# The whole is more than the sum of its parts – assessing writing using the consensual assessment technique

Zahn, Daniela; Canton, Ursula; Boyd, Victoria; Hamilton, Laura; Mamo, Josianne; McKay, Jane; Proudfoot, Linda; Telfer, Dickson; Williams, Kim; Wilson, Colin

*Published in:*
Studies in Higher Education

*DOI:*
10.1080/03075079.2019.1711044
10.1080/03075079.2020.1786275

*Publication date:*
2021

*Document Version*
Author accepted manuscript

*Citation for published version (Harvard):*
Zahn, D, Canton, U, Boyd, V, Hamilton, L, Mamo, J, McKay, J, Proudfoot, L, Telfer, D, Williams, K & Wilson, C 2021, 'The whole is more than the sum of its parts – assessing writing using the consensual assessment technique', *Studies in Higher Education*, vol. 46, no. 9, pp. 1907-1918. https://doi.org/10.1080/03075079.2019.1711044, https://doi.org/10.1080/03075079.2020.1786275

# The Whole is more than the Sum of its Parts – Assessing Writing Using the Consensual Assessment Technique

Daniela Zahn[1], Canton, Ursula[2]; Boyd, Victoria[2]; Hamilton, Laura[3]; Mamo, Josianne[2]; McKay, Jane[2]; Proudfoot, Linda[2]; Telfer, Dickson[2], Williams, Kim[2]; Wilson, Colin[2]

*[1] Independent researcher, Italy; [2] Glasgow Caledonian University, Glasgow, UK;*

*[3]Royal Veterinary College, London, UK*

Ursula Canton, School of Engineering and Built Environment, Learning Development Centre, Cowcaddens, Glasgow G4 0BA; Ursula.canton@gcu.ac.uk; 0044 141 2731177

No potential conflict of interest was reported by the authors.

Correction notice:
This is an updated version of the accepted author manuscript which includes corrections made to the original version of record.

This article was originally published with errors, which have now been corrected in the version of record available at:
https://doi.org/10.1080/03075079.2019.1711044

The correction notice was published on 1 July 2020 and is available at:
https://doi.org/10.1080/03075079.2020.1786275

# The Whole is more than the Sum of its Parts – Assessing Writing Using the Consensual Assessment Technique

Evaluating the impact of Academic Literacies teaching (Lea and Street 1998) is difficult, as it involves gauging whether writers 1) gain better understanding of what influences written social interactions, and 2) improve their ability to manipulate language to address readers. Self-report can assess the first; the second can only be evaluated by examining texts and their effect on readers. Texts are commonly assessed with rubrics-based tools focussing on textual features, but their insensitivity to communicative context and readers' perception makes these inappropriate for an Academic Literacies framework (Canton 2018). Consensual assessment, used by Amabile (1996) for creativity, offers a potential solution (Canton 2018). This paper presents a new instrument based on consensual assessment and empirically tests it. Intra-class Correlation Coefficients (ICCs) found moderate (Koo and Li 2016) to good (Cicchetti 1994) agreement among raters, which offers proof of concept for capturing the readers' perception of the complex interactions in writing.

Keywords: Consensus; writing; 'Academic Literacies'; assessment

## Introduction

The concept of Academic Literacies has made an invaluable contribution to research (e.g. Lea and Street 1998, Lillis 2001), pedagogical practice (e.g. Bharuthram and McKenna 2006) and organisational or disciplinary structures (e.g. Cleary and O'Sullivan 2015) by shifting the perception of student writing from a skill to the question of "epistemology and identities" (Lea and Stierer 2000, 35). Academic Literacies have inspired critical engagement with the discourse communities involved in academic writing and their potential to attribute systematic exclusion to the supposed deficiency of non-traditional students who do not bring the required skills (Coffin and Donohue 2001: 65; Lillis and Turner 2001). The pedagogy it has inspired focuses on critical inquiry into writing practices (Adler-Kassner and Wardle 2015) that makes tacit assumptions explicit (Jacobs 2005: 484) and allows students to critically interrogate them (Lillis 2001: 154). At the same time, it aims at increasing students' participation, teaching them to "deploy a repertoire of linguistic practices appropriate to each setting and to handle the social meanings and identities that each evokes" (Lea and Stierer 2000, 35).

"Small-scale ethnographically framed projects" in Academic Literacies research (Paxton and Frith 2014: 173, referring to Lea 2004) have returned insight into the role writing plays in the social interactions in Higher Education (HE).  It is, therefore, not surprising that evaluation of its impact often focuses on writers and their perception (Bharuthram and McKenna 2006; Li and Vandermensbrugghe 2011). The voice of writers is, after all, that of the most important experts on their understanding of discourse communities and the identities they wish to develop in them. On the other hand, there are two important reasons why it is surprising that there are relatively few studies that extend the scope of evaluations beyond the writer (e.g.  Borg and Deane 2011 and Wingate et al. 2011): the first of them relates to external pressures. Although fostering learning is often seen to be the main purpose of Higher Education, its second function, certifying a level of achievement, relies on quantitative assessments methods. Combined with the fact that accountability is, for better or for worse, becoming increasingly important in Higher Education (Dobbins et al. 2016), it can be difficult to defend an approach whose impact is mainly evaluated by the contribution the individual learner, or writer, perceives it to have. It leaves practitioners, who are convinced that most large-scale assessment methods do not appropriately reflect the complex nature of the social act of writing (Canton 2018), without sufficiently "incisive" methods to defend their approach (Leibowitz et al 1997: 15). These difficulties could be regarded as another reason to challenge the priority of quantifiable results in the current Higher Education system, but such an approach, as close as it is to the authors' hearts, can do little to address the more immediate challenge of resourcing teaching writing in an Academic Literacies based framework.

More importantly, however, the strong reliance on writer-centred forms of evaluating the impact of Academic Literacies interventions is surprising, because it does not adequately reflect one of the frameworks' central assumptions, the notion of identities: if teaching aims at enabling students to "deploy a repertoire of linguistic practices appropriate to each setting and to handle the social meanings and identities that each evokes" (Lea and Stierer 2000, 35), the focus on the writer, at the expense of the reader, is failing to acknowledge the social dimension of identities. As Illeris remarks, "ever since the term [identity] was taken up […] by the German-American psychoanalyst Erik Erikson (1950, 1968, 1982), its definition has been understood not just as a psychological but specifically as a psychosocial concept, that is, a concept explicitly including the combination and interaction between the individual and the social environment " (Illeris 2014: 151-2). In other words, identities cannot be developed in isolation from the feedback of a community, nor can they be established without its approbation. It does not only depend on whether students feel they have learned to "act, value, interact, and use language *in synch with* or *in coordination with* other people" (Gee 1999: 14) in an HE context. In order to successfully write in their new identities, their contributions also need to be accepted by the new discourse community.

To evaluate whether teaching has a positive impact on writers' abilities to communicate with their audiences, it is necessary to consider how their communicative performances are received. Most commonly  performance assessment (Weigle 2002: 46) is associated with a model of writing assessment that asks specifically trained raters to use rubrics for context-independent writing tests. However, the idea of performance assessment can be reinterpreted as capturing the intuitive reactions of expert members of the relevant discourse community. Consensual assessment offers a potential method to quantify such subjective, context-dependent judgements. Originally developed by Amabile and colleagues (1982, 1996; also Hennessey, Amabile, and Mueller 2011) to measure creativity, it relies on an operational definition that could be transferred to the notion of successful communication in writing: "A written text can be considered successful to the extent that appropriate observers independently agree it is successful within a specific communicative situation" (Canton 2018: 20). Such an operational definition is, by nature, context-dependent and any assessment based on it takes into account the tacit assumptions or "intuitive expertise" (Amabile 1996: 42) raters' have developed from their previous experience.

While the conceptual argument in favour of adopting the consensual assessment model for writing is detailed further in Canton (2018), the current paper represents a step towards developing such a method by testing whether consensual assessment of writing can work in practice[1]. In other words, the first question is whether the tacit agreement among raters can be quantified through the method that Amabile (1996) used to confirm the existence of a shared concept of creativity and consensus on the degree of a product's creativeness. In the field of writing, successful implementation of this method would mean achieving a good level of agreement among raters. Such a result could provide quantitative evidence that confirms the essential role of tacit knowledge that is "acted upon by expert practitioners but not explicated" (Turner 2001: 155). The paper also addresses a second question: whether a suitable measurement tool can reliably capture this tacit agreement.

Answering these two questions could, first of all, confirm existing insights into the nature of academic writing through an alternative, quantitative method to increase "understanding through methodological triangulation" (Turner et al. 2017: 244). Secondly, it contributes to developing a tool to evaluate the communicative quality of texts that is aligned to the conceptual assumptions of an Academic Literacies framework, which could offer new opportunities for evidence-based practice among writing practitioners.

[1] To present the empirical work in detail within the frame of a standard-length research paper, readers are encouraged to refer to the detailed comparison of different types of assessment and the conceptual argument in favour of transferring consensual assessment to writing in Canton (2018).

## Materials and Methods

### *Design of the Measurement Instrument*

Transferring Amabile's (1996) instrument to a new domain required carefully considered adaptations. In her instrument, the central question about the level of creativity of a product is combined with further questions designed to determine the independence of creativity from other factors, such as technical skill. For writing, understanding of the communicative situation and technical skills contribute to facilitating successful writing (Lea and Street 1998). To acknowledge this dependence means removing Amabile's questions to test independence in the context of writing. Instead we added further questions that allow the instrument to elicit a more nuanced picture of the numerous decisions writers make to communicate effectively. Although most readers are unlikely to identify specific linguistic features that facilitate or hinder communication, their reactions to the text can focus on wider-ranging aspects of the text. The questions thus need to achieve a balance between the following opposing demands:

1) specific enough to differentiate the impact of different decisions on the reader.
2) broad enough to capture the features above without reducing them to basic linguistic characteristics of the text, such as linking words, passive voice, verb tense etc. and oversimplifying the complexity of successful written communication.

The first distinction readers are likely to make stems from what Gee (1999, 11) identified as the 'magical' properties of language: it allows writers to communicate new thoughts in new ways; yet they need to do this within the possibilities of pre-existing structures to ensure communication happens. Communicating successfully thus means creating a text that helps readers understand new ideas while adhering to situationally appropriate conventions, and readers can judge both: how well a text facilitates their comprehension, and how appropriate it is for its context.

In addition to this distinction between two purposes of written communication, it is possible to distinguish between different aspects of a text. While it is important to avoid distinctions that presuppose familiarity with specific tools for linguistic analysis, a distinction between the following umbrella terms is commonly recognised in (academic) texts:

- 'content', i.e. the information conveyed in a text,
- 'structure', i.e. the order in which this information is conveyed and
- 'language', i.e. the linguistic means chosen to convey this information.

Although the specific instantiations of these concepts in different communicative situations and communities needs further exploration for novices (Lillis 2001), the use of these terms in the discussion of texts is wide-spread, including at secondary school level (AQA 2019). Asking readers to focus on these three concepts ensures that questions are formulated in a way that is simple and as clear as possible to understand for the intended raters (Howitt & Cramer 2014).

To reflect these distinctions, the instrument contains six items in two distinct categories:

- Category 1: Writer's Awareness of the Readers' Needs

  Three items to evaluate the effect of the writer's decisions regarding content, structure and language on the reader's ability to follow the text. In other words, the items examine readers' perception of the author's skill in choosing and manipulating the content, structural devices and linguistic possibilities in order to convey the intended meaning.

- Category 2: Writer's Context Awareness

  Three items to evaluate the appropriateness of the choices made with regard to content, structure and language for the given context, i.e. situational demands and purpose of the text. In other words, these items examine readers' perception of the author's skilfulness in adhering to the specific discourse community's conventions and her / his ability to convey social identity and group membership (Gee 1999) through text.

After several iterations of phrasing and rephrasing, an item validation was run.

### ***Item Validation***

#### *Background*

The validation tested whether there is a shared understanding and interpretation of the items through paraphrase matching. In addition to the intended paraphrase, we presented 3 further paraphrases shown in table 1.

[Table 1 near here]

These systematically varied two potential sources of interpretation error, namely the nouns associated with the features (content, structure, language), and the information associated with the verbs or a verb modifier, i.e. paraphrases were chosen that changed the thematic relationships within the sentence (see hyperlink below).

#### *Procedure*

An online questionnaire using Google.forms presented the paraphrases in a different randomised order for each participant. Participants were instructed to read each item and the associated paraphrases before deciding spontaneously which paraphrase best reflected the original.  Since the target group for the rating scale consisted of persons working in the areas of learning, academic and writing development, opportunity sampling was used via dhen@ jiscmail.ac.uk. Participants were not compensated for participation and could remain anonymous; however, a prize draw was held for participants who waived anonymity.

*Results*

36 participants took part in the validation study. Overall frequencies of responses are summarised in table 2

[Table 2 near here]

The frequencies above suggested that indeed the most frequently chosen response was the most appropriate paraphrase of the original question. An inferential analysis of this tendency ( Chi square;  effect of correct paraphrase with $\chi 2$ (1)= 3.44) only approaches statistical significance with p=.06, but odds ratios, which are more  appropriate for such a small sample size (Field 2019) show that the correct paraphrase was 27.5 times more likely to be chosen than the entirely misleading one. It was 4.9 times more likely to be chosen than the one where only verb information is misleading and 15 times more likely to be chosen than the one where only noun information was misleading. Thus, the validation has shown that the questions generally are interpreted along the intended lines.

### ***Scale***

In order to quantify tacit agreement, the type of response needs to be chosen. In her original study Amabile (1982a cited in ibid 1996) uses categories from 1-5, as well as continuous scales on which the end and 3 intermediate points are marked. Research in other areas that measure subjective impressions and perceptions, such as pain research (Couper et al. 2006), or food perception research (Lawless and Heymann 2010, 155), suggests visual analogue scales or line scales can provide a potential alternative. This option was chosen for this study, as the continuity of the scale reflects the perception of text quality as a continuum from better to worse. Unlike the 'bands' or 'categories' in many large-scale writing tests, it also reflects the lack of explicitly formulated descriptors in expert readers' tacit understanding of written communication. Being visually distinct from traditional rubrics this tool further encourages raters to adopt a different approach and to follow their intuition, rather than specific, explicit and externally provided criteria.

### ***Number of Raters***

Defining the number and criteria for suitable raters is a challenge that demonstrates the difficulties of an interdisciplinary approach. Using a small number of raters makes the instrument suitable and more user-friendly for research into a specific, local context. Nonetheless, it can also raise concerns about the statistical power of data collected (Kraemer and Blasey 2015). The current study adhered to the approximate number of ten raters Hennessey, Amabile, and Mueller (2011, 265) used in their studies, as the primary focus was to examine whether inter-rater reliability can statistically confirm the tacit consensus among readers Academic Literacies posits.  Future research can still explore steps to increase the user-friendliness for learning developers.

### ***Criteria for Rater Selection***

Another concern were the criteria on which rater choice was based. After initially using raters they expected to bring "intuitive expertise" (Amabile 1996, 42), later research suggests that consensual assessment of creativity is also possible with less specialist raters, although Amabile (1996, 72) recommends using "judges who have at least some formal training or experience in the target domain". For writing, Academic Literacies understanding of the discourse community is essential (Lea and Street 1998), a tenet that has been increasingly reflected in other approaches to writing, where the specialist skills required (Kellogg and Whiteford 2009) have been understood as contextually-specific expertise that can be challenging to transfer (Anson 2016). Therefore, it would seem wise to use raters who are highly familiar with the type of texts used.

The opportunity sample of raters for this study[2] was highly homogenous: all were learning and writing developers working at the same post-92 university in the UK. Although they were based in different academic schools at this university, they were all working with students on undergraduate programmes that are either accredited by professional bodies, or are designed for students in work, who are completing their degrees part time.

### *Selection of Texts*

To ensure that these raters felt equally "appropriate" as observers (Amabile 1982, 1001), texts were chosen from a student cohort working on the same, typical undergraduate assignment: texts in which students reported their research process and resulting insights into the concept of 'executive summaries,' a genre related to

[2] Eight of them are co-authors of the paper. The first and second author led the study and did not rate.

workplace writing that is often required as part of formal assignments in programmes that are closely aligned to professional practice.

Neither Amabile nor her collaborators on CAT stipulate a specific number of products, but to achieve statistical power (Lipsey 1990; Nakagawa 2005), especially with a relatively small number of raters, the number of texts should not be too small. For this study 29 texts from a student cohort on a single programme of study (in which the non-rater authors were involved) were used after students had finished the module and assessment had been completed. In addition, ethical approval and consent from the student authors were obtained.

***Procedure***

To ensure raters were familiar with the communicative context in which the 29 texts were produced, they were given the brief and a short contextual description, as well as the texts themselves and a rating sheet (see Appendix 1) with the questionnaire (6 questions) and an analogue scale ranging from 0 to 10 where 0 indicated *Not At All* and 10 indicated *Completely and Fully*. The pack for each rater also included the following instructions: to read the texts without conferring with other raters and to read all 29 texts once without rating them. This was done for two reasons: firstly, to mirror Amabile's (1996, 55) instructions as closely as possible, and secondly, to give the raters an opportunity to become familiar with the set of texts. Then, the raters were asked to use the scale and their experience as readers to rate each text, randomised in 3 presentation lists to account for presentation order effects (Krosnick and Alwin 1987; Eisenberg and Barry 1988).  Although Amabile (1996, 36) does not clearly describe how the two mechanisms, comparison between the texts and comparisons between a text and intuitive standards based on previous experience, interact, this procedure ensured that raters could have recourse to both of them. Raters could decide whether they completed the task in one or in several sittings within 2 months.

### *Analysis*

The rating packs with all texts and associated instructions were sent to ten raters, nine of which returned their results by the deadline. This resulted in an uneven distribution of raters across all lists: four raters in list 1, two raters in list 2, three raters in list 3. Furthermore, two texts were removed from the analysis as the same text appeared twice in the pack. This resulted in 27 texts per rater entering the analyses. Ratings on the analogue scale were manually measured and entered into Excel. The distribution of ratings per rater was checked and skewedness was found to be within acceptable limits, which means assumption of normality applied (King 2013). To account for the individual use of the rating scale, all raw data were normalised, i.e. a z-transformation was applied before all data analyses.

Our primary interest for this study centred around understanding whether agreement amongst a group of raters can be measured using the CAT as adapted in our questionnaire. To analyse whether tacit agreement can be determined quantitatively, intra-class correlation coefficients (ICCs) were computed. Firstly, our 2-way mixed-effects model assumes:

$$x_{ji} = m + r_i + c_j + e_{ij} \quad (1)$$

with $_i$ being the text rated (random effect), $_j$ being the rater (fixed effect), $r_i$ being the effect of text $_i$, $c_j$ being the effect of rater $_j$, and $e_{ij}$ being the random unexplained error associated with the model. Therefore, we estimated the ICC (Shrout and Fleiss 1979; McGraw and Wong 1996; Field 2005) using a consistency definition as

$$ICC = (MS_R - MS_E)/MS_R \quad (2)$$

where $MS_R = SS_{Rows} / df_{Rows}$ and $MS_E = SS_E / df_E$.

$SS_{Rows}$ is the sum of squares of the between-text variance. $SS_E$, the sum of squares of error, consists of two elements: the between-judge variability, i.e. how much a single rater's mean deviates from the mean across all raters, and the within-judge variability, that is calculated from how much raters' individual ratings of a text deviate from the mean rating of this text across all raters (Field, 2005). The ICC was calculated using average measures across all six items, because they better represent the way in which all six aspects contribute to the readers' overall impression of the text.

All analyses were performed using IBM's SPSS 25 software package.

## Results

Overall, the average measure ICC was .808 with a 95% confidence interval from .675 to .900 (F(26, 182) = 5.199, $p < .001$). That is, a moderate (Koo and Li 2016) to good (Cicchetti 1994) agreement between raters was found. The CIs appear wide as the number of data points is limited. However, they still fall within or very close to the defined levels of agreement. Our results, therefore, indicate that there is consensus on text quality amongst a group of raters and that it could be measured using consensual assessment as implemented in our questionnaire.

Analyses estimating the ICC for single score consistency were run to investigate the level of consensus for single items. The results indicated a level of agreement that is below the threshold of moderate or good agreement.

## Replication

To ensure reliability we ran a replication of the study under the following conditions: another set of 27 texts produced by a different year group on the same programme was sent to the 9 original raters. The raters followed the same procedure as described above (see part 3). 8 raters returned ratings. Analyses were carried out as described above (see part 4.) with the following results: average measure ICC was .810 with a 95% confidence interval from .680 to .901 ($F(26, 208) = 5.253$, $p < .001$). Again, individual results by item indicated a level of agreement that is below the threshold of moderate or good agreement.

## Discussion

The first question we asked in this study was whether consensual assessment can be adapted to quantify tacit agreement on successful writing as a complex, socially – and culturally bound concept. Overall, the nine independent, expert raters in the first study showed moderate to good consensus in their evaluation of the 27 texts analysed. Moreover, this result replicated in the second study. This suggests that the answer to the first research question is ‘yes’: the shared tacit knowledge that allows members of a discourse community to judge how successfully a writer communicates with them is sufficiently consistent to use an operational definition and can be quantified. Importantly, the quantification relies on averaging across all 6 items to create an overall evaluation of the text that consists of the individual aspects captured in each item.

This consensus across all aspects of the text is, however, not reflected in the results for the individual questions. The higher level of variation at the level of the individual question could result from mainly two general and different sources. The first source could be the instrument itself. That is, the pre-test has shown that the single questions tend to be interpreted along the intended lines, but there is still room for individual differences in how the raters interpreted each question. The second general source of error could stem from individual differences between raters. Despite the fact that the instrument is easy and intuitive to use, raters, esp. those trained in rubric-based assessment might find it (cognitively and epistemologically) challenging to switch from the supposed objectivity of well-defined rubrics to trusting their intuitive expertise as readers (see below). Further rater-related error could stem from rater specific factors, such as length or type of previous experience, but these cannot be explored further with our data. Future studies could address this by capturing more rater specific data.

The successful replication of the first study suggests that the answer to the second research question, whether a suitable measurement tool can reliably capture such tacit agreement, is also yes. The results thus demonstrate that our tool based on consensual assessment is appropriate to capture readers' impressions of how successfully different texts communicate. To emphasise the adaptation to writing, it will be given a new and memorable name: Technique for Writing Assessment by Consensus (TWAC).

Before examining the role the TWAC can potentially fulfil in the context of Academic Literacies teaching practice and research, it could be pointed out that although the TWAC captures tacit knowledge, its use has proven to be less intuitive than expected. Informal feedback from raters[3] demonstrated that various of them felt the lack of 'prior agreement among markers what constituted acceptable/unacceptable performance' led to an acute sense of uncertainty for them. Their feedback could be seen as surprising, as our raters not only brought a minimum of two years' experience as learning and writing developers, and were thus familiar with academic writing across various disciplines. They all brought some familiarity with the concept of Academic Literacies as well, which means that they were likely to be familiar with the role of shared tacit knowledge in academic writing (Elton 2010). Nonetheless, many of them had extensive experience with rubrics-based assessment as well, and it could be argued that their feeling that they needed some form of mediation or standardization before the reading task reflects the different understanding of assessment underlying this practice. Their reactions thus confirm the tension between the qualitative ethnographic tradition of Academic Literacies and that of positivist paradigm: despite their position as expert readers they did not perceive this experience-based knowledge of tacit conventions as a sufficient basis for 'know[ing] where to accurately place a cross' or at least felt that this 'intuitive' way of evaluating texts might be insufficient by the established standards of quantitative assessment practice. By translating the insight that tacit consensus shapes written communication in a discourse community into a quantitative result, the TWAC then does not simply offer corroboration of these insights in a different format, which might be more acceptable in the eyes of those who are deeply rooted in a positivist tradition. It also suggests that the differences between the two approaches might not be as irreconcilable as it seems, rather, that they can potentially complement each other.

## Conclusions and Further Research

The most important reason why a quantitative instrument such as the TWAC can complement existing Academic Literacies research and practice is its ability to formalise the gradual nature of the concept of successful writing. Qualitative research has revealed detailed characteristics of the nature and the place of writing in social structures (e.g. Ivanič 1998). Asking readers to translate their tacit knowledge into spatial perception on a continuous scale and then transforming this spatial representation into numerical values opens up new potential for comparisons: first of all, it offers the possibility to compare the level of agreement between different groups of readers. Comparing the level of consensus we calculated in this paper among a group of Learning Developers and a group of subject lecturers, for example, allows a comparatively rapid conclusion whether Learning Developers' perceptions (and consequently formative feedback to students) closely echo those of subject lecturers, who are ultimately marking texts. The *specific nature* of any differences identified in a quantitative study could be further explored through qualitative means, but the *extent* of such differences becomes more visible through numerical values[3]. Similarly, such a comparison between groups of readers could be used to evaluate whether an Academic Literacies based teaching intervention helps increase students' understanding of the tacit knowledge shared by discourse communities, as it could identify how much agreement there is between their evaluation of texts and that of experienced members of the relevant discourse community.

Secondly, once it has been established that readers agree in their evaluations of texts, it is also possible to draw conclusions on the communicative qualities of the texts

[3] The potential for such comparisons is currently explored by the first two authors in an ALDinHE funded study.

included in the evaluation. This would offer a more systematic way of identifying examples of good writing practice compared to individual recommendations. If further research could additionally demonstrate that the instrument is sensitive enough, it could also be used to examine changes in students' writing before and after an intervention. This would offer a new approach to assessing the impact of Academic Literacies interventions, as it would allow practitioners and researchers to reliably quantify how much students' greater insight into the social aspect of writing translates into greater ability to successfully participate in the written communication of their subjects. It would also allow comparisons between the impact of different types of interventions.

In theory, it could also be possible to compare how successfully writers communicate across a range of different texts. Since the instrument asks readers to assess how well texts communicate in a specific communicative situation, the consideration of context is already included in the task, which means that writers' texts produced for different audiences and in different situations could be included in an overall assessment, following suggestions for clinical skills comparisons with the help of generalisability theory (Bloch and Norman 2012). If further research could develop and test the analytical approach needed for such an endeavour, it could develop an instrument that can evaluate the impact of long-term writing interventions in situations where teaching introduces students to different discourse traditions, focusing both on developing analytical and critical skills and the ability to choose whether, and if so, how to participate in them.

As the previous paragraphs show, most of the applications of consensual assessment depend on further research into the way in which the increased comparability of context-sensitive assessment can be utilised. In many cases the research needed to develop such applications is still significant, which means that the current study cannot yet claim to developed consensual assessment of writing as a mature technique. By demonstrating that the agreement of how successfully writers communicate can be quantified, the paper has, however, confirmed that the tacit knowledge of expert readers can be quantified and thus confirmed many of the insights gained by Academic Literacies researchers. This may only be the first step towards fully developing the potential of the TWAC for Academic Literacies research and teaching practice, but these initial results of this study provide a promising empirical basis for such further development.

## References

Anson, Ch. M. 2016. "The Pop Warner chronicles: A case study in contextual adaptation and the transfer of writing ability." *College Composition and Communication* 67(4): 518 -549. http://www.ncte.org/library/NCTEFiles/Resources/Journals/CCC/0674-jun2016/CCC0674Pop.pdf

Amabile, T. M. 1996. *Creativity in Context*. Boulder, CL: Westview Press

Amabile, T. M. 1982. "Social psychology of creativity: A consensual assessment technique." *Journal of personality and social psychology* 43(5): 997 -1013. doi: 10.1037/0022-3514.43.5.997

AQA. 2019. "Subject Content. GSCE English Language (8700)*". AQA*. https://www.aqa.org.uk/subjects/english/gcse/english-language-8700/subject-content.

Bharuthram, Sh., and S. McKenna. 2006. "A writer–respondent intervention as a means of developing academic literacy." *Teaching in higher education* 11(4): 495-507. doi: 10.1080/13562510600874300

Bloch, R. and G. Norman. 2012. "Generalisability for the Perplexed. A Practical Introduction and Guide: AMEE Guide n. 68." *Medical Teacher*. 34(11): 960-992. doi: 10.3109/0142159X.2012.703791

Borg, E., and M. Deane. 2011."Measuring the outcomes of individualised writing instruction: a multilayered approach to capturing changes in students' texts." *Teaching in Higher Education* 16(3) : 319-331. doi: 10.1080/13562517.2010.546525

Canton, U. 2018. "'It's Hard to Define Good Writing, but I Recognise it when I See it': Can Consensus-Based Assessment Evaluate the Teaching of Writing?" *Journal of Academic Writing* 8(1): 13-27. doi: 10.18552/joaw.v8i1.450

Cicchetti, D. V. 1994. "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment* 6(4): 284 -290. https://pdfs.semanticscholar.org/50d7/f68422d0c0424674f6b235ac23be8300da38.pdf

Cleary, L. and I. O'Sullivan. 2015 . "The Political Act of Developing Provision for Writing in the Irish Higher Education Context" in *Working with Academic Literacies. Case Studies Towards Transformative Practice*. Edited by Lillis, T. M., Harrington, K.,

Lea, M. R. and S. Mitchell. 355-64. Fort Collins & Anderson: WAC Clearinghouse and Parlor Press. https://wac.colostate.edu/books/perspectives/lillis/.

Coffin, C., and J. P. Donohue. 2012. "Academic Literacies and systemic functional linguistics: How do they relate?" *Journal of English for Academic Purposes* 11(1): 64-75.

Couper, M. P., R. Tourangeau, F. G. Conrad, and E. Singer. 2006"Evaluating the effectiveness of visual analog scales: A web experiment." *Social Science Computer Review* 24 (2): 227-245. doi: 0.1177%2F0894439305281503

Deane, P., N. Odendahl, T. Quinlan, M. Fowles, C. Welsh, and J. Bivens-Tatum. 2008. "Cognitive models of writing: Writing proficiency as a complex integrated skill." *ETS Research Report Series* 2008(2): i-36. doi: 10.1002/j.2333-8504.2008.tb02141.

Dobbins, K., S. Brooks, J.J.A. Scott,.M. Rawlinson, and R.I. Norman. "Understanding and enacting learning outcomes: the academic's perspective." *Studies in Higher Education*, 41 (7): 1217-1235. doi: 10.1080/03075079.2014.966668

Elton, L. 2010. "Academic writing and tacit knowledge." *Teaching in Higher Education*. 15 (2): 151-160. doi: 10.1080/13562511003619979

Eisenberg, M. and C. Barry. 1988. "Order effects: A study of the possible influence of presentation order on user judgments of document relevance". *Journal of the American Society for Information Science*, 39(5): 293-300.

Field, A. 2019. *Discovering statistics using SPSS statistics*. London: Sage.

Field, A. 2005. "Intraclass correlation." in *Encyclopaedia of statistics in behavioural science*. Vol 2. Edited by Everitt, B. and Howell, D.C., 948-54 Hoboken, NJ: Wiley

Gee, J. P. 1999. *An Introduction to discourse analysis. Theory and method*. London, New York: Routledge.

Hennessey, B.A. T. M. Amabile, and J.S. Mueller. 2011. "Consensual Assessment". In *Encyclopaedia of Creativity*. Edited by Runco, M. A. and Pritzker, St. R., 253-60 Amsterdam and Boston: Academic Press / Elsevier.

Howitt, D. and D. Cramer. 2014. *Introduction to research methods in psychology*. 4th edition. Pearson Education

Illeris, K. 2014. "Transformative Learning and Identity" *Journal of Transformative Education*. 12(2): 148-63. doi: 10.1177%2F1541344614548423

Ivanič, R. 1998. *Writing and Identity. The Discoursal Construction of Identity in Academic Writing*. Amsterdam: John Benjamins.

Jones, J. 2004. "Learning to write in the disciplines: the application of systemic functional linguistic theory to the teaching and research of student writing." *Analysing academic writing*: 253-273.

Kellogg, R. T., and A. P. Whiteford. 2009."Training advanced writing skills: The case for deliberate practice." *Educational Psychologist* 44 (4): 250-266. https://doi.org/10.1080/00461520903213600

Kim, H.-Y. 2013. "Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis." *Restorative dentistry & endodontics* 38 (1): 52-54. http://dx.doi.org/10.5395/rde.2013.38.1.52#

Koo, T. K., and M. Y. Li. 2016. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." *Journal of chiropractic medicine* 15 (2): 155-163. doi: 10.1016%2Fj.jcm.2016.02.012

Kraemer, H.C. and C. Blasey. 2015. How many subjects?: Statistical power analysis in research. Los Angeles: Sage

Krosnick, J.A. and D.F. Alwin. 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2): 201-219.

Lea, M. R., and B. V. Street. 1998. "Student writing in higher education: An academic literacies approach." *Studies in higher education* 23 (2): 157-172. doi: 10.1080/03075079812331380364

Lea, Mary R. and Barry Stierer (eds.) 2000. *Student Writing in Higher Education. New Contexts*. Buckingham: Society for Research into HE and Open Uni introduction by the editors: 1-43.

Leibowitz, B., K. Goodmann, P. Hannon, and A. Parkerson. 1997. The Role of a Writing Centre in Increasing Access to Academic Discourse in a Multilingual University. *Teaching in Higher Education* 2(1): 5-19. doi: 10.1080/1356251970020101

Li, L. Y. and J. Vandermensbrugghe. 2011. "Supporting the Thesis Writing Process of International Research Students through an Ongoing Writing Group." *Innovations in Education and Teaching International* 48 (2), 195-205. doi: 10.1080/14703297.2011.564014

Lillis, T.M. 2001. *Student Writing. Access, Regulation, Desire*. London and New York: Routledge.

Lillis, T.M and J. Turner. 2001. "Student Writing in Higher Education: Contemporary confusion, traditional concerns". *Teaching in Higher Education,* 6(1): 57-68. DOI: 10.1080/13562510020029608.

Lipsey, M.W., 1990. Design sensitivity: Statistical power for experimental research. Los Angeles: Sage

McGraw, K. O., and S. P. Wong. 1996. "Forming inferences about some intraclass correlation coefficients." *Psychological methods* 1 (1): 30. doi: 10.1037/1082-989X.1.1.30

Nakagawa, S., 2004. "A farewell to Bonferroni: the problems of low statistical power and publication bias". *Behavioral ecology 15*(6): 1044-1045. doi: 10.1093/beheco/arh107

Paxton, M., and V. Frith. 2014. "Implications of academic literacies research for knowledge making and curriculum design." *Higher Education* 67 (2): 171-182. doi: 10.1007/s10734-013-9675-z

Turner, J. 2001. "Academic Literacy and the Discourse of Transparency." In: Jones, Carys, Turner, Joan and Street, Brian (eds.). *Students writing in the University: Cultural and Epistemological Issues*. Amsterdam: John Benjamins.

Turner, S.F., L.B. Cardinal. and R. M. Burton, 2017. “Research Design for Mixed Methods: A Triangulation-based Framework and Roadmap”. *Organizational Research Methods*. 20(2): 243-267 [online] Available from: https://doi-org.gcu.idm.oclc.org/10.1177/1094428115610808

Scardamalia, M., and C. Bereiter. 1987. *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin* 86 (2): 420. doi: 10.1037//0033-2909.86.2.420

Wingate, U., N. Andon, and A. Cogo. 2011. "Embedding academic writing instruction into subject teaching: A case study." *Active Learning in Higher Education* 12 (1): 69-81. https://doi.org/10.1177%2F1469787410387814