, comparative judgement (CJ) is been introduced as an alternative method for the scoring of essays. Although CJ has already shown promising results concerning its reliability, the validity of this method is up to the present time hardly underpinned by empirical research. This paper presents an explorative study into the validity of CJ to assess ‘academic writing’ in a pre-Master’s course at a Flemish university (Belgium). The holding of two critical assumptions underpinning the use of CJ for the assessment of writing are investigated: 1) its holistic character and 2) the meaning of a ‘shared consensus’ among judges. To assess the holding of these assumptions, the implications

arising from both are examined in this study.

Firstly, the extent to which the final rank-order of essays represents the whole of academic writing was investigated. Results indicate that almost 70% of the arguments given are directly related to the competence description. Furthermore, all dimensions of academic writing implied by the competence description are backed up with the arguments of the judges. This is a precondition for construct validity (Messick, 1989). This is not surprising, because the judges were given the competence description at the beginning of the assessment. However, the prevalence of the different dimensions varies. Some aspects of academic writing, for example structure and analysis/synthesis, are considered more often by judges. Meanwhile, judges referred relatively infrequently to argumentation and to the quality of the discussion in the essays. Similar findings are presented in the studies of Huot (1993) and Wolfe and Kao (1996) on the criteria judges used to holistically score single essays. Another explanation for this finding is that novices in academic writing do not master the writing skills to adequately integrate argumentation and discussion in their essays. In that case, this study shows how CJ enables judges to differentiate between essays on characteristics rooted in the quality of the essays. This is an advantage over rubrics as the latter allow less flexibility in this regard and consequently lowers validity.

Secondly, about a quarter of the arguments given related to construct-relevant dimensions of academic writing not defined within the competence description. Judges point, for example, to content aspects or to the style of essays when deciding on the winner of a comparison. This provides evidence that during the CJ process, judges use their expertise to make decisions (Jones et al., 2015). In the current study, the arguments not related to the competence description are, however, very common aspects within the writing domain (as for example ‘style’, see Cumming et al., 2002; Sakyi, 2000). This

study shows that, according to the judges who participated in this study, style should be part of the competence description in the first place. However, the extent to which CJ enables the use of expertise is probably related to the specificity of the competence description: a more specified competence description may leave less room for the use of expertise. Future research should examine this hypothesis.

Both the arguments related to the competence description and the arguments eliciting expertise are construct-relevant for the assessment of academic writing. This replicates the findings of the study of Whitehouse (2012) that concluded that judges' justifications for their choice between two geographical essays were in line with the competence assessed. Whitehouse (2012) pointed to the existence of a community of practice originating from the UK testing culture to explain this result. Within the current study, however, such a community of practice is lacking. Judges were researchers who had not received any training or did not share any experience regarding the assessment of academic writing. This suggests that CJ may also work in contexts without a community of practice.

One of the advantages of CJ, as claimed by its proponents, is its allowance and even appreciation of judges' differences in conceptualization of what constitutes good academic writing. This study provides, for the first time, evidence that judges do differ in their use of expertise, and in focus and broadness of their judgements. Comparing judges' relative use of the three main categories of argument indicates that although most judges gave about one-third of their arguments indicating their expertise, some judges seems less able to tap into their expertise. Judges also differ in the extent to which they use general arguments. Similar to what is found in the literature on rating (e.g., Eckes, 2008), this suggests the existence of 'judging profiles'. Within this study, it was, however, not possible to examine the existence of such profiles, or explore their

potential antecedents (e.g., characteristics of judges, judgement strategies) due to the small number of judges. Future research should include more judges to allow an exploration of judging profiles in CJ.

Judges also differ in their focus and broadness when deciding on the winner of a comparison. This study found that judges strongly differ in terms of the aspects of academic writing that they focus on. Looking at the arguments related to the competence description, most judges focused in the first place on the structure of the essays. But the second aspect paid attention to varied between judges. The same conclusion can be made for arguments only related to competence. Every judge focused mainly on style. However, the second focus varied between judges. This is in line with results from the rating literature, which show that, although judges may assign more or less similar scores to essays, they differ in the criteria they use to make these decisions (Gebril & Plakans, 2014; Lumley, 2002). A main source of these differences is the background of the judges (Pula & Huot, 1993). Within CJ, further research effort should be devoted to examining the impact of judges' characteristics on their decision-making.

In addition, the variation between judges in terms of broadness turns out to be significant. Judges strongly differ in terms of the number of arguments provided per comparison. It is, however, unclear how these results can be explained. Some judges may have a less broad conceptualization of competence. Another possibility is that judges differed in approach towards the judgement task. Judges taking a 'real' comparative judgement approach might be more intuitive, and make quick judgements, which results in a focus on a limited set of aspects and renders fewer arguments per comparison than judges using a more analytic approach (e.g., assessing both essays separately before comparing them and arriving at a decision). Furthermore, the differences in focus and broadness might also be an artefact of the data collection

method itself. Some judges might be less willing or able to specify their reasons for picking an essay than others. Further studies should elaborate on these findings and examine the different explanations given.

Finally, this study explores the differences in conceptualization between misfitting judges, deviating most from the shared consensus, and the other judges. Due to the small sample of (misfitting) judges, results are tentative. The misfitting judge who takes more decisions that deviate from the shared consensus (judge G) provides on average more arguments per comparison than the average non-misfitting judge. The other misfitting judge (judge H) makes fewer surprising decisions compared to the other judges. Judge H focuses strongly on a limited set of aspects. Again, this seems to suggest differences in judgement approach. Unfortunately, the data does not allow further exploration of this hypothesis. Furthermore, misfit statistics need to be taken with caution, because they are always relative, in comparison to the other judges (Whitehouse & Pollitt, 2012). This study is the first to connect misfit to the argumentation that judges provide for their decisions. Since differences in conceptualisation between judges are valued within CJ, judges having a rich conceptualisation of the competence to be assessed may be beneficial to the validity of the final rank-order. Taking this perspective, judge G can be viewed as one of the most valuable judges within the current study. Consequently, labelling this judge as ‘misfit’ is highly questionable. Future studies should further explore this perspective on misfit as well as examine other statistical measures that are indicative of misfit.

As this study is among one of the first to explore the validity of CJ by challenging the holding of two of its critical assumptions, it cannot provide a conclusion on the general validity of CJ. Further research is needed to identify and underpin the validity claims made by the proponents of CJ. Besides, this study has some limitations.

First, the small sample of judges does not allow extensive investigation of the differences and similarities among judges' focus and broadness. However, this study shows that more insight into differences among judges is needed in order to be able to set up valid assessment procedures. The small sample also makes it impossible to generalize the results. Second, the conclusions are drawn based upon the written justifications of the judges. This leads to problems, as it is for example shown that a small percentage of judges' arguments relate to the general quality of the essays. Establishing the relationship of these arguments to the competence assessed is impossible. This is problematic, as the extent to which judges' arguments reflect construct-relevant aspects is a critical requirement to underpin the construct validity of CJ (Messick, 1989). It is unclear whether this is caused by characteristics of the comparison (e.g., difficulty), by the holistic character of the judgements, or by the way qualitative data was collected. This implies problems concerning the distinction between what judges say they judged and what they really took into consideration. Future studies should use methods such as eye-tracking and think-aloud to replicate, complement and elaborate on this study.

Conclusion

In this study, a first effort is made to underpin the validity of comparative judgement (CJ). The implications arising from two critical assumptions underpinning the validity of CJ are examined. This study is set up within a pre-master course on academic at a Belgian university.

Firstly, examination of judges' arguments that justify their decisions points out that the rank-order of essays generated by CJ represents the whole of academic writing. Full construct representation is a precondition for claiming construct validity (Messick, 1989). Secondly, as judges also provided construct-relevant arguments unrelated to the

competence description, this evidences that CJ enables judges to use their expertise while judging. Furthermore, judges were found to vary in their focus, broadness and use of expertise while judging. This underpins, for the first time, the claim made by CJ's proponents that CJ allows differences in conceptualisation between judges. Finally, it was examined whether judges labelled as 'misfit' have a divergent conceptualisation regarding academic writing. Indications were found that this might be to a certain extent the case.

In sum, this study adds to the CJ-literature by providing evidence that the method is not only able to generate reliable scores, but also to provide valid scores with regard to academic writing, even when a community of assessment practice is lacking. Since the use of CJ in the assessment domain is growing, more validation studies are necessary, and the findings of this study need to be replicated and elaborated on in the future.

Acknowledgements

This work was supported by Flanders Innovation & Entrepreneurship and the Research Foundation under Grant 130043.

References

Azarfam, A. Y. (2012). Basic Considerations in Writing Instruction & Assessment. *Advances in Asian Social Science*, *1*, 139–150. Retrieved from http://worldsciencepublisher.org/journals/index.php/AASS/index

Baker, B. A. (2012). Individual Differences in Rater Decision-Making Style: An Exploratory Mixed-Methods Study. *Language Assessment Quarterly*, *9*, 225–248. doi:10.1080/15434303.2011.637262

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*, 209–220. doi:10.1080/02602930801955978

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs the Method of Paired Comparisons. *Biometrika*, *39*, 324–345. doi: 10.2307/2334029

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.

Bramley, T. (Ed.). (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge, UK: Cambridge Assessment.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. doi: 10.1191/1478088706qp063oa

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

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice*, *20*, 127–144. doi: 10.1080/0969594X.2012.741059

Cumming, A., Kantor, R., & Powers, D. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal*, *86*, 67–96. doi: 10.1111/1540-4781.00137

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185. doi: 10.1177/0265532207086780

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. doi:10.1016/j.asw.2014.03.002

Greatorex, J. (2007, September). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work* (pp. 5–8). Paper presented at the meeting of BERA, London.

Heldsinger, S. A., & Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*, 1–19. doi: 10.1007/BF03216919

Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research*, *55*, 219–235. doi:10.1080/00131881.2013.825159

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, *42*, 443–460. doi:10.1007/s13384-014-0168-6

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson, & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). New Jersey: Hampton Press.

Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: the case of fractions. In A. M. Lindmeier, A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education, Vol. 1* (pp.113-120). Kiel, Germany: PME.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*, 1774–1787. doi: 10.1080/03075079.2013.821974

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

Jones, I., Inglis, M., Gilmore, C. K., & Hodgen, J. (2013). Measuring conceptual understanding: the case of fractions. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/handle/2134/12828

Kane, M. T. (2013). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*, 115–122. doi: 10.1111/jedm.12007

Laming, D. (2003). *Human judgment: The eye of the beholder*. Andover: Cengage Learning EMEA.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*, 81. doi: 10.1037/h0043178

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*, 246–276. doi: 10.1191/0265532202lt230oa

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11. doi: 10.3102/0013189X018002005

Moss, P. A. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing*, *1*(1), 109–128. doi: 10.1016/1075-2935(94)90007-8

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

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*, 237–265.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*, 159–179. doi:10.1080/02602930801956059

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. Kunnun (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9, pp. 129–152). Cambridge: Cambridge University Press.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, *7*(17), 137–146. Retrieved from http://pareonline.net/getvn.asp?v=7&n=17

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

Thurstone, L. L. (1928). Attitudes Can Be Measured. *American Journal of Sociology*, *33*, 529–554. doi:10.1086/214483

Turner, H., & Firth, D. (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, *48*. doi:10.18637/jss.v048.i09

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools*, *43*, 653–672. doi: 10.1002/pits.20176

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. Retrieved from https://cerp.aqa.org.uk/research-library/testing-validity-judgements-using-adaptive-comparative-judgement-method

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/using-adaptive-comparative-judgement-obtain-highly-reliable-rank-order-summative-assessment

Wolfe, E. W., & Kao, C.-W. (1996, April). Expert/Novice Differences in the Focus and Procedures Used by Essay Scorers. Presented at the Annual Meeting of the American Educational Research Association, New York.

Table 1. *Relative proportion of the three categories of argument including quotations (N = 967).*

Table 2. *Arguments provided by judges A - K (%).*

Figure 1. Two misfitting judges (*M* infit = .87, *SD* infit= .19).