The Gradingly Marking Process: Research Background

July 2020

Table of contents

Executive summary

Research has shown that learners lack confidence in writing, with some justification as for many it is the least proficient skill in language learning. While preparation and practice are key to improving writing, it is often a challenge for language learners to gain access to high-quality, accurate and standardised feedback on their work. Teachers have little time available for detailed engagement and the feedback they offer may vary considerably.

Automatic Writing Evaluation (AWE) promises greater autonomy for students while potentially freeing up busy teaching staff to devote their feedback efforts to aspects of writing that require more personal attention. In addition, research shows that AWE encourages learners and provides them with the ability to review their own writing leading to greater motivation to improve.

Research also shows that variability in both classroom teacher and test rater responses to learners' writing is a persistent problem which can never be entirely eliminated. Most high stakes language tests specifically include writing a composition as one of the key components used for assessing language proficiency. Although the weight of the writing score may vary between different exam boards, it usually significantly contributes to the final grade. Differences in the marking criteria and scoring rubrics themselves will inevitably influence the outcome or result of the assessment. The Gradingly methodology has been to take as universal an approach as possible to the criteria and scoring rubrics to rationalise the variations among the different organisations and test providers.

Deviations and inconsistency in marking the writing skill have the potential to skew the final test results. To address these problems, educational and examination bodies are looking towards automating the grading processes, aiming to reduce the marking time and cost as well as increase the accuracy of the results. Automated scoring tools have been under development for over 30 years and evolved together with the progress in a field of Natural Language Processing (NLP).

There is no doubt that the implementation of automation in order to address marking problems has its merits and as shown across a number of research projects. This approach can enhance the reliability of human scoring and assist in minimising problems inherent to manual marking such as inconsistency, errors and routine. However, removing humans from the essay grading process and replacing them with a fully automated approach, would not be beneficial in all cases. It would potentially result in affecting the depth of the assessment or insufficiently handling non-typical cases, where human judgement cannot be yet replaced by technology. As the most recent research shows, computational methods improve the reliability of human marking, they do not substitute it altogether.

However, the future is promising as the progress in ML (Machine Learning) and NLP (Natural Language Processing) driven technologies is impressive and new tools are released at an unprecedented pace. Therefore, it is likely that the scope of automatic grading will be expanded to areas currently reserved for humans. This would allow for the development of a truly automated grading solution that is consistent, accurate, standardised and provides learners with high quality feedback.

Part one - The value of language grading and robust feedback during the learning process for writing

Many language learners have difficulty in producing linguistically accurate, communicatively convincing and discursively competent writing in their target language (Hinkel, 2002, 2004; Silva, 1993). Specifically, for new language learners, it is the skill in which most students are least proficient (Nesamalar, Saratha & The, 2001). Research by Berman & Cheng (2001) found that students themselves also find writing when learning a language more difficult than other skills such as listening and reading. Even though writing is one of the most important elements of language learning (Bjork, 1999; Razali and Jupri, 2014), learners face various challenges which often leave them feeling dissatisfied with their ability in, or mastery of, this skill. These include issues in grammar and syntax (Kaur, Ganapathy & Sidhu, 2012), vocabulary (Haider, 2012) but also other factors such as facing a lack of imagination and specific vocabulary of the given topic (Puteh, Rahamat, Karim, 2010).

Writing is therefore particularly challenging for several different reasons. Firstly, writing assessments attempt to test the know-how of the language from all aspects, with a specific focus on authenticity, context and communications (Crusan, 2002). Secondly, because of the importance of writing and the fact that many cognitive and linguistic strategies are required from learners (Rao, 1997), learners display feelings of self-doubt and anxiety in writing (Thomas, 1993). In official language test environments, learners are aware that essay writing is a direct assessment of both general language ability and of specific writing ability in a range of contexts. As a result, according to official test data from IELTS (IELTS Performance for test takers, 2020), results show that the writing component in the test is 9.19% lower for female test takers and 9.34% lower for male test takers compared to the speaking, listening and reading skills. In another research conducted by Zheng (2010), descriptive statistics showed that the writing component was significantly lower with a mean of 89.22 compared to 164.19 for listening and 166.19 for reading. This pattern, where the writing component is systematically graded lower, supports the case for more help to students in this area.

Since writing attempts to test knowledge of language as a whole, not the individual components of language, its successful production is the result of a wide range of related and integrated skills. As Bereiter (1986) says, “writing a long essay is probably one of the most complex, constructive acts that most human beings are ever expected to perform”. Preparation and practice are therefore key to improving writing skills. Regardless of the learning environment, individual or in a class setting, good feedback is very important for learners to improve on their writing (Wang, Shang & Briody, 2013). Providing learners with feedback has a clear impact on the performance of writing; for example, in research by Ismail et al. (2008), feedback to learners meant they made significant improvements on grammar.

Nevertheless, learners often face limited opportunities to receive accurate assessment of, and commentary on, their writing performance (Bjork, 1994, 1999) and many individual learners might not have any direct access to quality teacher feedback. Even those that are in a classroom environment, might face a lack of considered written reaction and response to their writing from busy or overworked teachers operating with time constraints (Razali and Jupri, 2014).

Therefore, a key challenge for language learners is to gain access to high-quality, accurate and standardised feedback. Even if feedback is given by teachers, it may not always be consistent or useful and the type of feedback can have a significant effect on the writing accuracy and the ability to improve (Bitchener, Young, Cameron, 2005). According to Zamel (1985): ‘ESL writing teachers misread student texts, are inconsistent in their reactions, make arbitrary corrections, write contradictory comments, provide vague prescriptions, impose abstract rules and standards, respond to texts as fixed and final products, and rarely make content-specific comments or offer specific strategies for revising the text.’ He also makes that point that these teachers see themselves as language teachers rather than teachers of writing as a specific skill, ‘What is particularly striking about these ESL teachers’ responses, however, is that the teachers overwhelmingly view themselves as language teachers rather than writing teachers.’

An important factor in providing this feedback is that the interpretations and uses of assessments requires validation, rather than the assessments themselves (Kane, 1992, 1998). Louw (2006) found that standardised feedback is more effective compared to regular feedback and stated that inconsistency can cause issues for students. In a study by Louw (2011), the following problems with feedback were identified:

1. The lack of consistency in technique and error identification by markers.

2. Incorrect focus by markers.

3. Unclear comments by markers.

4. Students' inability to understand and use feedback independently.

5. The amount of time it takes lecturers to comment effectively on students' texts.

6. Lecturers are not always consciously aware of how to provide students with effective writing pedagogy through feedback. This is especially relevant in content subjects where the lecturers are not trained in writing.

(Kasanga, 2004; Louw, 2006; Spencer, 1998; Deng, 2009.)

One type of feedback, form-focused feedback, defined by Ellis (2001, p. 2) as 'any planned or incidental instructional activity that is intended to induce language learners to pay attention to linguistic form' is showing direct impact on composition improvement according to a research by Zohrabi (2012). Attempts have been made through various methods, including but not limited to, online essay grading services by individual teachers, for example the online tool Write & Improve that provides diagnostic feedback at different levels of granularity (Andersen et al., 2013), Criterion (Burstein, Chodorow, & Leacock, 2003) and via the online learning environment PTE Success, where students can receive AWE (Automatic Writing Evaluation). However, according to Hockly (2018), current limitations in the field of NLP (Natural Language Processing) draw into question the effectiveness of these AWE systems. Stevenson and Phakiti (2014: 51), cite 'paucity of research, the mixed nature of research findings, heterogeneity of participants, contexts and designs, and methodological issues in some of the existing research ... as factors that limit our ability to draw firm conclusions concerning the effectiveness of AWE feedback'.

Regardless of automatic or human feedback, there is no doubt that practising writing on a frequent basis is how it can be developed and improved (Ismail, 2007). Human feedback and human opinion of their writing when it is available, tends to be valued by learners as long as they can trust the humans involved. However, there is no doubt that AWE also possesses multiple key advantages, including multiple submissions, visible progress of results and analytics on the writing performance. For example, AWE offers the opportunity to pinpoint recurring patterns of errors in student writing that would normally not be possible for a human to identify on a reasonable timescale (Wible et al., 2001, pp. 308-310). This, in particular, is a great prospect for language preparation in the future.

To sum up, writing is a difficult and complex linguistic skill needing practice to master. However, that practice cannot be as valuable as it should be for the learner if useful feedback is not readily available for each effort made. AWE promises greater autonomy for students while potentially freeing up busy teaching staff to devote their feedback efforts to aspects of writing that require more personal attention, such as custom annotation, audience awareness and communicative effectiveness. In addition, AWE encourages and provides learners with the ability to review their own writing (Chapelle, 2008). These aspects contribute to a higher motivation for language writing (Grimes and Warschauer, 2008).

Part two - Human marking: Issues and benefits

For the past five decades and more, teachers, assessors and researchers have struggled with the issue of how to provide language-learners with consistent, valid, reliable, and therefore useful, measurement of the quality of their writing, measurement which can also point to ways in which they can improve and reach the standard they are aiming for. A great deal of work has been done on this in the interim (Hamp-Lyons, 1991), for example, but nonetheless it is striking that this statement in Hirsch (1977) is still apt: 'The assessment of writing ability is the single most important snag to practical progress in composition teaching and research'. Davida Charney (1984) concurs, saying 'Teachers, administrators, testing agencies, and researchers all need a valid, reliable method of assessing writing ability.' This is a summary of the situation which is as relevant today as it was then.

Mastering the productive skill of writing, becoming confident that their writing is not only just about good enough but actually of a high standard, is one of the most difficult, and most desired goals for a language learner. However, in order to direct and help learners to achieve this, a clear formulation of what we mean is needed and as Hirsch (1977) also states: 'We cannot get reliable, independent agreement in the scoring of writing samples unless we also get widespread agreement about the qualities of good writing.' Agreement is therefore the first problem. As Youn-Hee Kin (2009) puts it, 'Rater variability is a potential source of measurement error. Rater-involved assessment …engages subjective judgments making complete rater consensus close to impossible.'

Human beings have diverse backgrounds, circumstances and assumptions. They bring these to their reading and assessment of learners' writing. In 'Assessing Writing', Weigle (2002, pp. 71-72), summarises the research on rater variables and the effects these variables have on the process of evaluating compositions and assigning scores. She lists:

- the amount of composition teaching or rating experience a rater has
- the academic discipline the rater comes from
- the cultural background of the rater
- the kind of training the rater has been given
- the expectations a rater might have of their students

As Fairbairn and Dunlea's (2017) review of the literature for their British Council report on research development for the Aptis system says, 'The subjectivity of the marking creates rater effects' (Myford & Wolfe, 2003). Fairbairn and Dunlea (2017) point out that 'researchers categorise rater effects differently and disagree about how to analyse the data' but neatly summarise the list of 'the main rater effects' as:

- leniency/severity where raters rate too high or too low
- inconsistency where raters apply the rating scale in a different way to what is intended
- a halo effect where raters are unable to distinguish between different categories and allocate similar scores to everyone
- a central tendency or restricted range where raters avoids extreme ratings or one part of the rating scale
- bias where raters mark a particular group of people in a particular way
- logical errors where raters mark related features of the speaking or writing performance in the same way
- basic errors where raters make marking mistakes perhaps due to fatigue

Any student learner or test-taker looking at these lists would be justified in having some concerns about the possibility of consistency among human markers and the reliability of judgement from rater to rater or teacher to teacher.

For most high-stakes tests, such as Aptis and IELTS, continuous training, multiple marking and as robust a system of quality assurance as they can devise has been the answer. This has involved the use of carefully devised criteria, rubrics and scoring guides, with detailed and robust rating scales. As Schaeffer (2008, p. 467) outlines in his earlier review of the research, however, 'although rater training reduces random error and makes raters more self-consistent, it cannot eliminate rater variability' and points out that it has been observed 'that there is a tension between raters' internal reactions to a paper and their efforts to apply the scale, which persists in spite of rater training.'

Teachers themselves are often aware of issues with judging their students' writing fairly and systematically. They know that they may be allowing prior expectation to influence them or lack of knowledge about how other teachers are doing it to limit their ability to provide the most relevant and appropriate feedback. As Hyland (2003, p. 216) says, 'Teachers are often the only evaluators of their students' writing and so they want to feel confident that they are responding consistently across student scripts and that other teachers would evaluate the work in a similar way.' He goes on to say, 'Unfortunately, however, raters may be influenced as much by their own cultural contexts and experiences as by variations in writing quality. Even when texts are double-marked, research has found that raters can differ in what they look for in writing and the standards they apply to the same text.'

A valid, reliable solution that is not subject to potentially unfair variability is clearly to be desired and in spite of the increased consistency among human markers that scoring rubrics can bring about (Jonsson & Svingby, 2007) only a technological solution could introduce the kind of absolutely consistent, unbiassed accuracy that is needed. Nonetheless, it is agreed that human perception of what good writing is can still play a role. As Bob Broad (2003) argues 'In pursuit of their normative and formative purposes, traditional rubrics achieve evaluative brevity and clarity. In doing so, they surrender their descriptive and informative potential; responsiveness, detail and complexity in accounting for how writing is actually evaluated.' Stuart Riddle (2015) echoes these concerns. Without human input in marking, '…what happens with inferential meaning or drawing on rich contexts, background knowledge, prior learning, cultural and social discourses? These are all part of the complex tapestry of human meaning-making in reading and writing.'

He goes on: ‘As one example, the NAPLAN marking guide refers to the use of classical rhetorical discourse in persuasive writing, including: Pathos – appeal to emotion; Ethos – appeal to values; Logos – appeal to reason. I have not yet come across a computer except in science fiction films that has emotions or values that could be appealed to in any persuasive sense. ‘It is because of this that, as teachers and students around the world have discovered, even the best and most sophisticated computer systems for marking writing can, if someone has that deliberate intent, be gamed or duped into accepting surreal input which is technically correct but actually nonsense.

In summary, automatic marking of writing can give us consistency, accuracy and reliability without bias in a way humans cannot. Rater variability is a persistent problem which can never be entirely eliminated. However, humans can give us an overview of the writing – its sense and impact – in a way computers cannot.

Part three - Defining the criteria with which to assess writing

Variability in evaluating writing is not just a result of variability in the markers' personal relationship with each example. While there is a long history of research, there is no single overall theory or methodology supporting the creation of criteria, scoring rubrics, or marking/rating scales, for writing assessment. Methodologies for developing rating scales may be either data-driven or theory-driven or indeed a hybrid or fusion of these traditions. Using teacher input has been one way forward. For example, Holzknecht et al. (2018, pp. 53-54) describe how four categories of the rating scale used for grading writing performances of the Austrian school leaving exam are arrived at by using teacher input to create them. Building on past research, i.e. scholars' understanding of the nature of language ability and the production of written texts, is another way forward.

Lim (2012) describes the process one leading test provider, Cambridge ESOL, used for their assessment scales and mark schemes. This involved a review of the literature, a review of other testing organisations' way of proceeding and a review of the descriptors related to particular levels of the CEFR. First, they identified the interaction of three elements in writing: cognitive, linguistic and sociolinguistic. Then, 'a number of assessment scales from other Cambridge English exams and from other test providers were also reviewed to determine the state of the art, so to speak, and they all seemed to reflect [these] elements.' Taking proposals from their reviewers, Cambridge decided that they should adopt an approach using analytic criteria, and, 'as to what the analytic criteria should be, as a result of the various reviews, it became clear that having separate sub-scales for each of the elements of language ability would be best, so as to ensure a proper and balanced coverage of the construct.'

The final scale of four analytic assessment criteria consists of ‘one criterion for each of the cognitive, linguistic and socio-linguistic elements, plus one criterion for task achievement’:

Reviewer	**Final**
● Content and Development	● Content
● Communicative Achievement	● Communicative Achievement
● Organisation and Linking of Ideas	● Organisation
● Range and Control	● Language

The descriptors to score the writing using these four criteria were designed to achieve coherence across exams and levels by relating to particular levels of the CEFR, though a certain circularity is explicitly acknowledged: ‘The CEFR levels are based in part on Cambridge ESOL’s suite of exams (Hawkey, 2009; North, 2004; Taylor; Jones, 2006) so a relationship already exists between them.’

Differences in the marking criteria and scoring rubrics themselves will inevitably influence the outcome or result of the assessment. While one test conflates all language issues into 'language control' making this one third of the marks available, another teases it all out into 'grammar, general linguistic range, vocabulary range, and spelling', making it over half of the marks available. Still another has as its categories, 'task response, coherence and cohesion, lexical resource and grammatical range & accuracy', making the specific language element half the marks.

At Gradingly, we too have extensively reviewed the literature and the criteria of all the leading test providers who include writing in their tasks and linked their descriptors to the Common European Framework of Reference. A summary of the main ones has been provided in table 1 below. While there are significant differences in the wording and weighting of the criteria chosen, as Lim notes above, they all reflect basically the same elements. The colour-coding shows all addressing to greater or lesser extent the green, blue and peach coloured areas i.e. content, organisation and language use. The Gradingly methodology has been to take as universal an approach as possible and to rationalise the variations among the different organisations and test providers. To this end, our final three main criteria have been divided into six sub-criteria and as the table shows, they cover all the areas addressed by any of the analytic or the holistic scales a learner may encounter from an exam provider or classroom preparation situation.

IELTS *Each criterion scored 0 - 9*	*Task Achievement or Task Response*	*Coherence and Cohesion*	*Lexical Resource*	*Grammatical Range and Accuracy*			
Cambridge *Each criterion scored 0 - 5*	*Content*	*Communicative Achievement*	*Organisation*	*Language*			
Pearson PTE *Each criterion scored 0 - 3*	*Content*	*Form (no. of words in sentences in continuous text.)*	*Development, structure, coherence*	*Grammar*	*General linguistic range*	*Vocabulary range*	*Spelling*
APTIS *Task-specific holistic rating scale 0-6 overall*	*Task fulfilment/topic relevance*	*Grammatical range and accuracy*	*Punctuation*	*Vocabulary range and accuracy*	*Cohesion*		
ISE *Each criterion scored 0 - 4*	*Task fulfilment*	*Organisation and structure*	*Language control*				
TOEFL iBT *Holistic rubrics covering these areas on a 0-5 overall scale*	*Topic and task*	*Organisation and development*	*Unity, progression and coherence*	*Facility in use of language*			
GRADINGLY *Each criterion scored 0 - 10*	*Content: relevance*	*Content: argument*	*Structure: cohesion and coherence*	*Structure: organisation*	*Grammar: range*	*Grammar: accuracy*	*Vocabulary: range*	*Vocabulary: accuracy*

Table 1.

Thus, with a fusion of objective, robust automatic marking with careful human overview, Gradingly aims to capture a broad and comprehensive coverage of English language teaching and assessment of writing, to provide evaluation and feedback with as independent and universal application in the industry as it is possible to create.

Part four - Natural Language Processing (NLP): Development and applications

Popular language assessment approaches include writing composition as one of the key components used for assessing language proficiency. Although the weight of the writing score may vary between different exam boards, it usually significantly contributes to the final grade. Any deviations and inconsistency in marking the writing skill have the potential to skew the final test results.

Grading of the writing component plays a crucial role in robust testing, however, it comes with a range of challenges, which can be broadly grouped into a linguistic and operational category. The first group involves marking accuracy, reliability, and consistency whereas the latter is related to issues such as the marking time, cost, human resources management, markers training, monitoring, moderation etc.

To address these problems, educational and examination bodies are looking towards automating the grading processes, aiming to reduce the marking time and cost as well as increase the accuracy of the results. Automated scoring tools have been under development for over 30 years and evolved together with the progress in a field of Natural Language Processing (NLP).

NLP is a subfield of Artificial Intelligence (AI) and computational linguistics, where the human language is transformed into a numerical form, allowing computers to process and analyse human text or speech. Examples of application classes include machine translation, speech recognition, question-answering systems, contextual recognition, text summarisation, categorisation, sentiment analysis and a range of text analytics tasks (Sarkar, 2019, pp. 62–65). The practical implementation of NLP systems is very broad and widely used in modern IT and online systems.

Sample use cases involve search engines, home assistants, spam filters, chatbots or recommending systems. NLP technology has been successfully applied in educational and language testing environments. Systems designed to mark a written content benefit from a range of NLP techniques, for example, spellchecking, grammar parsers, plagiarism detection, etc (Lane, Howard & Hapke, 2019, p. 8).

A range of NLP algorithms and techniques provide a set of tools, allowing certain issues pertinent to the manual essay grading to be addressed. One of the key concepts behind the AES applications is building a model containing a discrete set of statistical features contributing to the final grade. Hussain et al. (2019) divided these into two categories:

- Handcrafted features - involves a human expert preselecting a set of features based on which the grading is conducted.
- Automatically featured - so-called end-to-end models, where the features are determined directly by the system without human guidance, usually constructed using Neural Networks and Deep Learning

End-to-end models represent the essay in the form of vectors, which are processed through layers of neural networks and wired to output ratings. This approach often results in a 'black box' system, where the underlying logic of Neural Networks is difficult to understand by humans. Yet, it is important to note that automatically featured systems tend to perform better with extracting deeper semantic features of written content (Liu, Xu & Zhu, 2019). In contrast, handcrafted engineered systems have the advantage of being more explainable for humans and can be expanded by incorporating additional features. Associating the results with the marking rubrics are also easier. Examples of manually handcrafted features can include statistical analysis of certain types of errors, usage of language structures, writing mechanics, presence of chosen discourse elements, structure, range of vocabulary, the specific style features, the correctness of usage etc. (Shermis, 2018, pp. 185–186). The number of features contributing to the final grade varies between systems from a few to several hundred.

Automatic marking models can be developed using both supervised and unsupervised approaches. In the first case, the model is calibrated or trained on a set of preselected and marked essays. The objective of this phase is to build the grading model capable of analysis of the writeup and producing scores correlating to the human markers. The size of the corpora required for training purposes varies depending on underlying application architecture as well as the type of implemented algorithms. The number of essay items usually fluctuates from as little as a hundred to tens of thousands pre-marked samples. An alternative approach relies on measuring the distance between the model answer produced by an expert or teacher and the one submitted by a candidate. This solution is frequently used in systems generating feedback, as the model can point weaknesses, misconceptions or omissions compared to the benchmark answer (Suzen et al., 2020. p. 4).

NLP systems have become increasingly accurate in detecting malformed grammar structures or issues such as spelling errors, capitalisations, incorrect collocations, or assessing the range of vocabulary etc. However, there are aspects of language processing where technology lacks the required level of accuracy. Natural Language Understanding (NLU) and retrieving semantic meaning by machines still poses a major challenge. Over recent years, notable progress has been made in the field of sense disambiguation, contextual word embeddings and related sub-fields. However, the current state-of-the-art algorithms are not yet able to match human-level performance in language understanding or inference (Wang et al., 2018).

Human speech is ambiguous, where a number of meanings can be associated with the same sentence. Language processing tasks which humans tend to do naturally and effortlessly, such as understanding negations, applying sarcasm, understanding intentions or recognising named entities are difficult to grasp by computational methods. During communication, a human writer usually assumes that the reader possesses a general wisdom people usually acquire during their lives. NLP systems unaware of this context and knowledge base, are not always able to distinguish illogical utterances, which otherwise would be spotted by human marker instantly. Furthermore, some researchers have pointed out the possibility of adversarial submissions in order to score higher in writing tasks graded automatically. In some cases, deception tricks could be as simple as generating prompt-irrelevant samples or repeating several times a well-written passage of text (Liu, Xu & Zhu, 2019, p. 6).

In the case of text produced by other language speakers, further complications are introduced by the fact that other language speakers are prone to making linguistic mistakes. This causes deterioration of the accuracy of certain NLP algorithms. Apart from shortcomings pertinent to NLP technology, AES systems suffer from the limited ability to measure the level of creativeness and quality of the writeup (Hussein, Hassan & Nassef, 2019).

As mentioned earlier, AES systems before deployment undergo a calibration or training phase, where the final scoring is compared to a benchmark. Appraising the scoring performance, despite existing multiple frameworks, is not a trivial task (West-Smith, Buttler & Mayfield, 2018). It requires thorough consideration as to which essays should be selected in the system training pool, which human raters should be selected for benchmarking and how to handle non-standard cases and discrepancies in scoring (Raczynski & Cohen, 2018). Lack of the above considerations may result in training models based on biased input, ultimately leading to sub-optimal results.

Conclusion

Writing is a difficult and complex linguistic skill needing practice to master. However, that practice cannot be as valuable as it should be for the learner if useful feedback is not readily available for each effort made. Automatic Writing Evaluation (AWE) promises greater autonomy for students while potentially freeing up busy teaching staff to devote their feedback efforts to aspects of writing that require more personal attention. In addition, research shows that AWE encourages learners and provides them with the ability to review their own writing leading to greater motivation to improve.

Research also shows that variability in both classroom teacher and test rater responses to learners' writing is a persistent problem which can never be entirely eliminated. Most high stakes language tests specifically include writing a composition as one of the key components used for assessing language proficiency. Although the weight of the writing score may vary between different exam boards, it usually significantly contributes to the final grade. Differences in the marking criteria and scoring rubrics themselves will inevitably influence the outcome or result of the assessment. The Gradingly methodology has been to take as universal an approach as possible to the criteria and scoring rubrics to rationalise the variations among the different organisations and test providers.

There is no doubt that the implementation of automation in order to address marking problem has its merits and as shown across a number of research projects. This approach can enhance the reliability of human scoring and assist in minimising problems inherent to manual marking such as inconsistency, errors and routine. However, removing humans from the essay grading process and replacing them with a fully automated approach, would not be beneficial in all cases. It would potentially result in affecting the depth of the assessment or insufficiently handling non-typical cases, where human judgement cannot be yet replaced by technology. As the most recent research shows, computational methods improve the reliability of human marking, they do not substitute it altogether.

However, the future is promising as the progress in ML (Machine Learning) and NLP (Natural Language Processing) driven technologies is impressive and new tools are released at an unprecedented pace. Therefore, it is likely that the scope of automatic grading will be expanded to areas currently reserved for humans. This would allow for the development of a truly automated grading solution that is consistent, accurate, standardised and provides learners with high quality feedback.

Bibliography

Andersen, E., Yannakoudakis, H., Barker, F., Parish, T. (2013). Developing and testing a self-assessment and tutoring system. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, 32–41

Bereiter, C., Scardamalia, M. (1986). Educational relevance of the study of expertise. Interchange 17, 10–19.

Berman, R., Cheng, L., Cheng, L. (2001). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. Canadian Journal of Applied Linguistics, 4(1–2), 25–40.

Bitchener, J., Cameron, D., & Young S. (2005). The effect of different types of corrective feedback on ESL student writing. Journal of Second Language Writing, 14 (3), 191-205.

Bjork, R. A. (1994). Metacognition: Knowing about knowing. Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. Cambridge, MA: MIT Press.

Broad, Bob, (2003). What We Really Value: Beyond rubrics in teaching and assessing writing. Logan, Utah: Utah State University Press.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), Automated Essay Scoring: A cross disciplinary approach (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates

Chapelle, C. (2008). Utilizing technology in language assessment in E. Shohamy (ed.). Encyclopedia of Language Education (Second edition). Heidelberg: Springer.

Charney, Davida (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. Teaching of English, 18 (1), 65-81.

Crusan, Deborah. (2002). An Assessment of ESL Writing Placement Assessment. Assessing Writing, 17–30.

Deng, X. (2009). The case for writing centres. English Language Teaching World Online. Retrieved from http://blog.nus.edu.sg/eltwo/2009/08/12/the-case-for-writing-centres/html.

Ellis, R. (2001). Investigating form-focused instruction. Language Learning, 51(1), 1-46.

Fairbairn, J., & Dunlea, J., (2017). British Council Aptis Technical Reports: Speaking and Writing Rating Scales Revision. Retrieved from www.britishcouncil.org/aptis

Grimes, D. and M. Warschauer. (2010). Utility in a fallible tool: a multi-site case study of automated writing evaluation. Journal of Technology, Learning, and Assessment 8 (6), 1–43.

Haider, G. (2012). An insight into difficulties faced by Pakistani student writers: Implications for teaching of writing. Journal of Educational and Social Research, 2 (3), 17–27.

Hamp-Lyons, L. (1991). Assessing Second Language Writing in Academic Contexts. Norwood, NJ: Ablex

Hawkey, R. (2009). Examining FCE and CAE: Key Issues and Recurring Themes in Developing the First Certificate in English and Certificate in Advanced English Exams. Cambridge: UCLES/CUP

Hinkel, E. (2002). Second language writers' text. Mahwah, NJ: Lawrence Erlbaum.

Hinkel, E. (2004). Teaching academic ESL writing: Practical techniques in vocabulary and grammar. Mahwah, NJ: Lawrence Erlbaum.

Hirsch, E. D., Jr (1977). The Philosophy of Composition. Chicago: University of Chicago Press.

Hockly, N (2018). Blended learning. ELT Journal, 72(1), 97-101.

Holzknecht, F., Kremmel, B., Konzett, C., Eberharter, K., Konrad, E., & Spöttl, C. (2018). Teacher involvement in high stakes language testing. New York City: Springer.

Hussein, M., Hassan, H. & Nassef, M. (2019). Automated language essay scoring systems: a literature review. Peer J Computer Science 5, e208.

Hyland, K. (2003). Second Language Writing. Cambridge:CUP

Ielts.org. 2020. IELTS Performance For Test Takers 2018. Retrieved from https://www.ielts.org/teaching-and-research/test-taker-performance

Ismail, Noriah & Maulan, Sumarni. (2008). The Impact of Teacher Feedback on ESL Students' Writing Performance. Academic Journal of Social Studies, 8, 45-54.

Jonsson, A., & Svingby, G., (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review 2 (2): 130-144.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527–535.

Kane, M. T. (1998). Choosing between examinee-centered and test-centered standard setting methods. Educational Assessment, 5(3), 129–145.

Kasanga, L. (2004). Students' response to peer and teacher feedback in a first-year writing course. Journal for language teaching. 38(1), 64-100.

Kaur, S., Ganapathy, M. & Kaur Sidhu, G. (2012). Designing Learning Elements Using the Multiliteracies Approach in an ESL Writing Classroom. Southeast Asian Journal of English Language Studies, 18(3), 119-134.

Kim, Youn-Hee, (2009). Exploring rater and task variability in second language oral performance assessment. Frankfurt am Main: Peter Lang Gmbh

Lane, H., Howard, C. & Hapke, H.M. (2019). Natural language processing in action: understanding, analyzing, and generating text with Python. Shelter Island, NY: Manning Publications Co.

Lim, G. S. (2012). Developing and validating a mark scheme for Writing. Cambridge ESOL. Research Notes (49), 6-9.

Liu, J., Xu, Y. & Zhu, Y. (2019). Automated Essay Scoring based on Two-Stage Learning. arXiv:1901.07744 [cs].

Louw, H. (2006). Standardising written feedback on L2 student writing (Unpublished Masters dissertation). North-West University (Potchefstroom Campus). Retrieved from http://scholar.google.co.uk/scholar_url?url=https://repository.nwu.ac.za/bitstream/handle/10394/1718/louw_henk.pdf%3Fsequence%3D1&hl=en&sa=X&scisig=AAGBfm3HcNRB5ePjRAZrhy3h3-dWfblyiw&nossl=1&oi=scholarr

Louw, H. (2011). Mark Write: standardised feedback on ESL student writing via a computerised marking interface (Unpublished PhD thesis). North-West University. Retrieved from https://repository.nwu.ac.za/bitstream/handle/10394/6687/Louw_H.pdf?sequence=1

Myford, C.M, and Wolfe E.W., (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 1 Journal of Applied Measurement, 4 (4), 386-422.

Nesamalar, C., Saratha, S. & Teh, S. (2001). ELT Methodology: Principles and Practice. Selangor: Penerbit Fajar Bakti.

North, B., (2004). Europe's framework promotes language discussion, not directives. In Guardian Weekly. Retrievd from www.guardian.co.uk/education/2004/apr/15/tefl6

Puteh, S. N., Rahamat, R., & Karim, A. A. (2010). Writing in the second language: Support and help needed by the low achievers. Procedia Social and Behavioral Sciences, 7, 580-587.

Raczynski, K. & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? Applied Measurement in Education, 31 (3), 233–240.

Rao, Z. (1997). Training in brainstorming and developing writing skills. ELT Journal, 61 (2).

Razali, R., & Jupri, R. (2014). Exploring teacher written feedback and student revisions on ESL students' writing. IOSR Journal of Humanities and Social Sciences, 19(5), 63-70.

Riddle, S., (2015). Who needs teachers when computers can mark exams? The Conversation. Retrieved from theconversation.com/who-needs-teachers-when-computers-can-mark-exams-41076

Sarkar, D. (2019). Text Analytics with Python: a practical real-world approach to gaining actionable insights from ... your data. Place of publication not identified: APRESS.

Schaefer, E., (2008). Rater Bias Patterns in an EFL writing assessment in Language Testing 25 (4), 465-493.

Shermis, M.D. (2018.) Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring. Applied Measurement in Education, 31 (3), 177–190.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications. TESOL Quarterly, 27, 665-77.

Spencer, B. (1998). Responding to student writing: strategies for a distance-teaching context. (Unpublished thesis). University of South Africa. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.836.1910&rep=rep1&type=pdf

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. Assessing Writing, 19, 51-65.

Suzen, N., Gorban, A., Levesley, J. & Mirkes, E. (2020). Automatic Short Answer Grading and Feedback Using Text Mining Methods. Procedia Computer Science, 169, 726–743

Taylor, L., and Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). Research Notes (24), 2-5.

Thomas, J. (1993). Countering the 'I can't write English' syndrome. TESOL Journal, 2, 12-15.

Wang, Y., Shang, H., Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing, Computer Assisted Language Learning, 26 (3), 234-257.

Wang, A., Singh, A., Michael, J., Hill, F., et al. (2018.) GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Association for Computational Linguistics, 353–355.

Weigle, S. C. (2002). Assessing writing. Cambridge: CUP

West-Smith, P., Buttler, S. & Mayfield, E. (2018). Trustworthy Automated Essay Scoring without Explicit Construct Validity. Retrieved from: https://aaai.org/ocs/index.php/SSS/SSS18/paper/view/17574/15381

Wible, D., Kuo, C., Chien, F., Liu, A., Tsao, N. (2001). A Web-based EFL writing environment: integrating information for learners, teachers, and researchers. Computers and Education, 37(3-4), 297-315.

Zamel, V., (1985). Responding to Student Writing. TESOL Quarterly, 19, 79–101.

Zhen, Y., (2010). Chinese University students' motivation, anxiety, global awareness, linguistic confidence, and English test performance: a causal and correlational investigation (Unpublished Thesis). Queen's University. Retrieved from https://eprints.soton.ac.uk/357330/

Zohrabi, M., & Rezaie, P. (2012). The role of form-focused feedback on developing students' writing skill. Theory and Practice in Language Studies, 2, 1514 –1519.