eWrite

Assessment framework

Australian Council *for* Educational Research

Contents

January 2023

Introduction

ACER's eWrite program is an online assessment that allows teachers to accurately and efficiently measure students' writing abilities, to diagnose gaps, strengths and weaknesses in student learning, and monitor student progress over time. eWrite provides students from years 4 to 10 with eight appropriate and engaging writing prompts across the narrative, report, descriptive, and persuasive genres. Students' writing is automatically scored against the established rubric and detailed reporting provides immediate diagnostic feedback on individuals and groups against the following criteria:

- Orientation and engagement
- Register
- Text structure
- Ideas
- Vocabulary
- Paragraphing
- Sentences
- Sentence punctuation
- Punctuation within sentences
- Spelling

To do this, eWrite draws upon natural language processing, computational linguistics, and machine learning technologies to evaluate student essay responses to the task prompts. eWrite's Automated Essay Scoring (AES) system has been used successfully to assess over a million pieces of Australian student writing over the past decade. It is a system that takes advantage of technologies that allow for greater accuracy and more immediate feedback by mimicking the process human experts use to score essays: they study scored samples or exemplars, learn to apply the rubric, and apply scores consistently over time.

eWrite should be considered as one tool among many that teachers can use to assess and support students to improve their writing. Humans remain uniquely qualified to interpret and assess writing. eWrite brings efficiency and objectivity to the essay-scoring process. The use of automated scoring to supplement and assist teachers' own judgements about students' writing abilities reduces teacher workload and saves time in the collection, analysis and use of data to inform teaching and learning.

More information about eWrite's development, the 'training' of the automated scoring system, and the validity and reliability of the results can be found in Appendix 1.

Rationale for eWrite

The skill of writing is vital to effective communication, critical thinking, and academic success, as well as participation in tertiary education and professional settings. Writing allows students to articulate their thoughts and ideas clearly, enhancing their ability to express themselves and engage with others. It promotes critical thinking by encouraging students to organise their ideas logically, analyse information, and develop coherent arguments. Writing empowers students to communicate persuasively, think critically, and succeed academically, equipping them with a lifelong skill essential for personal and professional growth.

Progressive achievement and learning progressions

An emphasis on learning progression supports the design and use of eWrite to improve students' writing skills. eWrite allows teachers to collect evidence of student learning; to identify where students are in their learning at a given point in time; to monitor growth over time; and to reflect on student attainment.

The value of an integrated approach to assessment and student learning has become widely acknowledged. There is now a wide variety of formative, diagnostic assessment tools used in Australian classrooms. Summative assessments, such as NAPLAN, are also often used to inform teaching and learning. As Dylan Wiliam (2011) makes clear, 'any assessment is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers to make decisions about the next steps in instruction'. In his Report, David Gonski (2018, Finding 7) refers to the compelling evidence that 'tailored teaching based on ongoing formative assessment and feedback is the key to enabling students to progress to higher levels of achievement.' ACER's eWrite assessment provides indicators of student writing achievement via scale scores and accompanying achievement band descriptions. Upon completing their assessments, students are allocated individual criteria scores, as well as a scale score that represents their overall ability in writing. The eWrite scale is divided into achievement bands from which the skills and understanding represented at each level are described. The achievement bands provide valuable evidence-based information about the concepts and skills students have achieved, are consolidating, and are working towards. As the Gonski report recommends, reporting on assessment should have an emphasis on achievement and growth and that growth should be measured against learning progressions (Gonski, 2018, Recommendation 4). Masters (2013) also expresses the idea that learning should be assessed by measuring growth over time and against empirically derived learning progressions. The eWrite reports provide targeted formative feedback, allowing the student data to be sorted and analysed in a variety of ways. Using the eWrite data and the achievement band descriptions, teachers can structure learning specifically to students' needs, rather than where they are expected to be.

Progressive achievement in eWrite

The progression of writing skills is well-documented and begins with children identifying the difference between print and pictures, then moving to an understanding of letters and words as the symbolic representation of language, and on to more complex expressive language use in writing, and then to the more adult skills of reflection and generating written communication for specific purposes and audiences.

eWrite assists in informing teaching by assessing student writing skills across 10 separate criteria. Student performance can be analysed by criterion, so that teachers can identify learning needs and monitor progress at a formative level. The aspects of writing that comprise the marking guide are assessed across four writing genres – Narrative, Persuasive, Report, and Descriptive – that require students to apply different writing styles and formats to engage the reader and effectively address the task.

The eWrite achievement band descriptions are evidence-based, developed from valid and reliable assessment data that have identified a 'typical' trajectory of writing development. This can provide teachers with confidence in the data they are using to target areas of learning, and to identify how students' progress over time.

Construct

Definition

A construct is a description of an ability that can be measured on a single dimension (with a single numeric variable). It often refers to 'what students know and can do'. A mathematical model is used to transform observations (e.g. students' responses to writing prompts) into measurements. A careful definition of ability/proficiency helps ensure that the assessment and reporting are consistent and legitimate.

The skill of writing is vital to effective communication, critical thinking, and academic success, as well as participation in tertiary education and professional settings. Writing allows students to articulate their thoughts and ideas clearly, enhancing their ability to express themselves and engage with others. It promotes critical thinking by encouraging students to organise their ideas logically, analyse information, and develop coherent arguments. Writing empowers students to communicate persuasively, think critically, and succeed academically, equipping them with a lifelong skill essential for personal and professional growth.

eWrite aims to measure the essential skill of writing. The assessment requires students to utilise a variety of processes in a range of contexts and writing styles or genres. The eWrite marking guide or rubric is designed to capture the range of skills competent writers employ in communicating, including technical aspects, such as spelling, grammar, and punctuation, as well as the organisation of the text and the use of language to produce an effective written response to a given prompt.

Structure

eWrite's scoring system is trained to evaluate essays written in response to specific writing prompts covering a number of writing genres. Writing pedagogy in Australian schools is typically genre-based and in primary and secondary schools writing programs sequence and map the range of genres and subgenres that students are expected to master.

The selection of genres for inclusion in eWrite was driven by key syllabus and curriculum documents. Four task types were identified as important in school writing: narrative writing, report writing, persuasive writing, and description writing.

All tasks were developed with open topics so that students would not have to draw on specialised curriculum knowledge and instead could write using their own experiences and knowledge to generate a response.

Persuasive writing: pieces of writing that aim to persuade the reader to believe in a certain point of view (related to the prompt). The following definition is taken from the NAPLAN Persuasive Writing marking guide:

> *The purpose of persuasive writing is to persuade a reader to a point of view on an issue. Persuasive writing may express an opinion, discuss, analyse and evaluate an issue. It may also entertain and inform. The style of persuasive writing may be formal or informal but it requires the writer to adopt a sense of authority on the subject matter and to develop the subject in an ordered, rational way. A writer of a persuasive text may draw on their own personal knowledge and experience or may draw on detailed knowledge of a particular subject or issue*
>
> *(2013 National Assessment Program Literacy and Numeracy Persuasive Writing Marking Guide).*

Narrative writing: pieces of writing that create and narrate events, while also being able to entertain and emotionally move an audience. The following definition is taken from the NAPLAN Narrative Writing Marking Guide:

> *A narrative is a time-ordered text that is used to narrate events and to engage, entertain and emotionally move an audience. Other social purposes of narrative writing may be to inform, to persuade and to socialise. The main structural components of a narrative are the orientation, the complication and the resolution. Essential features of a narrative are the representation and development of character(s) and setting.*
>
> *(2022 National Assessment Program Literacy and Numeracy Narrative Writing Marking Guide).*

Report writing: a report is a specific form of writing, written concisely and clearly and typically organised around identifying and examining issues, events, or findings from a research investigation. It is usually a formal document that elaborates on a topic using facts, charts, and graphs to support its arguments and findings.

Descriptive writing: The primary purpose of descriptive writing is to describe a person, place, or thing in such a way that a picture is formed in the reader's mind, paying close attention to details; teaching students to write more descriptively will improve their writing by making it more interesting and engaging to read.

The eWrite marking guide

The eWrite scoring guide is an analytic, criterion-referenced guide. There are ten criteria, each of which is defined by a skill focus and is elaborated by between three and five scoring categories that represent the continuum of development that underlies the criterion:

Orientation and engagement: the degree to which writing orients and engages the reader

Register: consistently used, appropriate register or tone

Text structure: how well a piece of writing is structured overall

Ideas: the effectiveness of ideas used in writing

Vocabulary: the use of words, including variety and complexity

Paragraphing: the effective use of paragraphs to structure the writing

Sentences: the correctness of types of sentences

Sentence punctuation: the 'marks' used to separate words into sentences

Punctuation within sentences: the 'marks' used to separate letters, words, phrases, and clauses

Spelling: the writer's ability to spell different types of words correctly

The identification of these assessment criteria considered curriculum documents from both Australia and New Zealand. In Australia, the Statements of learning (English) and draft national curriculum (ACARA 2010), were used as a guide to the identification.

In addition to allowing test developers to assess the quality of the prompts, pilot scripts were used to inform the scoring guide development. This method of guide development complies with the principle of ensuring that the scoring guide reflects the demands of the task. Using curriculum documents, test developers identified features of writing that are considered significant in the development of writing skills and then pilot scripts were ranked on each of these features. Once scripts were ranked, shifts in quality could be identified. Where a number of scripts manifested particular qualities, these qualities were described, giving rise to ordered categories and their descriptors.

In this way, ordered categories are created from observations of salient writing features for each task. These features relevant to each task then comprise the assessment criteria and categories of a marking guide. This guide and a selection of the pilot scripts ('exemplars') constitute the method for obtaining consistent, valid, and reliable judgements when scoring student writing.

It is important to note that the Automated Essay Scoring works best when scoring the technical aspects of language. The criteria of 'Orientation and engagement' and 'Ideas' are both more authorial and therefore less technical, than the other criteria, but nonetheless eWrite has a record of assessing these criteria with an acceptable degree of accuracy.

Assessment design

Measuring the construct

eWrite works a lot like the holistic scoring systems commonly employed to score large-scale writing assessments. As part of the development process, eWrite is trained with a set of pre-scored essays on a particular prompt, with known scores assigned by human raters, or scorers, for those same responses. So, similarly to how human raters are trained to score a new prompt, eWrite is given a training set that includes many essays that were previously scored by expert human raters.

In the development phase, scoring for each eWrite task is subject to the same tests as an expert human rater. For instance, after a human rater is trained, they are asked to score a set of essays that have previously been scored by experts. The agreement rate between the new scorer and the 'known' scores is compared. If the human rater meets the criteria for acceptable agreement, the human is allowed to score new essays. Similarly, after eWrite is trained to score responses to a new prompt, it is asked to score a set of essays that were previously scored by experts. Just as in the human scoring process, the agreement between eWrite and the expert scores is evaluated.

After the model is created, the eWrite scores are compared to expert scores on a validation set of essays. It typically takes 300–500 human-scored essays to provide a sufficient training set. If eWrite and the experts agree, the model is ready to be put into use to score new essays submitted to that prompt. If not, the rubric is reviewed, and more essays are gathered to fine tune the process until the results are consistent.

The final eWrite writing tasks were selected after piloting and trialling with over 5000 students. Tasks were scored during this process using an analytical scoring guide for each of the criteria. As eWrite evaluates each scored essay, it learns how the writing features are associated with the score.

In the training set, the scores are also Rasch analysed and a measurement scale calibrated. The scores are then loaded into the online scoring system to create a computer model that replicates hand scoring. The system is designed to 'mimic' the way human scores have been assigned to particular features of writing. In other words, the system infers the writing rubric and the essay features associated with each score.

eWrite analyses the semantic, syntactic, and discourse-related features in a piece of writing. These text-related features are identified as larger categories called Latent Semantic Dimensions (LSD) (Vantage Learning, 2003a), described in five broad categories:

Focus and coherence: analyses the features that emphasise a single point of view, cohesiveness and consistency in purpose, and main ideas in an essay.

Organisation: analyses transitional fluency and logic of discourse. Examples include the introduction and conclusion, coordination and subordination, logical structure, logical transitions, and the sequence of ideas in an essay.

Development and elaboration: examines the breadth of the content and the supporting ideas in an essay (e.g. vocabulary, elaboration, word choice, concepts, and support).

Sentence structure: focuses on sentence complexity and variety such as syntactic variety, sentence complexity, usage, readability, and subject-verb agreement.

Mechanics and conventions: analyses whether the essay includes the conventions of standard English such as grammar, spelling, capitalisation, sentence completeness, and punctuation.

Delivery

Choosing the right test

eWrite provides writing tasks for the following text types:

Narrative	Description	Report	Persuasive
Task B (years 5–8)	Task D (years 5–8)3	Task C (years 5–8)	Task A (years 5–6)
Task H (years 4–8)			Task E (years 7–8)
			Task F (years 5–10)
			Task G (years 5–9)

The suggested year levels for the tasks are based on two pieces of information. Firstly, some prompts may not be easily accessible to younger students, while others may tend to limit the ideas that older students are capable of generating. The suggested year levels are guides to the maturity of thought necessary for creating effective responses to the prompt. Secondly, the trial populations from which the scoring algorithms were developed differed slightly, and the suggested guidelines reflect the majority of students in the trial samples. There is a chance that using the prompts with students outside the suggested year levels may affect the accuracy of the scoring, so it is better to be cautious when using the prompts with students older or younger than the suggested guidelines.

Students interact directly with the online assessment. It is designed for students to complete independently under teacher supervision. All instructions are given on screen and students type plain text straight into the online system. Time for planning and editing is built into the assessment, which scaffolds the students' writing experience.

The test time is 20–25 minutes, which is an optimal time period for such an assessment and in keeping with many other writing assessments conducted by ACER. It can be easily completed within one lesson.

Reporting

For each assessment, the following results are provided:

- a raw score (a total of the individual criteria scores)
- an eWrite scale score
- a band designation
- a breakdown of scores for each assessment criterion

eWrite scale scores are estimates of student ability, as measured by the assessment. The scale scores from eWrite can be used to make direct comparisons between eWrite tasks, which means they can be used to track student progress in writing over time. When comparing scale scores, it should always be borne in mind that every test score has some degree of measurement error. This means that two scores that are close together may not indicate a significant difference.

The individual criterion scores can be used to gauge specific strengths or weaknesses in sub-skills of writing, such as sentence punctuation. These scores in particular can be used as the basis for teachers to provide detailed feedback to students about what they did well in their writing, and what their next steps ought to be in order to improve.

The eWrite scale is divided into bands that cover the range of student writing achievement from year 5 to year 8, as it relates to this assessment. The eWrite scale is divided into seven bands (Band 3 or below to Band 9 or above). The bands were established using student distribution data from the trial phase and the relative difficulty of the categories on the marking guide. They provide a way to aggregate and summarise the performance of a group of students on eWrite assessments. A class teacher might make decisions about differentiation in future writing activities based on band groupings, for example by assigning modified tasks to groups of students who scored in the same band. For example:

> **Achievement band 6 | Scale score 430 to 479**
>
> *Students in this band are able to express ideas with generally accurate use of sentence punctuation and spelling. They provide sufficient elaboration of the main ideas using mostly precise vocabulary. Their text structures reflect the typical structures required of the text type.*

Students in a band are typically able to demonstrate all the skills in lower bands. If a particular criterion is not mentioned in a band's description, teachers should interpret that as the student having demonstrated the ability to the level described in lower bands. For example, sentence punctuation is not described in any band above Band 6, but teachers can safely assume that typical students in Bands 7, 8 and 9 and above will demonstrate accurate sentence punctuation because the Band 6 description includes the ability to 'express ideas with accurate use of sentence punctuation'.

Appendixes

Appendix 1

The development of eWrite, trial design and assessment validity

> *The following information relates specifically to the development of the five original eWrite tasks (Tasks A–E) in 2011. Further development following similar processes resulted in the later addition of three more tasks in 2016 (Tasks F–H).*

Initially, fifteen writing tasks were developed for a small-scale pilot study: four each of narrative, persuasive (argument) and description writing, and three of report writing.

A pilot study is important for two reasons. Firstly, test developers can judge whether tasks are accessible to students, regardless of their age, background, or experience, and secondly, whether the tasks elicit writing across a range of abilities: from weaker writers who struggle to come up with ideas and from more sophisticated writers who have higher level thinking skills and write with depth and flexibility.

Students in years 5–8 from a range of schools participated in the pilot. Each task was completed by all year groups.

Based on qualitative feedback from pilot test administrators, the classroom teachers, and the pilot markers, five tasks were selected for the original trial: two persuasive writing tasks (one more suited to younger students, and one more suited to older students), and one each of a description, narrative and report writing task.

Development of eWrite marking guide

The identification of assessment criteria in the development of the eWrite marking guide considered curriculum documents from both Australia and New Zealand, In Australia, the Statements of learning (English) and draft national curriculum (ACARA 2010), were used as a guide to the identification.

In addition to allowing test developers to assess the quality of the prompts, pilot scripts were used to inform the marking guide development. This method of guide development complies with the principle of ensuring that the marking guide reflects the demands of the task. Using curriculum documents, test developers identified features of writing that are considered significant in the development of writing skills and then pilot scripts were ranked on each of these features. Once scripts were ranked, shifts in quality could be identified. Where a number of scripts manifested particular qualities, these qualities were described, giving rise to ordered categories and their descriptors.

In this way, ordered categories are created from observations of salient writing features for each task. These features relevant to each task then comprise the assessment criteria and categories of a marking guide. This guide and a selection of the pilot scripts ('exemplars') constitute the method for obtaining consistent, valid and reliable judgements when marking student writing.

The connection between the category descriptors and the annotated work samples is therefore necessary and binding. The wording of the categories does not stand alone; just as it was developed from samples of writing, so it is interpreted by the same writing. Thus, the marking guide consists of two equally important components: the marking criteria and their categories; and sample scripts and their commentaries.

The eWrite trial

The eWrite trial was designed to fulfil two main purposes:

1. to provide enough cases for each writing task to develop a common writing measurement scale for all tasks using Rasch analysis. To achieve this the following was required:
    a. at least 250 cases per year group
    b. an equating design that allowed task difficulty to be compared
    c. a marking design that allowed rater-effects to be estimated

2. to provide a set of at least 300 scored cases (the 'training set') for each writing task to 'train' the Vantage Intellimetric® automated scoring system . To achieve this the following was required:
    a. sufficient coverage across each score point of the marking guide, including extreme score points
    b. at least two human ratings per training set script to offset individual human inconsistencies and to provide accurate data for the training process
    c. typed (rather than hand-written) essays

By training the Vantage Intellimetric® automated scoring system, a scoring model capable of automatically scoring unknown scripts of the same task types and topics used in the trial is created.

Trained test administrators in nine independent schools administrated the five selected tasks to students in years 5–8. Writing tasks were presented and students typed their responses.

Responses were then collated and double-marked. Discrepancies in the two sets of human scores were adjudicated. Data from one set of scores were Rasch analysed to determine the quality of the marking guide and to develop the common measurement scale.

Together, the two sets of adjudicated scores and associated scripts (the 'training set') were used to train the Vantage Intellimetric® automated scoring system to mark scripts in a way that reflects the approved, adjudicated human marking. The two sets of adjudicated scores were averaged for the training.

This process of 'matching' computer scoring to human scoring allows the scores generated by the computer marking to be mapped to the measurement scale.

For the trial objectives to be achieved, the trial equating design was required to:

- equate four genres (narrative, persuasive, description, report) to a common scale through common persons equating using a common marking guide
- provide reliable estimates of task difficulties
- produce reliable writing measures for the cohort tested

The trial design involved common persons equating and common item equating. Each student completed two tasks (common persons equating) and each task was completed by three or four groups (common item equating). These mechanisms allow for task difficulty to be assessed within and across year groups. Because the sample was not representative, the relative abilities of each year group, although calculable, were of no consequence to this project.

To control for student fatigue and for class effects, each pair of tasks assigned to each class group was rotated amongst the students.

Trial marking

All markers were trained using the same training material in a face-to-face presentation by the same marker trainer. The training material consisted of the marking guide and a set of exemplars for each task. After training in the use of the guide and before commencing live marking, markers completed a set of practice scripts that had been previously scored and linked to the marking guide via annotations. Exemplars and practice scripts were used to guide consistent application of the marking criteria.

The Intellimetric® automated scoring system

The Intellimetric® automated scoring system from Vantage Learning uses Natural Language Processing, Statistics and Machine Learning understandings (Edelblut & Mikulas, 2005). The system is 'trained' with a set of previously-scored writing scripts (the training set). The training works inductively: the system uses the scores to 'learn' the marking guide and how the human markers have applied their judgements to the scripts that accompany the scores. The system learning is then applied to new scripts of the same topic that have unknown scores.

The system 'trains' through the systematic interaction of the way the set of scores and the features of writing relate to each other and by the accumulation of these relationships, as more scripts from the training set are analysed. The system 'builds' itself as it progressively analyses material from the training set.

Vantage Learning claims that, during training, the system analyses more than 400 semantic, syntactic, and discourse features. These features cover content and structure. Content covers the nature and breadth of concepts and ideas, the support and elaboration of ideas, consistency of purpose, vocabulary, sequencing, transitions and the logical structure of ideas, and the relationship among parts of the response (beginning/end; introduction/conclusion). Structure covers semantic and syntactic features, for example, grammar, spelling, capitalisation, punctuation sentence correctness, completion, complexity and variety, and readability.

Of the scripts and their accompanying scores submitted in the training set, 50 are withheld from the 'training'. A 'scoring model' is created using the remaining scripts and scores. Once the model is created, an exercise to judge the validity of the scoring model is conducted. The 50 withheld scripts are submitted and scored by the model. The computer scores are compared to the human scores and the success of the scoring model can be gauged.

As the training set and the scores originate in Australia and New Zealand, and the system was trained from this set, the automated scoring of new, unknown scripts is based on Australian and New Zealand English language usage and spelling rather than American usage and spelling.

Conclusion

The results of 'the two human markers before adjudication' provide a reference point for judging the results of the 'after adjudication and the automated scoring system' and is the important comparison between these two sets of data. The comparison shows the differences in the extent to which human markers replicate each other's scoring, and the extent to which the automated scoring system replicates the approved, adjudicated human scoring.

The data from 'the two human markers after adjudication' (the training set) shows the magnitude of agreement likely to be achieved when two highly skilled human markers in a controlled environment mark the same scripts. When comparing these results across task types, insights into the task types to which human markers find it easier, and more difficult, to apply the marking guide consistently are evident. The data are indicative of the quality of the training set and one way of viewing them is as an 'aspirant' set of data for the ASS to replicate.

The validation process demonstrated that for all five tasks, across almost all criteria, scoring correlation was strengthened on the 'automated scoring system and training set' compared to 'two humans before adjudication'. The only exceptions were for the Paragraphing criterion, which saw a very small decrease in correlation for Tasks C and D and neutral correlation for Task E.

On all tasks, the rate of agreement on total scores is greater between the automated scoring system and training score sets than between 'two humans before adjudication. This demonstrates that the automated scoring system is more likely to replicate a set of human scores than another human marker is.

Appendix 2

Further reading

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 4-35. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1640/1489

Edelblut, P., & Mikulas, C. (2005). *A comparison of the Accuracy of the Bayesian and Intellimetric® Automated Essay Scoring Methods to Score Essays Written in English.* Vantage Learning.

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167-178. Retrieved from https://www.sciencedirect.com/science/article/pii/S1075293507000037

Masters, G. (2013). Reforming education assessment : imperatives, principles and challenges. *Australian Education Review, 57*, 1-68.

Ramesh, D., & Sanampudi, S. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review, 55*, 2495–2527. doi:https://doi.org/10.1007/s10462-021-10068-2

Rudner, L. M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetricSM Essay Scoring System. *The Journal of Technology, Learning, and Assessment, 4*(4), 3-20.

Vantage Learning. (2003). *Assessing the accuracy of IntelliMetric for scoring a district-wide writing assessment.* Newtown, PA: Vantage Learning. Retrieved from https://www.vantage.com/pdfs/research/RB806.pdf

Wiliam, D. (2011). *Embedded Formative Assessment - practical strategies and tools for K-12 teachers.* Bloomington: Solution Tree.