Journal of Education, 2026
Issue 107, https://ojs.sabinet.co.za/index.php/journal_of_education
http://dx.doi.org/10.17159/2520-9868/i107a07

CC BY NC ND

,

Construct validity in AI-mediated assessment of student teachers' paragraph writing

Monica Mawoyo

Kelello Collectives, NPC, Craighall, Johannesburg, South Africa
monica@kelello.org
https://orcid.org/0009-0001-7598-1638

Nicky Roberts

Curriculum Studies, Education Faculty, University of Stellenbosch, Stellenbosch, South Africa
nicky@kelello.org
https://orcid.org/0000-0002-1910-0162

Ruth Emsley

Department of Languages, Faculty of Humanities,
University of Limpopo, Mankweng, Polokwane, South Africa
ruth.emsley@ul.ac.za
https://orcid.org/0000-0002-6638-871X

Yvonne Reed

University of the Witwatersrand, School of Education
Languages, Literacies and Literatures Division, Johannesburg, South Africa
yvonne.reed@wits.ac.za
https://orcid.org/0000-0002-6157-3449

Mashaba Mashala

Department of Language Education, Arts and Culture, School of Teacher Education
University of South Africa, Pretoria, South Africa
mashajm@unisa.ac.za
https://orcid.org/0000-0003-2587-4903

(Received: 28 April 2026; accepted: 2 August 2026)

Abstract

To teach writing effectively, South African student teachers need to be proficient writers in English themselves. But paragraph-writing tasks used to assess their proficiency can no longer be assumed to measure what they claim to measure with the advent of AI-assisted writing tools. This paper reports on the PrimTEd English Assessment Project, in which student teachers' paragraphs and the accompanying marking rubrics were refined. In Cycle 1, 129 paragraphs, of which three were top-ranked, and three bottom-ranked by the Jumptrak ranking tool using the input of the researchers, informed rubric redesign. In Cycle 2, 539 paragraphs were comparatively ranked and independently marked using a revised rubric that incorporated originality, task engagement, and estimated AI use. Analysis therefore rests on purposefully selected anchor scripts and the study is best read as an illustrative pilot. Expert markers consistently prioritised task adherence, adequacy of development, and perceived authenticity, while their estimates of AI use varied widely and did not align with quality judgements. AI-assisted writing is therefore best understood as a threat to the construct validity of writing assessment rather than merely a detection or academic integrity problem.

Keywords: artificial intelligence, construct validity, design-based research, assessment of writing, initial teacher education, paragraph writing

Introduction

Writing proficiency is foundational to effective literacy teaching, yet a persistent challenge in South African initial teacher education (ITE) is that many student teachers are themselves not sufficiently proficient writers to teach writing effectively. Findings from the Initial Teacher Education Research Project (ITERP, 2012) showed that a substantial proportion of final-year Bachelor of Education (BEd) students demonstrated weak writing proficiency, raising the question: "If teachers of writing are not writers themselves, how can they authentically model, guide, and support learners' writing development?" Responding to this problem depends on being able to assess student teachers' writing proficiency credibly. However, the assumption that a piece of student writing is evidence of the writer's proficiency is now being disrupted by AI-assisted writing.

Large language models have disrupted assumptions underlying conventional paragraph and essay assessment by making it possible for students to produce linguistically polished responses that may not reliably reflect their own writing proficiency. This raises concerns about academic integrity, as well as a deeper assessment problem: whether existing writing tasks and rubrics remain valid for measuring what they intend to assess. The challenge posed by AI is therefore not just one of detecting machine-generated writing as has been the trend in universities when it was established that students were using AI. Detection addresses the symptom. The underlying problem is one of measurement: AI-assisted writing severs the assumed link between the polish of a text and the proficiency of the writer who produced it so that a paragraph may no longer evaluate the writing proficiency that the assessment was designed to support.

This paper takes construct validity as its central conceptual lens. Construct validity concerns whether an assessment task and its scoring criteria elicit and reward the capability they claim

to measure, and therefore whether the inferences drawn from scores are defensible (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Kane, 2013; Messick, 1989). The construct at stake here is the writing proficiency that student teachers require to model, scaffold, and teach writing. AI-assisted writing threatens this construct in two ways. It introduces construct-irrelevant variance, in that marks may be awarded for surface fluency that a language model can readily supply and that says little about the writer's own capability. It also exposes construct underrepresentation, in that prompts and rubrics oriented towards formal correctness may never have assessed the authorial judgement, contextual reasoning, and authentic voice that the construct entails. Fairness and functionality are treated in this paper as consequences of this validity problem rather than as separate concerns standing alongside it. Where scores carry construct-irrelevant variance, the assessment ceases to be fair: students who use AI may be rewarded for a proficiency they have not demonstrated, while students who do not use it are penalised for authentic but imperfect writing. Where the construct is underrepresented, the assessment ceases to be functional: it can no longer serve its diagnostic purpose of showing student teachers and their lecturers which aspects of writing proficiency have been developed, and which have not. Restoring construct validity is therefore the condition on which both fairness and functionality depend.

This paper is situated within the Primary Teacher Education (PrimTEd, 2019) assessment project, which seeks to strengthen assessment of student teacher readiness to teach in relation to its *Knowledge and Practice Standards for Languages and Literacies*. Within the assessments thus far designed, paragraph-writing tasks have been used to assess dimensions of writing proficiency that cannot be captured through multiple-choice formats alone, a format that most PrimTEd assesment items are based on. However, the growing possibility of AI-assisted writing has exposed tensions between rewarding formal language and literacy competence and recognising authentic engagement with a writing task.

Using design-based research (DBR), the paper reports on two iterative cycles of assessment design and redesign. It shows how empirical findings from rubric-based assessment of paragraphs written in response to prompts in a pilot (Cycle 1) informed prompts requiring contextualised personal engagement and revised rubrics foregrounding originality, task engagement, and authentic voice (Cycle 2).

We argue that the challenge posed by AI for assessment of paragraphs and essays is primarily one of restoring construct validity, and that fairness and functionality are restored with it. We propose that prompts requiring contextualised personal engagement, together with rubrics foregrounding authenticity and task engagement, offer a defensible basis for assessing student teacher writing proficiency in the AI-era. The study addresses two research questions:

- How can aspects of student teacher English writing proficiency be assessed fairly and accurately in the context of AI-assisted writing?

- How can paragraph-writing prompts and rubrics be redesigned to strengthen construct validity in the assessment of student teacher English writing, with a view to assessment at scale?

Literature review

The review that follows is organised around the question: "What would a paragraph-writing task and rubric have to do in order to yield valid evidence of the writing proficiency that student teachers need in order to teach writing?" Four bodies of scholarship bear on that question. The literature on teachers as writers establishes the construct, that is, what the assessment is trying to measure. The PrimTEd standards specify the form that construct takes in South African ITE policy. The literature on fairness in writing assessment addresses the conditions under which evidence of that construct can be gathered equitably across a linguistically diverse cohort and identifies construct-irrelevant barriers that compromise it. The literature on AI and writing assessment identifies what has recently changed in the relationship between a written text and the proficiency that text is supposed to model. Construct validity is the thread that holds these bodies of evidence together.

Teachers as writers

The notion that teachers of writing must themselves be writers is well established in the literature. Teachers who teach writing need to be proficient writers, capable of modelling, guiding, and participating authentically in the writing process (Cremin & Myhill, 2020; Cremin & Oliver, 2017). Mendelowitz and Dixon (2016) emphasise the importance of teachers participating authentically in the writing process in South African classroom contexts. This view is reinforced by Mendelowitz et al. (2023) and McKinney (2017), who advocate for socio-cultural awareness of learners' backgrounds and interests as central to language teacher education.

In the Teachers as Writers research project (2015–2017), Cremin et al. (2020) describe teachers who model authentic engagement with writing. These include engaged readers, writing demonstrators, fellow writers, response partners, editors, and publishers of both their own and their students' work. It is possible to assess aspects of student teachers' knowledge about writing and the teaching of writing through multiple-choice questions, as the PrimTEd Assessment EFAL test has done since 2018. However, assessing their proficiency as writers who can enact the roles listed by Cremin et al. requires making judgments based on paragraph or short essay text. Paragraph-based assessment is therefore a necessary component of credible evaluations of student teachers' readiness to teach writing. This study treats writing teacher identity as inseparable from writing practices and argues that strengthening student teachers' writing proficiency is central to teacher education.

Policy context: PrimTEd standards

In the South African context, the PrimTEd (2019) *Knowledge and Practice Standards for Languages and Literacies* provide curriculum guidelines for ITE programmes. These

Standards may be critiqued for what Cremin and Oliver (2017) consider to be a somewhat narrow conception of what counts as writing, yet the expectations of student teachers are made explicit. PrimTEd (2019) English as First Additional Language (EFAL) Standards 3.1 and 3.5 specify that teachers of writing should be able to guide learners in writing and designing texts across multiple modes and genres, and to scaffold the writing process through prompting, demonstrating, modelling, offering feedback, and using structured instructional frameworks. PrimTEd (2019) EFAL Standards 5.3 and 5.6 emphasise that teachers must draw on learners' linguistic repertoires and respond appropriately to home, community, and environmental contexts that influence writing and language learning. These standards frame assessment of paragraph writing as a test of writing proficiency and as a measure of readiness to teach writing authentically and inclusively.

The PrimTEd Assessment Project has developed test items to assess student teachers' knowledge of language (own proficiency), knowledge of language for teaching (subject matter knowledge), and knowledge of how to teach the language (pedagogic content knowledge). While most PrimTEd test items are multiple-choice questions, paragraph writing has been included in the pilot test and in the standards-aligned test based on this pilot, which are the focus of this paper. The assessments are low stakes, intended primarily to provide diagnostic information to student teachers and their lecturers. Assessments are administered in both controlled and invigilated computer laboratory contexts (where AI use is partially mitigated), and in uninvigilated settings, particularly for distance learning students, where AI use is more feasible. These contrasting assessment contexts require paragraph-writing tasks that remain valid even when AI assistance is possible, a challenge that increasingly applies to formative writing tasks in teacher education**.**

Fairness in writing assessment

Fairness in writing assessment is a multidimensional construct that requires assessment tasks and criteria to give all test-takers an equitable opportunity to demonstrate their abilities, regardless of cultural background, linguistic repertoire, or socio-economic context (Hamp-Lyons, 2007). In diverse and multilingual teaching and learning contexts such as those found in South Africa, fairness is particularly important. Viewed through the lens of affirmative ethics, fair assessment recognises students' diverse linguistic and cultural assets as resources rather than deficits.

Fair and just assessment tasks are those that provide all test-takers with an equitable opportunity to demonstrate the intended construct, while minimising constructirrelevant barriers related to language, culture, access, or task ambiguity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Messick, 1989). Writing prompts should be grounded in personal experience, requiring students to draw on first-hand memories and specific personal details that promote authenticity and make generic or fabricated AI responses less viable (Cheng et al., 2024; Shepard, 2000; Wiggins, 1998). Such prompts strengthen construct validity by eliciting situated, meaningful responses rather than rehearsed or externally generated content. Equally important, prompts should be culturally and linguistically inclusive (Dorans & Cook, 2016;

Kunnan, 2008; Weideman, 2012), employ clear instructions (Bachman & Palmer, 2010; Meyer et al., 2014), and require reflection and interpretation and not simple recall, ensuring that difficulty lies in the construct being measured rather than in decoding the task (Kane, 2013).

In the context of generative AI, prompts can build in resistance through personalisation by requiring unique contextual markers, such as local place names, school routines, language-of-instruction practices, or features of the physical landscape, which are more difficult for AI tools to generate authentically or accurately (Cheng et al., 2024; Mislevy et al., 2003). Emerging scholarship suggests that fairness is better preserved through assessment design that foregrounds authenticity, contextual specificity, and meaningful human responses than through unreliable AI detection tools (Cheng et al., 2024; Liang et al., 2023).

The introduction of AI-generated text as a new variable in writing assessment adds complexity to questions of fairness, affecting the design of writing prompts and the criteria used to assess written responses. If rubrics privilege surface features of writing, such as grammatical accuracy, structural coherence, and lexical sophistication, they may inadvertently reward polished AI-generated responses while undervaluing authentic but imperfect human writing. This creates an equity paradox: students who rely on AI may receive inflated scores, while those demonstrating genuine developing competence may be disadvantaged (Cotton et al., 2024). Fairness therefore requires rethinking what writing rubrics value, shifting greater emphasis towards evidence of authenticity, reasoning, voice, and contextualised meaning-making. Generative AI therefore not only challenges assessment integrity but also requires a reconsideration of what constitutes just and valid writing assessment.

AI and writing assessment

The proliferation of large language models and AI writing assistants has significantly complicated writing assessment in higher education (Cotton et al., 2024; Perkins, 2023; Qi et al., 2026). Deepshikha (2025) notes that while AI assessment tools can reduce educator workload and enhance learning experiences, critical challenges persist, including algorithmic bias, data privacy concerns, lack of transparency, and the need for human oversight. Accordingly, the grading of written work must remain a human judgement, with AI supporting rather than replacing professional judgement.

AI-generated text is typically characterised by near-flawless grammar, highly structured and formulaic organisation, generic content that lacks personal anecdotes or cultural specificity, and low "burstiness" (consistent sentence length and rhythm). Human-written text, by contrast, tends to exhibit more variability, idiosyncrasy, and linguistic personalisation. These differences have direct implications for rubric design: criteria that prioritise formal correctness may systematically reward AI-generated writing over authentic student expression.

Students have gradually turned to AI-powered applications like ChatGPT to help them argue more effectively (Male et al., 2025). A study by Li et al. (2023) found that ChatGPT responses were of higher quality across all criteria used to assess students in a pharmacy course, raising concerns about whether existing rubrics can distinguish AI-assisted from human writing. Guo

et al. (2023) suggest that features used to differentiate AI-assisted from human-written text include cultural nuances, use of personal pronouns, and how emotions are expressed in writing.

These claimed textual signatures of AI-generated writing should nonetheless be treated with caution because the literature is far from conclusive. The features commonly listed as diagnostic, namely, fluency, formulaic structure, low burstiness, and the absence of personal anecdotes, are tendencies of particular models under particular prompting conditions and not stable properties of machine-generated text. They can be readily removed by a student who prompts iteratively, edits the output, or writes with a model instead of delegating to it. The evidence on human detection is sobering: Waltzer et al. (2024) found that both students and instructors performed poorly at distinguishing AI-generated from student-written essays and were frequently confident when wrong. Automated detection performs no better in equity terms, with Liang et al. (2023) demonstrating that detectors systematically misclassify the writing of non-native English speakers as machine generated. Giray et al. (2025) argue that detection-led responses reposition assessment as policing at a cost to trust that falls disproportionately on the students least able to absorb a false accusation.

Two implications follow for the present study. The first is that authenticity cannot be read off textual features with any confidence, which means that a marker's judgement of likely AI use is necessarily an uncertainty judgement rather than an observation of the script. The second is that the case for assessment redesign cannot rest on the claim that AI-generated writing is reliably recognisable. It rests instead the more defensible claim that surface fluency has become a substantially less trustworthy proxy for writing proficiency, and that tasks and criteria should therefore seek evidence of the construct that does not depend on that proxy. This is a construct validity argument, and it holds whether any particular script was AI-assisted.

Methodology

DBR as a framework for assessment development

DBR is a methodological framework developed to study learning and educational tools in authentic, complex contexts through iterative cycles of design, implementation, analysis, and redesign (Brown, 1992; Collins, 1992; Design-Based Research Collective, 2003). Unlike traditional experimental research, DBR does not seek to control variables or isolate effects. It embraces the complexity of real educational settings as a productive site for knowledge generation. A defining feature of DBR is its dual output; it produces both practical design products (improved tools, curricula, or interventions) and theoretical insights that generalise beyond the immediate context (McKenney & Reeves, 2018; Reeves, 2006).

DBR was selected for this study because the research problem is concerned with improving assessment in a complex, real-world educational setting. The goal was to produce improved tools for the assessment of student teacher paragraph writing, as well as theoretical principles about fair and functional writing assessment in the AI era (McKenney & Reeves, 2018).

Selection of paragraphs to mark

A central methodological feature of this study was the use of the JumpTrak Comparative Ranking Tool,[1] which facilitates random selection pairs of anonymised student responses for comparative ranking. Students responded online to three paragraph-writing prompts using the JumpTrak platform. Some completed the assessment in an invigilated computer laboratory, while others completed it under online invigilation from home. Although comparing AI use across these settings is beyond the scope of this paper, AI use was a concern in both contexts.

The five-member research team compared randomly selected pairs of anonymised responses using the seven-point comparative scale shown in Table 1. Each response was ranked between three and five times, providing sufficient data for JumpTrak to generate a weighted ranking of all responses from highest to lowest. The highest- and lowest-ranked paragraph for each question was then independently marked by all five researchers using a prompt-specific rubric, with scores averaged across markers. This combined process of comparative ranking followed by independent rubric marking provided a check on the consistency of rubric application while preserving expert human judgement as the basis for selecting the anchor responses.

Table 1
JumpTrak Comparative Ranking Seven-Point Comparative Scale

Scale Point	Comparative Judgement
1	Answer 1 is MUCH better than Answer 2
2	Answer 1 is MODERATELY better than Answer 2
3	Answer 1 is SLIGHTLY better than Answer 2
4	No answer is better than the other
5	Answer 2 is SLIGHTLY better than Answer 1
6	Answer 2 is MODERATELY better than Answer 1
7	Answer 2 is MUCH better than Answer 1

For paragraphs written in response to three prompts (Q1, Q2, and Q3), ranking was performed by five researchers comprising English language lecturers and former lecturers, a PrimTEd English Community of Practice member with a master's degree in English and a PhD in education, and a graduate with an English major. The top- and bottom-ranked paragraphs for each prompt were independently marked by all five researchers, with scores averaged across markers.

1 This tool is commercially referred to as the Jumptrak Peer Marker, which reflects its original design intention. However, in this study, the tool was used by expert lecturers and not by peers. To simply the method description, we refer to it by its functionality as a comparative ranking tool.

In Cycle 1, three paragraphs from $n = 129$ students were ranked, resulting in six paragraphs for independent rubric marking (the highest- and lowest-ranked responses for each of the three questions). These students were first-year BEd students from a public university of technology.

In Cycle 2, three paragraphs from $n = 539$ student teachers were similarly ranked, again producing six paragraphs for independent rubric marking. These students were also first-year BEd students, this time from a private college. The rubric-based analysis in this paper therefore draws on 12 purposefully selected (using comparative ranking) anchor scripts. Each script was independently marked by five markers, thereby providing 60 data points of rubric use.

The DBR cycles

The study proceeded through two iterative DBR cycles, each producing revised assessment artefacts and theoretical propositions about fairness, functionality, and construct alignment. Empirical tensions identified in Cycle 1 informed redesign decisions in Cycle 2. Table 2 summarises these iterative changes and the design principles generated.

Table 2
Summary of iterative design changes across DBR cycles

Dimension	Cycle 1 (Initial Design)	Empirical Tension Identified	Cycle 2 (Redesign Response)	Design Principle Generated
Writing prompts	Prompts requiring students to describe and provide reasons on personal topics	Prompts could enable superficial or potentially AI-assisted responses and did not consistently elicit contextual specificity	Prompts remained the same but assessment criteria required personal, experiential, and context-specific engagement	Prompts should invite authentic engagement and assessment criteria can reduce reliance on generic responses
Rubric criteria	Content and language criteria emphasising description, reasons, and linguistic control	Surface-feature criteria risked over-valuing fluency and under-valuing authenticity and task engagement	Rubric revised to include originality and task engagement alongside formal competence	Criteria should align with the construct of teacher writing readiness
Fairness assumptions	Fairness primarily conceptualised as accessibility and inclusivity	Accessibility alone proved insufficient when technological conditions altered the meaning of task performance	Fairness reconceptualised in relation to construct validity and authenticity	Fairness and functionality are interdependent
Assessment challenge	Concern about weak student writing proficiency	AI-assisted writing exposed tensions in judging authenticity, quality, and ownership of response	Assessment reframed as a construct validity problem rather than detection problem	Assessment tools require iterative redesign under changing conditions

Dimension	Cycle 1 (Initial Design)	Empirical Tension Identified	Cycle 2 (Redesign Response)	Design Principle Generated
Role of judgment	Rubric expected to carry primary evaluative weight	Marker judgments revealed holistic consideration of development, adequacy, and authenticity beyond rubric categories	Comparative ranking and revised criteria positioned human judgment as integral to validity	Human judgment remains central to fair assessment

The redesign reflected a broader reconceptualisation of the assessment problem, from concerns about prompt vulnerability and scoring limitations toward authenticity, construct alignment, and fair judgment. This iterative movement is central to the study's DBR logic.

Cycle 1: Initial problem identification and design

The starting point for Cycle 1 was the ITERP (2012) finding that many final-year BEd students demonstrated insufficient writing proficiency, alongside the PrimTEd Standards' emphasis on teachers *as* writers and teachers *of* writing. In response, three paragraph-writing prompts were designed to be accessible and inclusive across diverse student teacher cohorts: (1) a description of the student's favourite place in South Africa, (2) a description of their number one role model, and (3) a description of their most precious possession. Each prompt required a 130–150-word response.

The prompts were framed around familiar, personal topics to avoid culturally exclusive or class-specific assumptions and to provide equal opportunities for all test-takers to respond. An accompanying rubric assessed content (description and reasons) and language use (vocabulary, grammar, punctuation, and coherence) on a 15-point scale (see Table 3).

Table 3
Rubic for Cycle 1

Content	Content description	5
	Content reason	4
Language	Choice and use of vocabulary	3
	Grammatical accuracy, punctuation, and coherence	3
Total		**15**

At this stage, fairness was primarily conceptualised in terms of accessibility and inclusivity, while the rubric assumed that formal criteria would provide a sufficient basis for judging writing quality. Testing these assumptions in practice generated the empirical tensions leading that informed Cycle 2.

Cycle 2: Analysis, redesign, and construct realignment

During the piloting of Cycle 1, the research team identified frequent AI-assisted writing in student submissions—with many paragraphs being almost identical in terms of topic, vocabulary choices, and sentence structure—together with broader concerns related to how existing prompts and rubric criteria captured evidence of writing proficiency. The issue was not simply the presence of AI-assisted writing, but also that the original assessment design relied heavily on surface-feature criteria, which did not consistently elicit the contextual specificity and task engagement needed for valid judgments of student writing capability.

Cycle 2 began with a systematic review of literature on AI and writing assessment, alongside reflection on findings from Cycle 1. Prompts were not revised as they already reflected topics that could elicit personal, context-specific, and experientially grounded responses. The rubric was redesigned to include originality and task engagement alongside content and language criteria, with attention to marker perceptions of likely AI use treated as analytically relevant to judgments of authenticity (see Table 4). Cycle 2 represented not only a technical revision but a broader re-conceptualisation of the assessment problem as one of construct alignment, prompting redesign to strengthen fairness, functionality, and validity.

Table 4
Rubic for Cycle 2

	Content	**Missing**	**Poor**	**Average**	**Good**
1	Description	0	1	3	5
2	Reasons	0	1	3	5
	Language	**Poor**	**Average**	**Good**	**Excellent**
3	Accurate grammar and punctuation	0	1	2	3
4	Appropriate sentence structure and cohesion	0	2	3	4
5	Choice of vocabulary	0	1	2	3
	Originality	**Definitely Not Used**	**Possibly Used**	**Probably Used**	**Definitely Used**
6	Likely use of AI	0	−1	−2	−4

Data sources and analysis

Data across both DBR cycles were drawn from multiple sources: writing prompt development documents and revision records, rubric drafts and revision histories, and research team reflections on the piloting experience. Content analysis was used to identify patterns in rubric criteria and their relationship to AI susceptibility. Research team reflections informed the

iterative refinement of assessment tools. Formal grading data provided quantitative evidence for the reliability of the revised rubric across the three writing prompts (Q1, Q2, and Q3)

Results

DBR Cycle 1 outputs: Initial design artefacts and emerging limitations

The first DBR cycle produced two initial design artefacts: a set of three paragraph-writing prompts and an accompanying rubric (Table 5) intended to assess student teacher writing proficiency. The prompts were designed to be accessible and inclusive, while the rubric assessed content and language competence on a 15-point scale. At this stage, the design assumed that transparent criteria focused on description, reasoning, and linguistic control would provide a fair basis for judging writing quality.

Table 5
Rubic marking for Cycle 1 (n = 5 markers, assessing six paragraphs selected from n = 129 student teachers)

Question 1	**My Favourite Place**	**Average**	**SD**	**Min.**	**Max.**	**Range**
Top		88%	14%	73%	100%	27%
Bottom		35%	15%	13%	47%	33%
Question 2	**My Number One Role Model**	**Average**	**SD**	**Min.**	**Max.**	**Range**
Top		75%	3%	73%	80%	7%
Bottom		32%	15%	20%	53%	33%
Question 3	**My Most Precious Possession**	**Average**	**SD**	**Min.**	**Max.**	**Range**
Top		98%	3%	93%	100%	7%
Bottom		17%	7%	7%	20%	13%

The average mark across the five markers for the top paragraphs was Question 1 88% (SD = 14%), Question 2 75% (SD = 3%) and Question 3 98% (SD = 3%). For the bottom ranked paragraphs, the average scores were 35% (SD = 15%) for Question 1, 32% (SD = 15%) for Question 2, and 17% (SD = 7%), for Question 3. Marker agreement on rubric scores was generally high with an average standard deviation of 10%.

However, all the markers raised concerns about the rubric and their marking. The emergence of AI-assisted writing exposed limitations in an assessment design that relied heavily on surface-feature criteria and prompts that did not consistently elicit contextual specificity and task engagement. Empirical findings suggested that these criteria risked over-valuing fluency and under-valuing authenticity and adequacy of development, while generic prompts created

opportunities for responses that did not reliably reflect the construct of readiness to teach writing. These findings prompted redesign in Cycle 2 (Table 6).

Table 6
Rubic marking for Cycle 2 (n = 5 markers, assessing 6 paragraphs selected from n = 539 student teachers)

Question 1: My Favourite Place	**Average Across 5 Markers**	**SD**	**Min.**	**Max.**	**Range**	**AI likehood (marker range)**
Top	88%	14%	73%	100%	27%	100%
Bottom	35%	15%	13%	47%	33%	0%-50%
Question 2: My Number One Role Model	**Average Across 5 Markers**	**SD**	**Min.**	**Max.**	**Range**	
Top	75%	3%	73%	80%	7%	25%-100%
Bottom	32%	15%	20%	53%	33%	
Question 3: My Most Precious Possession	**Average Across 5 Markers**	**SD**	**Min.**	**Max.**	**Range**	
Top	98%	3%	93%	100%	7%	0%-50%
Bottom	17%	7%	7%	20%	13%	50%-100%

To explore these concerns further, markers also estimated the likelihood that each paragraph had been AI generated. The top ranked paragraph for Question 1 received an average score of 88% (SD = 14%) and all markers agreed that it was highly likely to be AI generated (100% likelihood). The top paragraph for Question 2 had an average mark of 75% (SD = 3%), and the AI likelihood ranged from 25% to 100%. For Question 3, the top paragraph scored an average mark of 98% (SD = 3%) and again the markers diverged on its AI likelihood (from 0% to 50%).

Considering the bottom ranked paragraphs: Question 1 scored an average of 35% (SD = 15%) and AI likelihood ranged from 0% to 50% amongst the 5 markers; Question 2 scored an average of 32% (SD = 15%), and the five markers were at both extremes: some considering it be no chance of AI likelihood, and others considering it to be definitely AI (100%); and Question 3 scored an average of 17% (SD = 7%), with AI likelihood scores ranging from 50% to 100%.

This variation across markers is itself a finding. The five markers were experienced language educators and practitioners applying the same rubric to the same scripts, and their content and language scores clustered reasonably tightly, with an average standard deviation of 10% across

the six paragraphs. Their estimates of AI likelihood were far more divergent, with estimates for one script spanning the full range 0% to 100%. Two interpretations are possible. Either AI-likelihood judgement is a skill that may improve with training, or—more plausibly, given research on both human and automated detection (Liang et al., 2023; Waltzer et al., 2024)—the textual evidence available to a marker simply underdetermines the judgement. These implications are considered further in the Discussion. The redesigned rubric shifted emphasis from formal correctness towards development, contextual authenticity, and meaningful engagement with the task. This shift reflects the study's argument that fairness and functionality depend on construct alignment.

Comparative analysis: AI-generated versus human-written paragraphs

Analysis of student submissions across both cycles revealed identifiable patterns distinguishing AI-generated from human-written paragraphs. These findings directly informed the DBR redesign process and are summarised in Table 7.

Table 7
Common claimed and contested characteristics of AI-generated and human-written paragraphs

Aspect	AI-Generated Paragraphs	Human-Written Paragraphs
Grammar and syntax	Near flawless, error-free	Often contain errors and inconsistencies before editing
Structure and coherence	Highly structured, formulaic	More varied, sometimes loosely structured
Creativity and personal voice	Limited, generic, lacks anecdotes	Rich in personal insights, unique style
Grading tendencies	AI grading often more generous	Human grading considers deeper originality
Detection clues	Repetitive phrasing, predictability	Higher burstiness and linguistic variation
Cultural nuance	Biased toward hegemonic cultures	More personalisation; recruits from cultural expressions

Table 8 summarises the key differences between the top- and bottom-ranked paragraphs across the three writing tasks. Scoring was out of a range from 3–5 for each dimension.

Table 8
Top- and bottom-ranking patterns

Question	Top-Ranked Passage	Bottom-Ranked Passage	Key Ranking Pattern	AI-Likelihood Pattern
Q1 Favourite place	High description (4.6), high reasons (4.6), language (2.8)	Very low description (0.6), reasons (0.8), language (0.3)	Markers rewarded vivid elaboration and penalised extreme brevity (23 words)	Top ranked attracted higher suspicion (3.2) than bottom (0.4)

Question	Top-Ranked Passage	Bottom-Ranked Passage	Key Ranking Pattern	AI-Likelihood Pattern
Q2 Role model	Strong description (3.8), reasons (4.0), language (2.5)	Lower description (2.2), reasons (3.0), language (1.3)	Markers rewarded personal specificity and task alignment over abstract polish	Bottom ranked attracted higher suspicion (2.6) than top (2.0)
Q3 Precious possession	Moderate description (2.6), strong reasons (4.4), language (2.2)	Lower description (1.8), Very weak reasons (0.6), language (1.8)	Markers rewarded developed explanation and penalised minimal response (29 words)	Bottom ranked attracted much higher suspicion (3.6) than top (0.4)

Across all three tasks, markers consistently prioritised substantive fulfilment of the task, adequacy of development, and perceived authenticity over linguistic polish alone. The strongest responses were rewarded because they aligned closely with prompt expectations, provided sufficient elaboration, and conveyed credible engagement with the topic. By contrast, the lowest-ranked responses were consistently constrained by underdevelopment, particularly lack of description and limited justification, which sharply reduced their writers' capacity to meet task demands.

Task-specific differences also emerged. In Question 1, vivid descriptive elaboration appeared most strongly associated with ranking while in Question 2, personal specificity and credible human-like voice appeared to be particularly valued. In Question 3, emotional depth and developed reasoning were more salient than descriptive richness. Across all tasks, markers consistently penalised responses that lacked sufficient development, regardless of whether those responses were perceived as human-written or associated with AI-like features.

The AI-likelihood ratings further complicate simple assumptions about AI and quality. Higher AI suspicion did not consistently align with higher scores, nor did suspect AI-associated responses uniformly outperform human-like ones. In Question 1, the top-ranked response attracted relatively high AI suspicion, while in Questions 2 and 3 bottom-ranked responses were associated with higher AI-likelihood ratings. These patterns suggest that originality did not operate as a stand-alone discriminator, and that marker judgments were shaped more strongly by adequacy of response, authenticity, and construct-relevant task engagement than by linguistic polish alone. These findings informed the redesign of the prompts and rubric in Cycle 2.

Design principles emerging from the DBR process

Consistent with DBR's dual output mandate, the two cycles produced revised assessment artefacts and a set of transferable design principles for fair and functional writing assessment in the AI era. These principles are:

- Writing prompts should require personal, experiential, and contextually specific responses that encourage authentic engagement and reduce reliance on generic AI-generated responses.
- Rubrics should include authenticity-focused criteria: originality, personal voice, and task engagement, alongside formal competence criteria.
- Rubric criteria should align with the intended construct of teacher readiness to teach writing, rather than surface linguistic polish alone.
- Assessment tools are dynamic design artefacts requiring ongoing revision in response to the educational and technological landscape.
- Linguistic polish should not be conflated with writing quality; assessment frameworks should value clarity, development, connection, and contextual authenticity, recognising that effective writing for teaching is not synonymous with polished perfection.

Discussion

The findings suggest that AI destabilises the relationship between the assessment task and the construct it is intended to measure. Viewed through a construct validity lens, the central issue is whether prompts and rubrics continue to capture evidence of the writing proficiency needed for teaching writing, rather than relying on surface textual features that may function as weak proxies for that capability. The DBR process demonstrated how these tensions can serve as empirical signals for redesign.

We suggest that fairness and functionality are not separate concerns, but interdependent conditions of valid assessment. A rubric may appear procedurally fair yet still be unfair if those criteria no longer represent the construct of interest. Likewise, one that overvalues fluency or undervalues authentic but imperfect writing risks distorting judgments of student capability. Fairness, in this sense, depends on construct alignment.

The comparative ranking findings reinforce this interpretation. Across the three tasks, markers did not consistently reward AI-like linguistic polish, nor did AI-likelihood ratings align straightforwardly with quality judgments. Instead, markers appeared to prioritise adequacy of development, substantive fulfilment of the task, and perceived authenticity. Responses were penalised most consistently when they were underdeveloped, especially where brief outputs limited the scope for elaboration, regardless of whether they were associated with human or suspected AI authorship. Authenticity, task engagement, and writing quality appear to operate as interdependent constructs requiring integrated treatment in assessment design.

This raises a difficulty for the redesigned rubric. The Cycle 2 rubric attaches a mark penalty to perceived AI use, while the argument of this paper is that AI compromises validity by introducing construct-irrelevant variance into scores. A penalty applied based on a marker's suspicion is itself a source of construct-irrelevant variance, and the AI-likelihood ratings reported above give reason to take that possibility seriously. Markers disagreed sharply, and no ground truth was available. Two students who wrote unaided might therefore receive different marks depending on which marker happened to read them, and the students most

exposed to error are those whose writing resembles the fluent, formulaic register that markers associate with AI, or, as Liang et al. (2023) show for automated detectors, those writing in an additional language. A criterion introduced in the name of fairness could in this way reproduce the inequity it intended to prevent. The evidence presented here suggests that the originality criterion should function as a prompt for markers to look for construct-relevant engagement, namely, personal, contextual, and developed response, rather than as a detection instrument with a penalty attached. Whether an explicit AI-use penalty can be justified at all is a question this study raises but does not resolve, and it is a matter to take up for the next design cycle.

A further implication concerns the continuing role of human judgment. The findings suggest that revised criteria alone do not guarantee fair assessment; lecturer expertise and professional judgment remain integral to making valid inferences about student writing. Marker perceptions of authenticity, adequacy, and task engagement were analytically significant in this study, indicating that human judgment should be understood as part of the validity argument. Technology awareness and professional judgment are therefore complementary dimensions of fair assessment.

Strengths, limitations, and recommendations

A strength of the study lies in its combination of iterative design, empirical analysis, and theoretical development through a DBR approach. The comparative ranking of 668 scripts across two cycles, together with independent scoring by a five-member expert panel, provides a stronger empirical basis than a single-site or single-cycle study would allow. The study also benefits from its location within the PrimTEd Assessment Project, which gives the findings relevance beyond a single institutional context.

Several limitations should nonetheless be noted. The study reports on an early-stage pilot, and the revised rubric has not yet been validated at full implementation scale. AI-likelihood judgments in this study were treated as indicators of marker perceptions, which were analytically relevant because the study examined how such perceptions may shape assessment judgments. Markers were not trained before making these judgments, and no ground truth about actual AI use was available for any script, so these estimates should not be read as measurements of AI use. The rubric-based analysis also rests on 12 anchor scripts, which is a narrow basis from which to generalise about rater behaviour. Finally, although the design principles are presented as transferable, their applicability beyond this South African ITE context requires empirical testing. Future research should investigate the validity and reliability of the revised rubric at scale and explore how peer ranking and expert judgment might best be combined in large multilingual assessment settings.

Conclusion

We have argued that fair and functional assessment of writing in the AI-era requires more than detecting AI or penalising linguistic polish; it requires rethinking what counts as evidence of writing proficiency in ways aligned to the purposes of teacher education. Drawing on two DBR cycles within the PrimTEd Assessment Project, the study offers promising but preliminary

evidence that prompts grounded in lived experience and contextual engagement, together with rubrics that value originality, task engagement, and formal competence, can strengthen the alignment between the assessment and the construct it is intended to measure. It is not claimed that the revised rubric produced fairer or more valid outcomes; it has not yet been validated at implementation scale, and the evidence presented here is based on a small number of anchor scripts. What is claimed is that it is better aligned with the construct, and that such alignment is the condition on which fairer and more functional assessment depends.

The central finding is not that originality alone distinguishes AI from human writing, but that fairness and functionality depend on aligning assessment with the construct being measured: the capacity of student teachers to write sufficiently well to teach writing. This shifts attention away from perfection or detection as ends in themselves toward judging the quality of response, depth of engagement, and readiness to model writing for learners.

The principal contribution of this study is conceptual. It reframes AI-assisted writing as a problem of construct validity rather than of detection or academic integrity and demonstrates how DBR can be used to redesign assessment in response to technological disruption. Practically, it offers a revised rubric and a set of transferable design principles as artefacts for further testing rather than validated solutions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Cheng, Y., Lyons, K., Chen, G., Gašević, D., & Swiecki, Z. (2024). Evidence-centered assessment for writing with generative AI. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (pp. 178–188). Association for Computing Machinery. https://doi.org/10.1145/3636555.3636866

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). Springer.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Cremin, T., Myhill, D., Eyres, I., Wilson, A., Oliver, L., & Nash. T. (2020). Teachers as writers: Learning together with others. *Literacy.54(3), 49–59.* https://doi.org/10.1111/lit.12201

Cremin, T., & Oliver, L. (2017). Teachers as writers: A systematic review. *Research Papers in Education, 32*(3), 269–295. https://doi.org/10.1080/02671522.2016.1187664

Deepshikha, D. (2025). A comprehensive review of AI-powered grading and tailored feedback in universities. *Discover Artificial Intelligence, 5*(1), Article 251. https://doi.org/10.1007/s44163-025-00251-7

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8. https://doi.org/10.3102/0013189X032001005

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge.

Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college writing. *Internet Reference Services Quarterly, 29*(1), 83–116. https://doi.org/10.1080/10875301.2024.2437174

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*. *arXiv*. https://doi.org/10.48550/arXiv.2301.07597

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 487–504). Springer.

Initial Teacher Education Research Project. (2012). *An investigation into aspects of South African initial teacher education* [Research report]. JET Education Services.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kunnan, A. J. (2008). Towards a model of test fairness. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment* (pp. 229–251). Cambridge University Press.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively? *Computers and Education: Artificial Intelligence, 4*, Article 100140. https://doi.org/10.1016/j.caeai.2023.100140

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Male, H., Timu, T. M., & Sabrina, D. (2025). The role of artificial intelligence (AI) in essay writing. *SIGEH ELT: Journal of Literature and Linguistics, 5*(1), 476–493. https://doi.org/10.36269/sigeh.v5i1.3115

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research* (2nd ed.). Routledge.

McKinney, C. (2017). *Language and power in post-colonial schooling: Ideologies in practice*. Routledge.

Mendelowitz, B., & Dixon, K. (2016). Transgressing the boundaries: Positioning student writing in South African higher education. *Journal of Writing Research, 8*(1), 93–127. https://doi.org/10.17239/jowr-2016.08.01.04

Mendelowitz, B., Ferreira, A., & Dixon, K. (2023). *Teaching language and literacy in South African primary schools: A sociocultural approach*. van Schaik.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. CAST.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report RR-03-16). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2), Article 7. https://doi.org/10.53761/1.20.02.07

Primary Teacher Education Project. (2019). *Knowledge and practice standards for primary teacher education graduates: Language and literacy* [Draft]. Department of Higher Education and Training. https://www.jet.org.za/projects/all-projects/primted/resources/language-and-literacy-resources-repository/language-and-literacy-teacher-standards-draft-v2-201810.pdf

Qi, J., Tuxworth, J., & Gomes, C. (2026). Navigating supervisory conversations about Gen AI: Taboo and supervisory power dynamics. *Higher Education Quarterly*. Advance online publication. https://doi.org/10.1111/hequ.70151

Reeves, T. C. (2006). Design research from a technology perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 52–66). Routledge.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity, 20*, Article 11. https://doi.org/10.1007/s40979-024-00158-3

Weideman, A. (2012). Validation and validity beyond Messick. *Per Linguam, 28*(2), 1–14. https://doi.org/10.5785/28-2-524

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.