Thinking Skills and Creativity 42 (2021) 100949

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc

What characterises creativity in narrative writing, and how do we assess it? Research findings from a systematic literature search

Richard D'Souza [*]

University of Exeter, Graduate School of Education, Stocker Rd, Exeter, Devon, EX4 4PY, UK

ARTICLE INFO

Keywords:
Creativity
Writing
Assessment
Narrative
Review

ABSTRACT

This paper reports findings from a systematic search of the empirical literature from 2000 to 2020 on the assessment of creativity in narrative writing. It seeks to synthesise the designs, methods and findings on how different disciplines have gathered relevant data to the question of how creativity in writing might be assessed, and feedback made more effective. It draws together the established knowledge base around two research questions. The methodology for the systematic involved searches on five academic databases for relevant keywords, producing 1796 papers. Initial screening of the abstracts identified 97 studies for further scrutiny, and for which full texts were accessed for secondary screening based on the inclusion criteria. The final 39 papers judged to satisfy the selection criteria were subject to in-depth analysis and synthesis. The findings of the review reveal that four main techniques have been utilised in efforts to assess creativity in writing, each with their own merits and limitations, and a paucity of research in several crucial areas. The review indicates that, while several disciplines have contributed to the knowledge base in this area, few interdisciplinary studies exist that draw together multiple techniques and provide clear answers for the research questions used in the study. Furthermore, there is little empirical evidence suggesting that assessment improves student creativity with regard to writing, and new research in the field would be advanced by addressing explicit definitions of creativity, the practices of writing 'experts', and writing considered within its social and cultural context.

1. Introduction

Assessment in education has been identified as a potentially powerful lever for enhancing practice (Hattie & Timperley, 2007; Wiliam & Leahy, 2015; Wisniewski, Zierer & Hattie, 2020), yet it is also a key area that requires development when it comes to articulating a pedagogy for creativity (Craft, 2005). Thus, there is currently a lack of clear, widely adopted advice on what constitutes 'best practice' in creativity assessment for education. In research, the challenge of assessing creativity has been described as 'significant, yet complicated' (Turkman & Runco, 2019); methodologically, a number of dilemmas are created for the creativity scholar – should the research focus upon the creative individual, their approach or creative process, the product produced (see Sternberg, 1988 for an exploration of *process* and *product* creativity), or how it is perceived and received by others? How far should the influences behind the product, and its reception, be understood in relation to the sociocultural context in which the act of creativity took place? Should the research address 'Big-C' (eminent creativity, with high cultural value) or the more every day, 'little-c' creativity that all humans can demonstrate (Craft, 2000; Kaufman & Beghetto, 2009)? Furthermore, what methods might adequately capture or measure these elements of creativity?

[*] Corresponding author at: 94A Leckhampton Road, Cheltenham, Gloucestershire, GL53 0BN UK.

https://doi.org/10.1016/j.tsc.2021.100949
Received 23 July 2021; Received in revised form 23 August 2021; Accepted 7 September 2021
Available online 12 September 2021

While the craft of writing may be the most accessible of the cultural domains in which creativity is studied (Csiksentmihalyi, 1997), research has understood this in different ways. Cremin and Myhill (2012) discuss discrepancies between 'creative writing', a more genre-bound term largely restricted to arts-based, school and university writing curricula, and 'creativity in writing', a more inclusive approach acknowledging the creativity of most writing and linking to broader definitions of creativity, considering social and cultural elements. There remains a lack of clarity in the definition of creativity with reference to creative writing (Turkman & Runco, 2019); it is also interdisciplinary, involving cognitive, social, personality, and biological aspects of psychology (Sternberg, 2009). Academics from the fields of linguistics and educational research have also contributed their insights. Language is intrinsically generative in nature (Chomsky, 1965), and perhaps the best example of 'everyday' creativity (Runco, 2014) because of the immense repertoire of possibilities it affords. Writing has, therefore, been described as a creative meaning-making process achieved through the active and continued involvement of the writer with the unfolding text (Emig, 1988). Writers can be understood as creative thinkers, problem-solvers and designers who are juggling with task constraints and drawing upon the creative and linguistic resources available (Sharples, 1999). Some research (see for example Dymoke, 2003; Wilson, 2010) has addressed pedagogies for poetry assessment in the light of this perspective for 'little-c' creativity choices in writing, and thus this paper seeks to contribute insights with regard to the narrative genres. This paper seeks to synthesise the body of research produced in recent years and outline the merits and limitations within the adopted methodologies of the studies. The review is directed by two overarching research questions:

1 *What is known of the characteristics of narrative writing judged to exhibit creativity?*
2 *What is known about the 'best practices' for assessing the creativity of writing in educational settings?*

2. Methods

2.1. Literature search strategy

The research questions outlined above necessitated a precise and systematic approach to reviewing the relevant empirical literature, and to narrow the focus of the works reviewed to those that specifically addressed the issue of assessing creativity in narrative writing. The review follows a broadly configurative synthesis logic (Sandelowski, Voils, Leeman & Crandell, 2012), characterised by deliberative construction of the data into patterns to create a richer conceptual understanding of the phenomenon (Newman & Gough, 2020) – in this case, the methods researchers have used to assess creativity in writing. Searches were run on five databases in November 2020 to identify studies relating to the assessment of creativity in writing: AEI (Australian Education Index), BEI (British Education Index), ERC (Education Research Complete), ERIC (Educational Resources Information Centre), and WOS (Web of Science). The search, while international in scope, was restricted to peer-reviewed studies published between 2000 and 2020. The key search terms used were *assessment* and variants (assess* OR mark OR marking OR grade OR grading OR feedback), *creativity* (creativ* OR narrat*), and *writing* (writ*). Initial searches generated too many studies unrelated to creativity in writing and for which assessment was not a focus for the study; a proximity operator (N5) was therefore added between the terms *creativity* and *writing* to ensure that these two terms were loosely connected in the search results. The final search necessitated that the core term *assessment* and its variants appeared in the abstracts of papers, revealing a total of 1744 results which were then downloaded to EndNote X9 before duplicates were removed. A summary of the results generated by the searches is given in Table 1.

A further 21 records were added through separate hand searches and consultation with other academics in the field. Initial screening removed unrelated and non-empirical studies, studies that did not involve writing, did not mention 'creativity', 'narrative' or their variants, had been published prior to the time frames specified, or did not address an appropriate demographic (relevant articles included participants aged 5 and above). A total of 245 articles were eventually retained, organized into three groups. A bank of 148 articles did not deal explicitly with the issue of assessing creativity in writing, but nonetheless applied some way of measuring progress or attainment in writing - for example, intervention studies that sought to show an improvement based on a particular input – served as contextual background literature providing a backdrop on how writing performance and quality had been measured in research more generally. The full texts of 97 potentially relevant articles were accessed (51 that addressed the education of children aged 5–15; and 46 that focused on post-16 and professional education). These articles were then subjected to further screening based on the selection criteria for the review.

Table 1
Electronic search results.

Boolean search terms	AEI	BEI	ERC	ERIC	WOS	Total
(assess* OR mark OR marking OR grade OR grading OR feedback) **AND** (creativ* OR narrat*) **AND** (writ*)	105	206	1680	1361	2098	5450
(assess* OR mark OR marking OR grade OR grading OR feedback) **AND** (creativ* OR narrat*) N5 (writ*)	47	89	1247	533	213	2129
(AB = (assess* OR mark OR marking OR grade OR grading OR feedback) **AND** (creativ* OR narrat*) N5 (writ*)	**41**	**73**	**1014**	**417**	**199**	**1744**

2.2. Selection criteria

The inclusion criteria required that studies focused on the issue of assessing the creativity of narrative writing genres rather than poetry or nonfiction text types, and were peer-reviewed reports of empirical investigations. Studies were excluded if they were not published in English; were EFL-focused; did not address narrative, or were not empirical investigations. A total of 39 studies met all of the inclusion criteria (25 for the age range 5–15; and 14 for post-16 and professional education). In studies that involved both teachers and their students (e.g. Cremin et al., 2017, in which both teacher and student voices are studied) the participants for whom findings were most pertinent to the research questions (in this case, the adult participants) determined which age range the file was added to. A PRISMA flow diagram (Moher, Liberati, Tetzlaff & Altman, 2009) outlining the literature search is given in Fig. 1.

2.3. Data extraction and synthesis

This review focuses on the assessment measures utilised by researchers to draw out their findings, rather than the known effects of interventions. An inclusive approach was therefore adopted, whereby the details of all the included studies were tabulated and compared prior to the synthesis of findings to reveal the variety of different approaches to assessing creativity in writing as well as gaps in the existing relevant literature. The review templates recorded key information about the sample, location, duration, key findings, text genre, and methodology used in the studies. Key themes were discussed with a senior researcher in the field of writing research, and the quality of the studies was reviewed with reference to the Critical Appraisal Skills Programme (CASP) quality checklists.

2.4. Characteristics of the field

Relevant studies were undertaken in the USA (16), UK (4), Canada (3), China (3), Spain (3), Australia (2), Turkey (2), Bangladesh (1), Mexico (1), Holland (1), Germany (1), Switzerland (1), and Iran (1). A range of writing served as the data gathered for analysis in

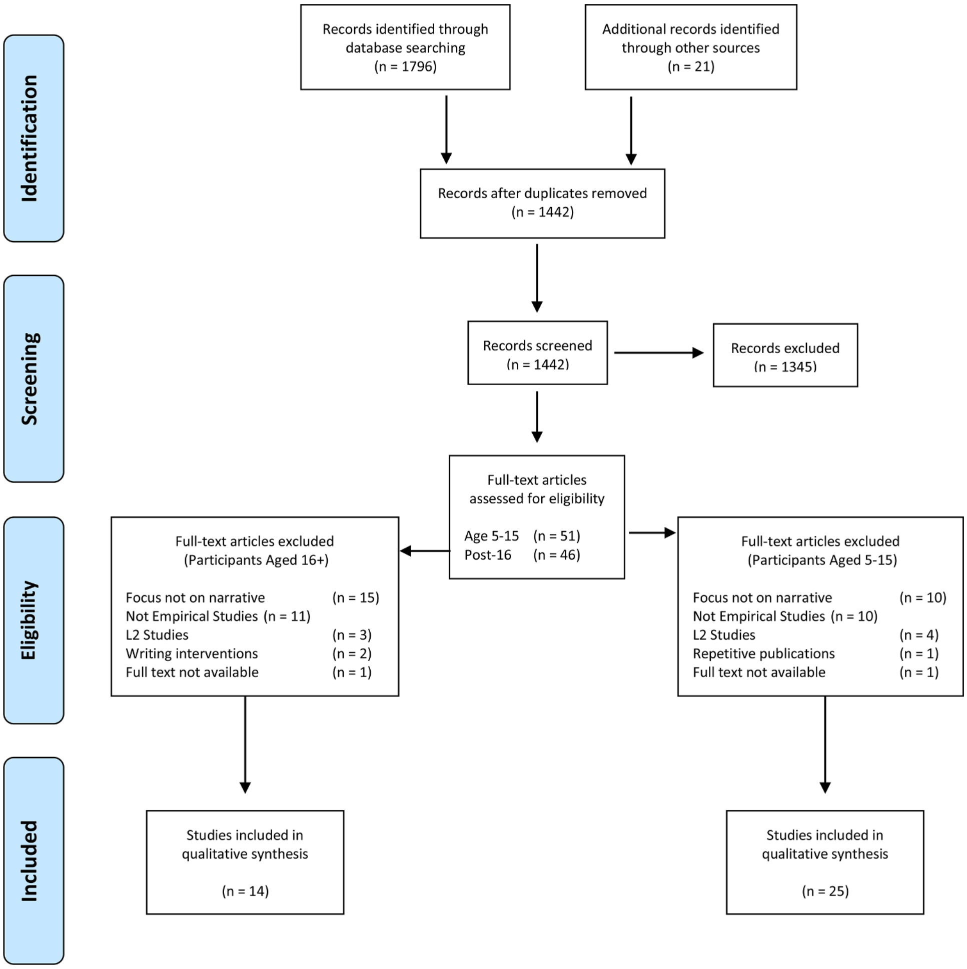
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1796)
Additional records identified through other sources (n = 21)
Records after duplicates removed (n = 1442)
Records screened (n = 1442)
Records excluded (n = 1345)
Full-text articles assessed for eligibility
Age 5-15 (n = 51)
Post-16 (n = 46)
Full-text articles excluded (Participants Aged 16+)
Focus not on narrative (n = 15)
Not Empirical Studies (n = 11)
L2 Studies (n = 3)
Writing interventions (n = 2)
Full text not available (n = 1)
Full-text articles excluded (Participants Aged 5-15)
Focus not on narrative (n = 10)
Not Empirical Studies (n = 10)
L2 Studies (n = 4)
Repetitive publications (n = 1)
Full text not available (n = 1)
Studies included in qualitative synthesis (n = 14)
Studies included in qualitative synthesis (n = 25)

Fig. 1. PRISMA flow chart of study selection.

the studies, including texts written by children, university students, and pre-service teachers, as well as published books and texts about writing. Participants and sample sizes were varied due to the number of studies that used multiple groups to rank or rate the texts. These included participants in the education sector (teachers, creative writing instructors, examiners, and students), in academia (researchers, psychologists), and those in the field of commercial publication (authors, agents, and editors). Review templates summarising the characteristics of the included studies is included as an appendix to this paper.

The methodological approaches adopted by the included studies varied from small-scale case studies and action research reports (3) to larger quantitative datasets analysed through statistical testing (27). The most frequently used research designs employed quantitative data for a nonexperimental measurement, such as studies that employed the Consensual Assessment Technique (CAT) – whereby experts rated writing samples and inter-rater reliability was tested for – or a survey. Twenty studies fell into this category, while a further seven adopted an experimental or quasi-experimental approach. Some studies tested the validity of rubrics, while others adopted a more experimental design to explore the effect of a rubric or change in curriculum emphasis. Four studies analysed qualitative data using grounded theory, and four used an explanatory sequential mixed methods approach to enrich quantitative data with qualitative investigation.

3. Findings

3.1. The criteria identified in narrative writing judged to exhibit creativity

The notion of writing as creative design suggests that just about all writing is creative, requiring the recognition and utility of the infinite possibilities within a language to creatively align the writers' knowledge of language, text and audience (Cremin & Myhill, 2012; Sharples, 1999). One aim of this paper is to draw out knowledge specifically related to the features of narrative writing that was judged to exhibit such creativity, distinct from generic appraisals of classroom writing. 24 of the studies included rubrics or established measures of narrative or creative writing to aid in reaching their conclusions: some used generic rubrics for narrative writing designed to comprehensively assess student work, while others utilised explicit definitions of creativity to design their own measures. It is this latter group that constitutes the focus of this section: six studies that explicitly addressed the creativity of the writing and sought to wrestle with evidence of creative elements in the writing, moving beyond studies that employed more generic, sometimes vague writing criteria to assess writing performance; instead, this approach attempts to draw together features from the writing to contribute to a knowledge base for creativity theory. Of the six studies that address this research question, two are drawn from the pool that addressed participants aged 5–15 (Cheung, Tse & Tsang, 2001; Kettler & Bower, 2017) and four are drawn from the post-16 education and professional writing pool of studies (Ashton & Davies, 2015; Form, 2019; Furst, Ghisletta, & Lubart, 2017; Zedelius, Mills & Schooler, 2019).

3.1.1. Originality and quality in writing

Of the six studies that looked closely at the creativity of narrative writing, one looked at the creativity in writing process while the other five addressed creative products. Furst, Ghisletta, & Lubart, 2017 conceived of the writing process as one that entailed a mixture of two sub processes, Generation (ideation, playing with ideas, thinking of possibilities) and Selection (critique, evaluation, and refinement), and sought to understand the roles these played throughout producing a piece of writing. The experiment tasked ten groups of participants with producing a narrative as instructed for the different phases of writing – for example, to avoid 'selection' behaviours, such as critically reviewing the work, during planning or drafting and to instead maximise 'generation' behaviours throughout the process. Four phases of writing (preparation, drafting, development and correction) scaffolded the different intensities of generation and selection for the different experimental groups. One group served as a control, and produced the text following their own process without interference or instruction. The experiment necessitated a rationale and justification for the definition of creativity used in the research:

"Generally speaking, creativity is defined as the confluence of both originality and appropriateness (e.g., Sternberg & Lubart, 1995) so our evaluation is based on these two dimensions. However, in the context of our writing task, we preferred the label of *quality* to appropriateness or usefulness, which may not be very relevant when referring to a fiction text […] we conceptualized a creative text as one that is both original and of high quality." (Furst, Ghisletta, & Lubart, 2017. 205)

Research assistants and student peers rated the texts 1–5 for quality (defined broadly by the researchers as constituting technical correctness and notions of style, coherence and aesthetic value) and creativity (involving surprising, unexpected, and original elements). The study found that the more creative works were produced with a high degree of 'Generation' or ideation at the start of the writing process, followed by increasing or high levels of criticality, or 'Selection'. This highlights the importance of the writer's rigour regarding their work after an initial spurt of generative thinking. However, none of the experimental groups outperformed the control group, who wrote normally. This suggests that those who responded to the stimulus without attempting to control the process demonstrated more creativity - insofar as they produced texts that were both original and of higher quality, by the authors' definition of creativity - than those who attempted to change or manipulate their writing processes as directed for the purposes of the study.

3.1.2. Examining rubric criteria for creativity

The remaining five studies examined creative products to reach their findings, with four of these producing rubrics to summarise them. Form (2019) explored the extent to which linguistic originality accounted for the popularity of bestselling books, while the remaining studies (Ashton & Davies, 2015; Cheung et al., 2001; Kettler & Bower, 2017; Zedelius et al., 2019) sought to capture a more holistic set of features for the creative texts used. This section of the literature review will trace the lineage of criteria used in these studies to address the data that generated them.

The Chinese Creative Writing Scale developed by (Cheung et al., 2001) is reportedly derived from the (Carlson Originality Scale, 1965), with additional input from scales by (Torrance, 1965) and (Guilford, 1967). Such instruments have helped establish the notion of 'divergent thinking' in creativity: the scale developed by (Carlson, 1965), for example, draws on iterative analysis of 5000 children's stories, with both the definition of creativity used and process of refining the scale described in the paper's methods section. Similarly, the work of (Torrance, 1965) draws on three sets of studies involving 20,000 school children (Gowan, 1965).

Cheung et al. (2001) use these works to create a new scale for use with their sample in Hong Kong, noting cultural differences between Chinese and the West alongside grammatical and semantic structural contrasts. While the notion of exploring creative acts within their social and cultural context is central to creativity theory (Cremin & Myhill, 2012) in this paper there is little discussion of how these differences were resolved. Leong (2010, 2011), for example, considers the question of creativity in the East as contrasted to the West, and concludes that several core differences exist in the approach to creative output. Leong argues that in the East, including in Hong Kong, much of arts education is still attempting to recover from misinterpretations of the underlying beliefs behind Confucius' teachings on education (Leong, 2011). Yet comparisons between 'East' and 'West' have become more challenging in a globalised world, especially where nationality may not constitute cultural membership (Craft, Gardner & Claxton, 2008), and crucially, creativity is often a reaction to potentially inhibitive cultural factors (Runco, 2014). The focus of the paper by (Cheung et al., 2001) is on establishing the validity of the rubric, though the authors could have offered valuable insight and demonstrated further rigour in adapting their scale to fit with what they justify as differences in language and culture, and contributed some of these details around socio-culturally determined notions of creativity in their discussion.

Zedelius et al. (2019) utilise rubrics, raters, and linguistic computations to explore 'uniqueness of expression' in a study on the features of creative writing, corroborating assertions by (Turkman and Runco, 2019) on the potential of computerised scoring for new creativity tests. The rubric used for this study centres around the themes of image, voice, characterisation and story identified by (Mozaffari, 2013), who reports identifying these qualities through a review of the literature on creative language – yet it is unclear from this study precisely what literature was reviewed, how appropriate material was sourced, or how these qualities were arrived at. The same criticism could be levelled at the generic rubric produced by (Vaezi and Rezaei, 2019), another paper included in this review,

Table 2

Summary of characteristics for creativity in narrative writing.

Meaning and Relevance	• 'Aesthetic': "This text has artistic or aesthetic value." *(*Furst, Ghisletta, & Lubart, 2017*. 208)* • Use of 'psychologically meaningful categories' identified through Linguistic Inquiry and Word Count (LIWC) text analysis tool *(*Zedelius et al., 2019*)*
Reader's Immersive Experience	• **Engagement and Flow** ○ 'Hook' (engagement established at the opening) *(*Ashton & Davies, 2015*)* ○ 'Fluency': ability to generate new ideas in each paragraph *(*Cheung et al., 2001*)* ○ Sentence variation: "prose switches regularly and seamlessly between short simple sentences and long complex sentences" *(*Ashton & Davies, 2015*. 318)* • **Clarity and lack of distraction** ○ 'Blocking' (concerning clarity on the spatial relationships between characters) *(*Ashton & Davies, 2015*)* ○ 'Coherence' *(*Furst, Ghisletta, & Lubart, 2017*)* ○ Text cohesion and readability identified through Coh-Metrix text analysis tool *(*Zedelius et al., 2019*)* ○ Few "distracting spelling mistakes" *(*Ashton & Davies, 2015*. 319)*
Development and Control	• 'Elaboration': "A response that includes complex details, metaphors, or sophisticated expressions used to make the language vivid and interesting; may use humour or connections or comparisons for elaboration". *(*Kettler & Bower, 2017*. 295)* • High degree of 'Generation' or ideation during preparation and drafting, followed by increasing or high levels of criticality, or 'Selection' *(*Furst, Ghisletta, & Lubart, 2017*)* • 'Image': (centred around detail and command of devices to convey a scene) *(*Zedelius et al., 2019*)* • 'Flexibility': ability to generate a wide variety of ideas and to develop time space, characterisation and story *(*Cheung et al., 2001*)* • 'Quality': "This text is well written." *(*Furst, Ghisletta, & Lubart, 2017*)* • Overall quality "viewed in light of the story's genre and intended audience [the story was] well written and engaging." *(*Ashton & Davies, 2015*. 318)*
Distinctiveness, Voice and Originality	• 'Originality': "A response that is very different from other students; characterized as quite eccentric, odd, novel, innovative, or original yet successful at communicating according to the prompt; very imaginative". *(*Kettler & Bower, 2017*. 295)* • 'Originality': "[The text] has something special, original." *(*Furst, Ghisletta, & Lubart, 2017*. 207)* • 'Surprise': "[The text] has surprising, unexpected elements." *(*Furst, Ghisletta, & Lubart, 2017*. 207)* • 'Creativity' *(*Furst, Ghisletta, & Lubart, 2017*)* • 'Voice': (visible and distinctive style) *(*Zedelius et al., 2019*)* • 'Originality': (unusual or unexpected choices) *(*Zedelius et al., 2019*)* • 'Originality': ability to embellish ideas using vivid image, novel themes, original plot/setting, unusual story structure, unusual ending, style and emotional tone. *(*Cheung et al., 2001*)* • 'Linguistic originality': Use of relatively rarer words (for example, 'wuthering') and new word formations (for example, 'Muggle'). *(*Form, 2019*)*

which was developed through iterative consultation with multiple groups of experts responding to Likert questionnaires and unstructured interviews on the validity and reliability of the rubric throughout its design. Here, metrics demonstrating the experts' approval for the rubric's comprehensiveness is reported, while the rigour and validity of the qualitative analysis of the conversations with those consulted is not demonstrated – for example, transparency over complexities and contradictory statements arising through analysis.

By contrast, (Ashton and Davies, 2015) contribute a rubric specific to the linked genres of science fiction and fantasy, though once again the design of the rubric is not discussed in detail – the main body of the paper discusses the methodology and results of an experiment exploring the potential of peer feedback. Consequently, once again the rigour and validity of the qualitative data informing the design of the rubric is not discussed. Nevertheless, the criteria described are potentially revealing and point to the value of genre-specific discussions with writers, editors and literary agents in identifying features of creative narrative writing – interestingly, the rubric produced does not give equal weight to all criteria, and focuses primarily on the effect of the writing on the intended audience. Some of these insights are precise, such as the idea of 'blocking' - concerning clarity over the spatial relationships between characters – and the writer's ability to engage the reader early in the piece. Furthermore, this study hints at the role of technical accuracy in the writing, with the rubric detailing 'distracting' errors, which may frame spelling and grammar mistakes in the context of creativity – they detract from the readers' experience of the product.

Finally, (Kettler and Bower, 2017) draw together three instruments for use in creating a comprehensive list of rating criteria: Scales for Identifying Gifted Students (SIGS); the Renzulli Scales for the Rating the behavioral characteristics of Superior Students – Creativity (Renzulli and Reis, 1991); and the Creativity Checklist developed by (Proctor and Burnett, 2004). Each of these methods utilises teachers' observations and ratings of students' personality characteristics to provide a list used by (Kettler and Bower, 2017) in their paper on measuring creative capacity in gifted students. This study (which sought to compare teacher ratings with student products in an exploration of the association of giftedness and creativity) could be critiqued for the use of a single teacher trained in gifted and talented education and with significant and positive relationships to the students' products. The criteria listed in the 'Creativity demonstrated in writing rubric' are revealing, however, and centre around notions of 'originality' and 'elaboration'.

Despite the issues over how these criteria are arrived at and the problematic role of culture in determining definitions and value systems for creativity, each of the studies nevertheless contributes valuable insight into the question of which features may constitute 'good' creative writing through the rubric items used, subcategories rated by participants, and word-level features identified for computational linguistics programmes to explore. There are a number of similarities across these studies, and (when reported) the levels of reliability and validity are all adequate as demonstrated through statistical testing. Table 2 represents a synthesis of the sub-criteria in the rubrics, rating categories and linguistic features used, organised into thematic groups:

3.2. *'Best practices' for assessing the creativity of writing in educational settings*

Numerous groups offered insights into different approaches to assessment within the studies, including teachers and pre-service teachers (in 15 of the studies), professional writers (6), psychologists (5), graduate students (4), university writing tutors (2), professional agents and editors (2), and writing examiners (1). Despite this range, the majority of studies used these participants for providing numerical ratings; few studies, with the notable exception of (Cremin et al., 2017) gathered qualitative data on this topic for analysis. Ashton and Davies (2015), for example, consulted with writers and agents in the genre of science fiction and fantasy and produced a rubric based on their views, but do not describe their views, the coding process, or discuss the nuances of this consultation further. Overall, few studies asked the participants about their practices when it came to assessing writing, and fewer still made use of observation to note differences between participants' claims in light of their practices. Consequently, little is known about the declarative and procedural knowledge of 'experts' regarding how they define creativity and what they value in creative products; this is particularly true of those immersed in the world of professional writing such as literary agents, editors, professional writers and university creative writing tutors.

Cremin et al. (2017) include some insights into authors' assessment and feedback practices in their mixed methods study, combining a qualitative dataset with results from a randomised controlled trial to explore the impact of partnerships between writers and teachers. In their reflections on how feedback had impacted on their own development, the professional authors who participated in this study highlighted perceptive criticism, encouragement, subjectivity and choice in the nature of feedback from peers and collaborators in the world of professional writing. When partnered with teachers in schools, these writers engaged heavily and directly with students on the 'grittier' aspects of drafting, critically addressing children's writing in the classroom forum and referring back to the children's authorial intent. Indeed, observations of adult feedback given at a writers' residential in this study noted that it sought to explore the writers' intention for each piece and was characterised by the sharing of the more experienced writer's sharing of craft knowledge in achieving this goal. In workshops, tutor and peer feedback was public and sought to precisely critique or praise the effect of the writing on the audience, the writer's particular skill, key features of the piece or holistic discussions of the piece as a whole. The study highlights public peer and tutor feedback, and individual conferencing as appropriate practices.

In (Humphry and Heldsinger's, 2019) report on features of writing perceived by examiners as essential to better writing performance, authorial aspects (audience, text structure, ideas, character and setting, and sentence structure) were rated as essential more frequently than conventions of writing (paragraphing, punctuation, and spelling). Despite this, the literature around the assessment practices of teachers when it comes to writing emphasises that this group of experts tends to focus more on correcting errors around writing conventions, such as spelling and grammar. In their study of 10,585 assessment messages by 41 teachers in Spain, (Lucero, Fernández and Montanero, 2018) revealed the predominance of specific grammar and spelling mistakes, which in turn were the best predictor of the overall grade given and neglected other semantic, rhetorical or pragmatic aspects of the children's writing. A similar study by (Peterson, Childs and Kennedy, 2006) with 108 teachers in Canada noted that teachers were generally reluctant to engage with the ideologies in students' writing and tended to indicate and make a greater number of corrections, particularly when the work was attributed to a male student writer. Consequently, other studies included in the review find that, in turn, students show a tendency to focus on these conventional elements too – for example, (Taggart and Laughlin, 2017) explored the students' emotive responses to such feedback. DeBenedictus (2009) highlights the importance of uncovering students' perceptions of the features 'quality' narrative writing. Gulley (2012) included groups for oral, written or mixed feedback, and categorised feedback using a form to comment on statement, content, organisation, sensory details, grammar, and style. Overall, however, there is a paucity of research on the nature of feedback provided for creativity in the writing classroom, particularly at the university level.

Due to the prevalent use of (Amabile's, 1982) Consensual Assessment Technique by cognitive psychology researchers (in which panels of experts rate creative products, and interrater reliability is tested for), the majority of the studies reviewed sought to recruit judges from relevant backgrounds. However, these studies utilised their judgments rather than seeking to unpick their practices and understandings regarding creativity in writing. One of the frontiers for research in this area is examining who exactly qualifies as an 'expert', and where a balance can be struck between recruiting for purpose and recruiting for convenience when it comes to academic research. Kaufman, Gentile and Baer (2005) reported a strong degree of correlation between the ratings of gifted creative writing students and an 'expert' panel of teachers, writers, and creativity psychologists. Despite being older, a limited correlation between university students and published writers in rating stories for their creativity was then found by (Kaufman, Baer and Cole, 2009), with the experts tending to rate lower for creativity than the novice raters, who also demonstrated lower inter-rater reliability and less consistency in their ratings than the expert writers. Further research is needed to establish the parameters and nature of these differences and to unpick the craft knowledge understood by these expert groups.

4. Discussion

The range of different methods used to assess creativity in writing reflects the interdisciplinarity of the field. Broadly speaking, quantitative tests designed to measure creativity were preferred by those working in the field of cognitive psychology; academics from the field of linguistics tended to run computations to unpick different linguistic elements within the writing corpus; educational researchers and creative writing academics adopted a number of different approaches, including quantitative data gained through writing tests and surveys, and qualitative data gleaned through the analysis of writing samples or interviews with participants. In general, research has approached the puzzle of assessing creativity in writing using four main methods: rubrics, ratings, computational linguistics and peer assessment. In this section, I shall discuss the merits and limitations of each method.

4.1. Consensual assessment technique

Research (Amabile, 1982, 1983, 1996; Baer, 1993, 1997, 1998; Kaufman, Baer & Gentile, 2004; Runco, 1989) has repeatedly demonstrated the high inter-rater reliability of the Consensual Assessment technique (CAT), assuring its validity – consequently, it has since come to be regarded as the 'gold standard' of creativity assessment (Baer & McKool, 2014). Ten of the studies included in this review utilised some form of the CAT to assess creativity in writing, largely authored by the same group of psychologists testing the technique in a variety of circumstances and fields: due to its grounding in creativity as a socially-constructed quality, the CAT is an interesting tool for testing notions of creativity in different contexts. For example,(Baer, Kaufman and Gentile, 2004) obtained high levels of inter-rater reliability using the consensual assessment technique for a panel of judges including teachers, writers and editors, and psychologists for three different types of writing – stories, personal narratives, and poems – suggesting the technique can be extended to nonparallel creative products.

However, without additional data the CAT reveals little more other than that these products are creative, and further methods are required to understand more precisely the complex elements that combine in products perceived to have creative value. Some of the included studies (for example, Furst, Ghisletta, & Lubart, 2017; Kelly, Knight, Peck & Reel, 2003; Ng & Yeung, 2011) circumvented this issue by modifying the technique slightly, combining it with the use of more precise criteria to yield more revealing insights. Several studies have also identified how specific procedural choices impact the CAT's reliability as a measure, and researchers' depth of knowledge about procedures and their effects still remains incomplete (Cseh & Jeffries, 2019). A further criticism is that it cannot yield standardized scores the way many achievement tests can: the creativity of the artifacts, ideas, or performances in a group is judged in relation to each other, not some pre-established standard (Baer & Kaufman, 2019). But the main limitation with the CAT is practical – it is logistically difficult to source appropriate judges (Kaufman, Baer, Cole & Sexton, 2008), and therefore to use in the classroom context.

Aside from the limitations imposed by to the requirement of multiple 'expert' judges and the nuances around who qualifies for this (Kaufman et al., 2005, 2009), the studies included that used the CAT raised a number of questions around the extent to which it can adequately capture the diverse manifestations of creativity, especially when considering the possibility that creativity can be determined by sociocultural differences: Alhusaini and Maker (2015), for example, found that teachers using the CAT rated the writing of white students as more creative than the work of Mexican American and Navajo students, found no difference between students from different grade levels, nor significant differences between boys and girls for open-ended stories. By contrast, (Kaufman et al., 2004) used the CAT with a larger pool of judges and found no significant differences between African American-Caucasian differences on any of three writing tasks; the only significant differences in creativity ratings occurred in poetry between Latino-Caucasian groups and Latino-Asian groups. In her discussion of Creative Writing at MA Level, (Newman, 2007) notes that students with different cultural memberships can find these nuances in value systems for creativity frustrating, and that in writing out of a different tradition they can find themselves sidelined. This is one further issue with the CAT: it perhaps unfairly disadvantages certain children where elements such as socio-economic status, writing in a second language, and differing cultures and environments affect socio-culturally determined notions of creativity.

4.2. Rubrics

While several of the included studies made use of rubrics (Beyreli & Ari, 2009; Humphry & Heldsinger, 2019; Johnson, 2003; Lucero et al., 2018; Montanero, Lucero & Fernández, 2014; Zedelius et al., 2019), or designed rubrics for use in their studies (Ashton & Davies, 2015; Cheung et al., 2001; Kettler & Bower, 2017), relatively few designed rubrics based on their findings or went into detail in justifying the criteria used in their rubrics. Where this was the case, such as in Vaezi and Rezaei's (2019) multi-phase rubric design project, issues remain around the criteria that are included in the final product and the transparency over how qualitative data was analysed. Rubrics are commonly used in writing classrooms, but findings from Vaezi and Rezaei (2019) and (Ashton and Davies, 2015) suggest that training markers in their use has a significant effect on their consistency and accuracy.

One of the limitations around the use of rubrics concerns where the criteria come from. In her analysis of the underlying beliefs and values about writing that underpin writing curricula in Canada, (Peterson, 2012) notes how criteria can favour some writing discourses, and sideline others. Rubrics can thus be highly influenced by political agendas and curriculum drivers, and can impose a framework onto the writing. Despite listing a set of criteria to look for in the work, they are often also subjective in nature and can result in writing that displays several desirable features, but which may lack other important elements. Further to this, many rubrics tend to give relatively equal emphasis to each of the criterion included, where there instead may be nuances around these, and may be influenced by different conceptions of creativity. Several of the studies thrown up by the literature search more generally reported rating for the *presence* of a desired story feature, rather than its execution or the design choices around this.

Gardner (2012) argues for genre-specific criteria in rubrics, finding that students scored significantly higher with a rubric more precise to narrative writing than a generic one. Yet genre (within the general umbrella of fiction) was not an overt concern for the majority of the studies included in the review, with most of them specifying no genre or a mix of genres (17). When genres were detailed, these were obscure; for example, recounting a memory or dream (6), the story of a town hero (2), a prompt involving seasons or the weather (2), science fiction (2), fantasy (1), fable (1), crime (1), and unspecified multimedia (1). There is a gap in the literature regarding the assessment of creativity in writing for most of the major narrative genres, such as romance, action, mystery, horror, or historical fiction. Furthermore, relatively few studies have identified characteristics for creativity in writing and, while there is a degree of consensus amongst them, further research is needed to produce a framework for assessment purposes in the writing classroom. Such criteria can be problematic, and difficult to identify in the context of academic research – (Zedelius et al., 2019), for example, define 'voice' as "unique and recognizable if you, when reading another piece by the same author, would recognise that it is written by the same author" (Zedelius et al., 2019. 891). In this instance, research would arguably have to work with several pieces of writing to identify such 'voice', and to be able to identify the writers in order to then zoom in on how this was achieved. Notions of originality, which also appeared in several rubrics, must be understood in context and defined further – is a text original if it creates an entirely new form, or alters an existing one? Similarly, testimonials from professional writers often discuss 'quality' and 'voice' as achieved through several drafts (see, for example, Bingham, 2014; Lamott, 1995; Magrs, Lodge, Jones & Bell, 2001), while research often works with a first draft produced under controlled conditions, aiming to assess blindly and for objectivity regarding the writers' intentions, influences and backgrounds, as well as the writing's social and cultural context, in order to generate findings on the subjects as a collective.

4.3. Peer assessment

The third method for assessing creativity in writing related to the collective and individual responses of children and adults within the writing classroom itself. Peer assessment and feedback featured in several of the included studies. Some studies observed and analysed the nature of these interactions (Cremin et al., 2017; Peterson, 2003; Rojas-Drummond, Albarrán & Littleton, 2008). Others (such as Ashton & Davies, 2015; Ng & Yeung, 2011) collected ratings from peers and analysed these as quantitative data. Peer assessment allows for some subjectivity and expression of the diversity of different effects on the reader when it comes to the 'little-c' creativity choices made by the writer. While it is easy to operationalise in the classroom, peer feedback can be more difficult to capture, analyse and code into a generalisable framework, in part due to this diversity of responses and the element of subjectivity. Arguably, it is therefore limited in terms of identifying particular features of the writing or measuring to some objective standard. Despite this challenge, (Peterson, 2003) claims to have been able to able to trace revisions made in the writing to peer interactions, and such interactions may allow for different social and cultural notions of creativity to be discussed and appreciated.

Cremin et al. (2017) highlight the potential of peer and expert feedback, but note a certain element of craft knowledge at play in the giving and receiving of constructive criticism, and sensitivity to the social, emotional and cognitive demands of writing and revising drafts. Indeed, craft knowledge appears to play a role in several of the studies, where participants received training before rating for creativity, using rubrics, or providing feedback – decisions made by the research team, presumably, to provide a vocabulary for creative writing or a hierarchy of ideas for the participants to draw upon when making their decisions. The importance of this is most sharply evident in (Ashton and Davies, 2015) in which the peer group trained in the use of the rubric was more closely aligned with expert's rating than the group that did not receive training. One potential pitfall of peer assessment is highlighted in this study, in that both student groups displayed a tendency to rate most texts towards the centre of the scale (central tendency rating error), deviating little from average ratings even for texts that the experts rated as much lower or higher in quality.

Table 3
Summary table showing the merits and limitations of each method for researchers.

Method	Critical Evaluation
Consensual Assessment Technique (CAT)	***Merits:*** • Widely recognized technique for evaluating creative products. • Has been validated repeatedly through high inter-rater reliability metrics that demonstrate consistency in 'expert' judges' ratings relative to non-expert groups. ***Limitations:*** • Logistical challenge of sourcing multiple appropriate judges makes this technique difficult to operationalize in different contexts, most notably the classroom. • Provides an overarching judgement on creative merit, relying on subjective criteria without necessarily offering insight into the creative characteristics or features of the products assessed. • Products are rated comparatively, rather than to an overarching standard: thus high or low scores are only representative of the selection of products that are rated. ***Enhancement Suggestions:*** • The CAT is best used in a research context and using the guidelines associated with the technique (see Amabile, 1982). • The limitations of the method for assessing creativity in writing can be curtailed by employing the CAT in combination with a technique that allows for deeper insight to be gained through articulation and discussion of the creative characteristics of the products assessed to be creative (for example, data gained through peer review, focus group discussion, or qualitative interviewing).
Rubric	***Merits:*** • Often used in classroom contexts for marking and grading, and represent a commonplace and practical tool for assessing writing. • In research, they can provide a framework for training participants in key concepts and in standardizing responses. ***Limitations:*** • The criteria included in rubrics, especially concerning creativity, can often be generic, vague, open to interpretation, or influenced by political and other agendas. ***Enhancement Suggestions:*** • Rubrics are versatile and practical tools, but require a great deal of consideration and research in grounding the features listed, and their respective gradations. These must be precise, clear, and ideally tailored to specific genres and text type outcomes. • While rubrics can be used as training tools for research and peer review, a certain degree of valuable insight and subjectivity can be lost by standardizing responses. It is perhaps best paired with a more open feedback system such as peer review, allowing more interpretive responses to influence final judgements.
Peer Review	***Merits:*** • Facilitates more immediate and varied feedback, allowing for subjective responses to be articulated back to the writer. • Responses can be insightful and detailed, and allow for broader interpretations of the writing to be voiced than is typical of a single marker or a scaffolded evaluation system. ***Limitations:*** • Feedback offered in this way can lack objectivity, be influenced by different biases through a lack of scaffolding, and risks being implemented inconsistently in different classroom and research contexts. ***Enhancement Suggestions:*** • The subjective elements in peer review are simultaneously a strength and a weakness of the method. Feedback can be enhanced by providing a language for the key constructs of creativity theory and a framework for responses in the classroom, with clear and consistent guidelines for implementation in different research contexts. • Peer review is perhaps best paired with another technique (for example, corpus linguistics or the CAT) to glean both quantitative metrics and the benefits of broader, more subjective responses to establish insight into how authorial choice and intentions factor into creative success.
Corpus Linguistics and Syntax-based Scoring	***Merits:*** • Provides revealing quantitative metrics based on word-level features of a text, and can demonstrate relationships and patterns offering insight into the relationships of these features to wider judgements of creativity.

(*continued on next page*)

Table 3 (*continued*)

	Limitations: • May not be practical for use in a classroom context, and can be laborious in research if being conducted without the aid of computational software to process large bodies of text and identify patterns and word-level characteristics. • Characteristics need linking to valid conceptions of creativity to bridge the linguistic elements with broader constructions of creative processes and products. ***Enhancement Suggestions:*** • Since corpus linguistics provide metrics on the word- and sentence-level characteristics of a text or group of texts, in order to link these to notions of creativity or suggest these features contribute towards creativity this method must be paired with another such as the CAT to more powerfully establish the creativity of the products and the linguistic elements that work towards creating this effect.

4.3.1. Corpus linguistics and syntax-based scoring

By contrast, the studies that paid closer attention to syntactic features of the writing attempted to zoom in on the linguistic features of the writing and use these as quantitative data. In one examination of the first 1000 words of 92 bestsellers' linguistic originality, (Form, 2019) found that the writers' use of relatively rarer words (for example, 'wuthering') and new word formations (for example, 'Muggle') were a significant predictor of the bestsellers' popularity, a finding with implications for the level of emphasis that writing teachers afford to aspects of style and vocabulary choices. This was an effect that remained significant after controlling for the year of first publishing and the gender of the author. Zedelius et al. (2019) also conceived of style as visible and distinctive through application of stylistic tools but established other metrics for the writers' voice, image and originality in their linguistic computations. The results of this study compared the computerised metrics of text cohesion and readability to human ratings of creativity, finding that linguistic features predicted the ratings to a significant degree. This suggests that at least some aspects of creative writing can be captured by computerised measures, and that different pieces of writing could possibly be compared to some objective standard despite their uniqueness.

While several of the studies appeared to focus more on writing proficiency and performance than creativity, syntactic features certainly seem to play an important role in the improvement of writing generally, with some studies such as (Sandiford and Macken-Horarik, 2020) examining grammatical realisations of dimensions of narrative texts that to account for embryonic developments in student writing. Similarly, (Yeung, Ho, Chan and Chung, 2017) found that spelling, transcription, and oral narrative skills were unique and important predictors of Chinese writing performance. Dockrell, Connelly, Walter and Critten (2015) and (Diercks-Gransee, Weissenburger, Johnson and Christensen, 2009) scored student writing for productivity and text accuracy, though findings are inconclusive with regard to the relevance of these measures for assessing the creativity of the writing used in these studies. In summary, word-level approaches appear to have the potential to provide some revealing results, but nevertheless could not be very effectively operationalised in the classroom and allows for little in the way of subjective interpretation of the creativity evidenced in the writing, instead attempting to identify objective features.

For researchers seeking to select a method appropriate for their research design in the field of assessing creativity in writing, a critical evaluation of each technique is offered in Table 3. Since each technique benefits from approaching creativity writing in subjective or objective ways, and leans towards producing qualitative or quantitative data, combining techniques can mitigate the limitations of any one particular method. From the studies reviewed in this paper, it is noted that relatively few thus far have combined methods when approaching the subject of assessing creativity in writing.

4.4. Contribution of the review to answering the research questions

The literature search has revealed a number of 'gaps' in the existing literature. Relatively few empirical studies exist that explore the assessment of creativity in the adult education and professional sectors and has highlighted the range of different approaches that researchers have taken toward establishing measures of creativity and data around creativity in narrative writing. While several of the studies contribute significantly to the knowledge base, a lack of research, particularly in interdisciplinary and high-quality qualitative studies, limits the potential of the review to answer its research questions conclusively. Relevant methods and evidence were embedded in a range of different studies that sought to explore the limits of the consensual assessment technique, the nature of giving and receiving feedback on writing, and attempts to find objective linguistic features of creative writing: few of the studies included address the research questions posed by this review directly.

4.4.1. What are the characteristics of writing deemed to exhibit creativity?

The relevant studies broadly suggest that the perceived creativity within a piece of writing are the ways in which a writer controls their craft to immerse their intended audience in the unique experience of their work, without 'breaking the illusion'. The critical and evaluative elements of writing are found to be crucial in achieving this, alongside a fluent, flexible and distinctive style achieved through innovative use of language, different sentence structures, and new ideas within paragraphs as well as across texts. Clarity in the mind of the reader as to the way events unfold in the narrative is achieved through mastery of spatial relationships between characters, and vivid description. Finally, the writer works to intentionally engage and surprise the reader in a variety of ways – generating and playing with ideas, successfully taking risks and, as the evidence so far hints at, even demonstrating some enjoyment of the process.

4.4.2. What is known of the practices of 'experts' in assessing creativity?

While several groups have been identified in research for use as 'experts' in this field, relatively few studies have attempted to unpick their definitions and values regarding creativity in writing, and have instead sought to harness their judgement in identifying creative products. Further research will need to address some of the complexities around such judgements and the reasons for them, unpicking differences in what participants state about creativity in writing in the light of how this manifests in their feedback, judgements, and practices. Relevant studies within the final pool of studies suggest that distinctions exist regarding the writer's command of audience, structure, ideas, character and setting as opposed to the technical aspects of writing, with teachers demonstrating a tendency to focus on correcting these latter errors and professional writing tutors and examiners privileging the former in the context of the writers' intentions for their piece, and the choices made to achieve this.

4.5. Limitations of the research reviewed

Overall, the evidence base around effectively assessing and providing feedback in relation to assessing the creativity of narrative writing is not strong for any one particular method, and the issue of a lack of an agreed-upon definition for creativity persists in problematising synthesis of the findings. There is a paucity of research across cultures, with the vast majority of studies being conducted in the USA and other Anglophone countries. In a similar way, the genre of a piece of narrative writing appears to have been important in the studies reviewed, for example in the precise items identified and included in the rubric produced by (Ashton and Davies, 2015), yet creativity within several key narrative genres have not been studied in the recent empirical literature. The majority of the studies (17) did not specify the genre of writing used for the research, used a mix of creative narrative writing, or made use of open tasks whereby students could interpret a stimulus how they wished and, therefore, produced a range of outcomes. Six studies required students to use dreams or memories as a basis for their narrative writing: other genres that guided writing used for the studies included stories based on the weather or seasons, science fiction, stories whereby a character becomes a hero, and other highly specific scenarios.

Many of the 'expert' voices that could add value to this discussion have not been attended to adequately by the field; there has been a tendency to testify the validity of a given rubric without demonstrating rigour over how the criteria had been arrived at. Few high-quality qualitative or mixed-methods studies have been conducted exploring or providing a basis for further research. The majority of studies used only quantitative data in their analysis: some of these involved experiments or quasi-experiments (7), while others used a non-experimental design such as the CAT or a survey to gather data for analysis (20), and provide limited contributions to the research questions used in this review in unpicking evidence of creativity in writing. Just five studies combined both quantitative and qualitative datasets for analysis, but without directly addressing the research questions of this review even these are limited in how far, and how accurately, we can know anything on the topic of effectively assessing the creativity of narrative writing. Seven of the studies reviewed focused only on qualitative data; however, rigour in qualitative research was a recurring issue with several of the studies, or the literature used to design the instruments used in them, particularly with regard to how themes and criteria were identified and adequacy of the discussion around this. There was also a lack of studies that harnessed the use of multiple techniques for assessing creativity in writing. Due to the merits and limitations of each other techniques identified in this review, it is suggested that more studies that adopt an interdisciplinary approach utilising several measures for creativity, attending to the social and cultural aspects that influence it, will no doubt enrich and broaden the field and aid in the design of evidenced creativity assessment frameworks that could be operationalised in the writing classroom. A viable progression for research in this field is offered in Fig. 2; the literature search

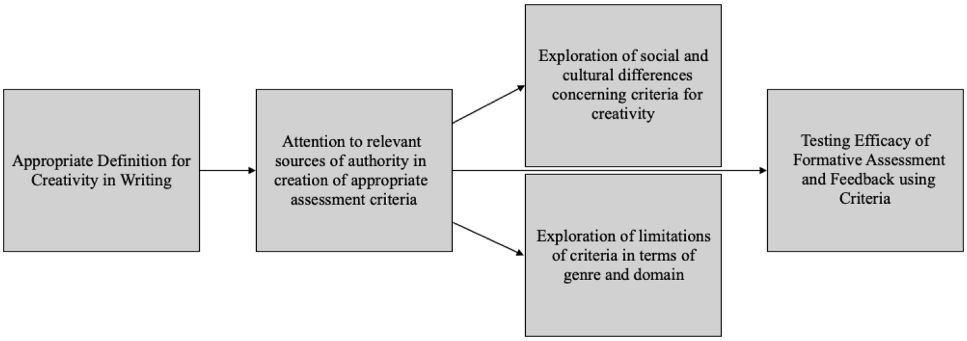
Appropriate Definition for Creativity in Writing
Attention to relevant sources of authority in creation of appropriate assessment criteria
Exploration of social and cultural differences concerning criteria for creativity
Exploration of limitations of criteria in terms of genre and domain
Testing Efficacy of Formative Assessment and Feedback using Criteria

Fig. 2. Summary of gaps and progression for the field.

reported here suggests significant gaps in all stages of this progression, and the need for further sustained research in addressing the issue of effective assessment for creativity in writing.

5. Conclusion and implications

This systematic search and review of studies from 2000 to 2020 has offered a clearer idea of the existing approaches and frontiers of the research field as it stands, and some of the existing debates within it. It has revealed that few papers have attempted to assess 'little-c' creativity in writing, as distinct from a more generic, genre-bound definition of 'creative writing', and reasserts issues caused by the lack of a consistent definition for creativity in writing and how this problematises our understanding of the features of writing deemed to exhibit creativity. Four key research methods have been identified in the literature for assessing creativity in writing: rubrics, linguistic computations, peer feedback, and the Consensual Assessment Technique (CAT), each with their own merits and limitations, which have been discussed in reference to the wider body of creativity and writing research. The review finds that relatively little is understood for what is valued in writing by 'experts' in the world of professional writing, and has discussed gaps in the body of research concerning genre and the potential importance of genre in articulating precise criteria. While there is also currently limited evidence that formative assessment for creativity in writing improves student creativity, the role of feedback in determining progress and the potential impacts on the accuracy of peer assessment have been outlined. The next steps for research in the field have been discussed in the light of the existing literature.

Defining and outlining a strategy for effective formative assessment for creativity in writing remains an issue for the area of developing a pedagogy for creative teaching and learning; qualitative aspects of research in this area generally require a greater degree of rigour, direction and specificity in their approach in order to add value to this field and give credence to the validity claims of quantitative instruments and assessment criteria. It is recommended that research make greater use of the experts identified in this field, producing high-quality qualitative studies that reveal rich insight into the topic under discussion from those immersed in the world of professional creativity in writing, with close attention to the texts produced and the ways in which qualitative coding arrives at themes and characteristics for creativity in writing. Finally, studies should attend to notions of creativity within its social and cultural context, exploring the impact of a diverse range of writers and texts. Research could also be conducted in a wider range of cultures and societies: the field is dominated by studies carried out in Anglophone countries, yet writing creatively occurs in all languages, and diversity in contexts and findings creates further opportunities and avenues for creativity scholarship to pursue.

6. Funding details

The Creative and the Critical: Effective Assessment of Creativity in Writing Economic and Social Research Council (Swindon) 2018-09-30 – 2026-09-29 | Grant GRANT_NUMBER: 2,096,294 URL: https://app.dimensions.ai/details/grant/grant.7837446

Data Access Statement

The research data supporting this publication are provided within this paper.

CRediT authorship contribution statement

Richard D'Souza: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing.

Appendix 1

Summary Table of Research Papers (Participants Aged 5–15)

Appendix 2

Summary Table of Research Papers (Participants Aged 16+)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
Alhusaini and Maker (2015), USA	To what extent could judges distinguish between technical quality and creativity when examining children's stories? What were the differences in the creativity of children in third, fourth, and fifth grades? What were the differences in the creativity of male and female students? What were the differences in the creativity of children from different ethnic groups? What were the interactions of children's ethnicities, genders, and grades in their creativity?	139 students (72 F; 67 M) aged 9–11. Mixed sample of White, Mexican American, and Navajo. 5 teachers (4 F; 2 M) acting as 'experts' for CAT Open-ended stories (no constraints) produced as part of DISCOVER assessments.	Consensual assessment technique used – experts rated the products on a 1–7 low/high creativity scale. Technical quality was also rated. Quantitative dataset. Cronbach's alpha, correlation between ratings; and *t*-test and both a 1-way between-subjects and 3-way ANOVA run. No differences found between students from different grade levels, nor significant differences between boys and girls. Teachers rated the work of white students as more creative. This can be attributed to differences in socio-economic status, the difficulty of writing in a second language for some students, and the influence of culture and environment on socio-culturally determined notions of creative products.
Baer et al. (2004), USA	Can the consensual assessment technique be extended to nonparallel creative products?	Texts written by 8th grade students: • 103 stories • 103 personal narratives • 102 poems 13 expert judges (5 writers/editors; 4 teachers; 4 psychologists)	Consensual assessment technique rating the creativity of the products using a 6-point scale. Quantitative dataset. Coefficient alpha interrater reliabilities calculated. Very high levels of inter-rater reliability were obtained, demonstrating that the consensual method can be validly extended to such samples.
Beyreli and Ari (2009), Turkey	What is the rate of concordance amongst raters using the same scale? What are the factors that lead to probable differences in assessment? What is the level of concordance amongst raters in terms of the external structure, language, expression, and organisational parts of the texts being assessed?	200 memoir narrative texts written by 6th and 7th grade students in Istanbul. 6 'expert' raters (2 teachers and 4 masters)	Analytic rubric comprised of ten items. • External structure: format, spelling, punctuation. • Language and expression: vocabulary, sentences, paragraphs, expression. • Organisation: title, introduction, story, conclusion. • Raters given two sessions of training using the rubric. • Quantitative measurements: Pearson correlation, Friedman coefficient and Kendall's W to measure levels of concordance and reliability. • Sufficient concordance amongst raters.
Bintz and Shake (2005), USA	What can we learn about the impact creating writing portfolios has on preservice teachers' understanding of writing portfolio assessment? How can we use findings to develop more informed instruction in literacy courses in our elementary teacher education program?	92 preservice teachers in 3 groups of students (inc. 1 control group of 15 students and the 77 other assigned to produce the portfolio). Mix	Experimental groups produced a portfolio similar to that required in grades 4–11 comprising a personal narrative, a poem, a short story, and a letter to a reviewer. A 3-page reflective paper was also included for the teachers to complete, and a 4-item Likert-style survey with a comment section administered at the end of the course. Mixed dataset: Grounded theory coding for qualitative data to identify recurring patterns of student responses in reflective survey comments and contents of the writing portfolios; percentiles and frequencies for quantitative data also explored. Findings indicate that active engagement with writing portfolios significantly and positively influence preservice teachers' competence in and confidence with writing portfolio assessment.
Cheung et al. (2001), Hong Kong	How valid and reliable is the Chinese Creative Writing Scale for use with primary school students in Hong Kong?	69 compositions from two grade 5 classes 2 independent raters with experience teaching, who received one training session using the scale. Guided fantasy: "If I were a flower…" topic stimulus.	Scale organised into three parts, scored from 0 to 2: • Fluency: ability to generate new ideas in each paragraph • Flexibility: ability to generate a wide variety of ideas and to develop time space, characterisation and story • Originality: ability to embellish ideas using vivid image, novel themes, original plot/setting, unusual story structure, unusual ending, style and emotional tone.

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			After receiving the training input each rater provided ratings on 20 pieces of writing, before discussing their understanding of the guidelines and reaching a consensus.Quantitative dataset: intra-class r and Cronbach's alpha used to determine interrater reliability and internal consistency. Pearson correlation and exploratory factor analyses to examine construct validity.Positive results suggesting the validity and reliability of the scale.
deBenedictus (2009), USA	Is the Make-It-Better/Make-It-Worse protocol a fruitful way of learning about students' writing?	6 4th Grade students (3 F; 3 M)3 versions (first draft, MIW version, MIB version) of 2 pieces of writing: • Narrative response to a wordless Peanuts cartoon • Persuasive letter to presidential candidates	Charts produced displaying phrases and sentences that changed across the course of the unit, showing what students had adapted in their work to 'make it worse' and then to 'make it better'. Author notes that the exercise revealed which strategies students had most control over and where differences in metacognitive knowledge and skill lay. The students exhibited particular attention to the visual appearance of their writing, their spelling, tone, and sentence complexity.
Diercks-Gransee et al. (2009), USA	Could reliable alternative curriculum-based measures of writing be identified for high school students? Do these alternative curriculum-based measures of writing correlate with holistic scores of writing proficiency for high school students? Do these alternative curriculum-based measures of writing proficiency correlate with the Wisconsin Knowledge and Concepts Examination results for high school students? Could these alternative curriculum-based measures of writing be used to accurately screen for learning disabilities or low performance on a statewide measure of language arts for high school students?	82 10th¯Grade Students (39 F; 43 M) 7 scorers (1 researcher and 6 psychology students) 2 narratives produced in response to 'story starter' prompts. Students given 30 s to think and 10 min to write. 'I stepped into a time machine…' 'It was a dark and stormy night…'	Samples were scored for four alternative curriculum-based measures: • Number of incorrect word sequences • Number of correct punctuation marks • Number of adverbs • Number of adjectives Statewide assessment results for the narratives also obtained. Quantitative dataset. Alternate-form bivariate Pearson product-moment correlation coefficients, holistic mean scores, bivariate Pearson product-moment correlation coefficients and cross-tabulation analyses used to analyse the data.Results revealed moderately strong alternate-form reliability and criterion-related validity coefficients for 'incorrect word sequences'. Although 'correct punctuation marks' was found to be reliable, the criterion-related validity evidence varied according to the type of criterion measure. Other findings indicated that 'incorrect word sequences' and 'correct punctuation marks' cut scores may have utility for specific screening purposes. The curriculum-based measures of 'number of adjectives' and 'number of adverbs', however, were not found to have the technical adequacy needed for predictive purposes.
Dockrell et al. (2015), UK	What are the potential uses of curriculum-based measures of writing as a means of evaluating writing products in a cohort of English primary school pupils?	236 primary school pupils (95 F; 130 M)3 typeset writing tasks: • Expository: "There are many things that make a day at my school interesting/ boring". • Narrative: "One day I had the best/worst day at school". • Letter outlining ideal house (WOLD standardised task)	Scripts scored by graduate psychologists or project directors after initial training. • Productivity: total words produced, number of correct word sequences, total punctuation marks, total number of complete sentences. • Text accuracy: proportion scores for words spelled correctly, correct word sequences, correct punctuation marks, correct sentences. Quantitative dataset. Cronbach's alpha, mean scores, standard deviations and repeated measure ANOVA conducted for each measure, with genre type and year group.The CBM –W measures were differentially sensitive to development and showed construct validity as evidenced by their association with the norm- referenced

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			test measuring writing quality. Change over time was also evident and significant differences between narrative and expository texts were found. Pupils with special educational needs scored significantly more poorly on the CBM-W.
Duran and Yilmaz (2019), Turkey	What is the primary school fourth grade students' writing story level? What do primary school teachers think about students' story writing skill? Which preferences do the fourth graders prefer based on plot, main character, time and place in their stories?	142 fourth-grade students in Manisa province. 22 classroom teachers. 3 field experts to mark the work. Open task (no genre specified)	Quantitative scale for evaluating story writing skill, involving allocated points for pre- and post-writing approach as well as writing process, and the 'shape'/technical accuracy of the piece. 0–5 rating scale used for each item. Semi-structured interviews with the teachers on their opinions towards the writing. 6 categories identified: strong aspects, lovable aspects, challenging aspects, unlovable aspects, activities to develop writing the story, and blind sides. Scanning of the stories themselves to determine plot, main character, time and place. Mixed dataset. Frequencies and averages calculated as well as qualitative codes. In the study the students were the most successful at criterion of using imagination when they wrote stories. Students stated that they didn't like writing stories when the plot was limited. Punctuation, paragraphing and the order of events emerged as key issues. Suggests that students will be more successful if they're free in terms of the plot. The fourth graders preferred to write stories set in forests, about friendship, with characters named Ali or hero narrators, and set in the summer months.
Gardner (2012), UK	How accurately does the Assessment of Pupil Progress evaluate students' ability to write narrative texts?	3 groups of samples of narrative writing (69; 42; 37) Unspecified.	Pupils' writing marked at 3 stages using two assessment criteria: • Assessment of Pupil Progress (APP) generic criteria: statements for spelling, vocabulary, technical accuracy, sentence variety, paragraph cohesion, structure, appropriacy, and imaginativeness. • Assessment of Narrative Writing (ANW): descriptors for plot, narration, characterisation, setting, words and grammar, textual organisation, experience and meaning, affective reader response. Mixed dataset. Descriptive statistics for pupil scores and gain scores with extracts to illustrate effect of 'best fit' statements and descriptors.Findings suggest a significant number of pupils demonstrate higher levels when assessed against the ANW criteria than against the APP criteria.
Humphry and Heldsinger (2019), Australia	Which features of writing do examiners perceive to be germane when judging differences between performance levels of narrative writing?	5 NAPLAN writing examiners with teaching experience Subset of 37 pieces of narrative writing from a larger bank of assessed work – randomly selected from different grade levels to represent a range of abilities.	Standard NAPLAN rubric marking criteria comprising range of marks available for audience, text structure, ideas, character and setting, vocabulary, cohesion, paragraphs, sentence structure, punctuation, and spelling. Raters viewed pairs at relatively similar levels and made decisions on which was higher quality. A total of 454 comparisons made. Quantitative dataset. Mean averages generates for each rubric criterion. Chi Square and Pearson's Rho used for measures of variance. Conventions of writing (paragraphing, punctuation, spelling) were seldom recorded as being essential to a better rating performance, while authorial aspects (audience, text structure, ideas, character

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			and setting, and sentence structure) were rated as essential more frequently.
Johnson (2003), Bangladesh	What is the level of children's reading and writing ability in Bangladesh? Which profiles of literacy are diagnostic? Where is innovation and development assistance best targeted? What are ways we might begin to assess children's potential for meaning-making beyond the linguistic mode alone?	2657 Grade 5 children (1414 F; 1243 M) for writing task. Stratified random sample of 687 children (349F; 338 M) for reading task. 'Continue the story' task from tiger and stork story prompt (30 mins).	Writing scale with 3 dimensions: Global: characters, setting, coherence, chronology, overall global score. Local: cohesion and connective devices, overall local score. Formal accuracy: handwriting, spelling, overall accuracy score. Reading scores collected for fluency and accuracy based on reading record sheets. Levels 1–3 assigned based on broad indicators of achievement for both writing and reading scores. Mean averages and graphs generated for each dimension. The paper concludes that large-scale studies of children's literacy are necessary and will continue to provide important sources of information for governments attempting to alleviate poverty and create equitable access to education and other social services. On the other hand, data obtained in this fashion masque the potential of children as creative meaning-makers. The study reported here shows that teacher-based assessment, expanded to recognise modes of meaning-making other than language, can be a vital, additional source of information for those interested in children as learners.
Kaufman et al. (2004), USA	What differences can the Consensual Assessment Technique reveal on the issue of creativity for racial and gender groups?	13 experts (teachers, psychologists, creative writers) Samples of 8th Grade writing: 103 personal narratives 103 poems 104 stories (No uniform genre)	6-point low, medium, high creativity scale for CAT approach. Portfolio mix of students' best work selected to represent the different racial groups selected from wider bank of writing used in a previous NAEP study (1998). Quantitative dataset: Cooefficient alphas to determine interrater reliability and ANOVA to determine significant differences. No significant differences for gender on any of the tasks and no significant African American-Causasian differences on any of the writing tasks. Only significant differences in creativity ratings occurred in poetry between Latino-Causcasian groups and Latino-Asian groups.
Kaufman et al. (2005), USA	Do gifted student writers and creative writing experts rate creativity the same way? How much did raters in each group (novice and expert) agree with each other? What is the relationship of the novice raters to the expert raters?	8 gifted creative writers who had been accepted to a highly selective arts school. 13 expert judges consisting of middle school teachers, published creative writers, and psychologists who studied creativity. 27 short stories and 28 poems selected from the 1998 NAEP Classroom Writing Study.	Participants independently rated the poems and short stories on a scale of 1–6 (not creative/highly creative) and submitted by mail. Quantitative dataset. Coefficient alpha interrater reliability analysis for RQ2; Pearson correlation coefficients for RQ3. Strong agreement amongst experts, and agreement amongst gifted students within accepted standards. Strong degree of correlation between the ratings of gifted students and experts ($r = 0.78$ for poems and $r = 0.77$ for short stories)
Kettler and Bower (2017), USA	What is the relationship between the teacher's rating of student creativity and students' creative writing samples? Do identified gifted students score higher than general education students on the creative writing samples? Are there gender differences in the teacher's ratings of student creativity and/or in the creative writing products?	155 Grade 4 students (76 F; 79 M) in an urban district in a southwestern state. 41 identified as gifted and talented by the school. 1 teacher 4 research assistants	Teacher provided ratings for each student on three different scales (Creativity Checklist, Renzulli Scales – Creativity and SIGS – Creativity). Creativity Demonstrated in Writing rubric designed (0–3 scores for two criteria: 'originality' and 'elaboration') and 4 research assistants trained to use it with the students' writing. Quantitative dataset. Scores for all participants and participant groups tabulated. Pearson correlation coefficients and multiple regression analyses used. Identified gifted students scored higher than general education

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			students in both teacher ratings of creativity and written products. Female students scored higher than male students in both teacher ratings of creativity and written products. These findings suggest that teacher ratings moderately predict creative student products. Identified gifted students scored consistently higher than general education students in creativity, and females scored slightly higher than males on creativity measures.
Lucero et al. (2018), Spain.	What is the nature of teachers' annotated feedback messages on texts written by students? Which aspects of writing a story are considered more important in writing assessment?	10,585 assessment messages by 41 schoolteachers (19 secondary teachers and 22 primary) written on short stories composed by 393 students aged 10–15, from 14 Extremadura schools. Title stimulus: 'The boy who became the hero of his town'.	Stories distributed randomly amongst the 41 teachers for marking. Compositions analysed by the researchers using the marking criteria from two standardised writing tools: • PROESC (writing process test) ○ Content: 'Where and when'; 'characters'; 'event with consequences'; 'coherent ending'; 'creativity'. ○ Coherence: 'Logical continuity'; 'unity literary figures'; 'complex sentences'; 'vocabulary'. • RAS (rubric to assess stories): Organisation and content; grammatical aspects. Mixed dataset. The assessment notes were categorized according to code, place, extension, assessment content, and implicit meta-linguistic content. Frequency of assessment categories quantified and compared.Findings reveal the predominance of direct correction of specific spelling and grammar mistakes. The frequency of these corrections is, additionally, the best predictor of the global grade given by the teacher for the composition. Concludes that teachers of compulsory education approach assessment of narrative texts from a conception which places excessive emphasis on more local and superficial aspects of the composition in detriment of other semantic, rhetorical or pragmatic aspects. Some teachers, however, mostly in Secondary Education, also recorded non-corrective assessment content, such as markings, questions, suggestions for expansion, or justifications. Certain assessment patterns are evident, which combine other types of evaluation (semantic-organizational, or superficial).
Marcos, Fernandez, Gonzalez and Phillips-Silver (2020), Spain	Can students' creative thinking be enhanced through a structured program of reading and writing activities in the context of a cooperative learning classroom? Does a relationship exist between improvements in creative thinking and improvements in academic performance?	60 students aged 9–10 in Almería region, split equally into experimental and control groups Experimental group received a 7 week intervention involving a range of literary modes and activities. Control group taught using standardised textbook.	CREA test used as a pre and post-test measure of creative thinking for both groups. Academic achievement measure by grade-point average (GPA) with levels 1–5. Quantitative dataset. Tests for normality and Mann-Whitney U/ Wilcoxon rank tests run for the variables. The results revealed a significant increase in creativity scores in the experimental group as compared with the control, and a moderate positive correlation between creative thinking and academic achievement. The present findings are consistent with the idea that creative thinking (divergent thinking) can be enhanced with reading and writing activities implemented through cooperative learning in school-age children.
Montanero et al. (2014), Spain	How does iterative co-evaluation with a rubric improve narrative text production processes in primary school students?	128 students in Badajoz province, split into rubric group (62) and control group (66) 256 narratives (128 pre-intervention and 128 post-intervention) on theme of 'a boy that becomes a hero in	2 sessions. Students in rubric group co-evaluated and provided feedback on classmates' writing in both sessions. Control group received individualised corrections from the teacher.All writing then analysed using:

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
		his town'. 4 primary teachers	• PROESC standardised test (see Lucero et al., 2018, above); • A narration assessment rubric with reference to 7 items: setting, theme or initial event, plot (episodes and resolution), creativity and interest, sentences (readability, punctuation, construction), vocabulary and spelling. • By 4 teachers grading the writing on a 1–10 scale (blind/in a random order). The process of iterative co-evaluation with rubric resulted in a clear improvement in the organization and contents of the narrative texts (mainly in the description of the characters and in the story setting) while the students assessed by their teachers improved, significantly, their grammar and, above all, spelling mistakes.
Morris, Greve, Knowles and Huot (2015), USA	What are the consistent, ongoing themes in the books on writing assessment? What differences and/or similarities can we note about the books published on writing assessment by examining them diachronically? Can we confirm the largely positive, progressively advancing status for writing assessment available in the current literature about the field?	34 books devoted to writing assessment 4 coders (graduate students and scholars)	Books coded by the researchers with regular feedback and collaboration. The main categories were: subject, audience, authorship, purpose, and methodology. Findings tabulated to compare differences within the categories for books published before 2000 and after 2000. Notes a rise in theory and practice as a subject, in teachers/administrators as an audience, and in composition-based authorship. Suggests that writing assessment is a stable area of study with increased scholarly activity by new scholars, primarily composition authors writing for teachers and administrators.
Peterson (2012), Canada	What are the underlying beliefs and values about writing and learning to write that underpin writing curricula across the provinces of Canada, the Northwest Territories and the Yukon Territory?	8 curriculum documents, available online, from each province/territory.	Curriculum documents analysed using six key established writing discourse codes: skills; creativity; process; genre; social practices; and sociopolitical. Qualitative dataset. References to each discourse code tabulated by province/territory. The analysis showed that the process discourse predominates in all writing curricula. Elements of the skills, creativity and genre discourses are present with varying emphases across the provincial and territorial curricula. However, there is minimal to no evidence of the social practices and sociopolitical discourses.
Peterson (2003), Canada	How are social meanings constructed through peer interactions produced in students' revisions of their narrative writing?	4 Grade 8 students (2F; 2 M) selected for case study from 33 in the class. 1 Teacher Open writing task (no genre)	13 weeks with visits two to three times per week. Extensive field notes, drafts and final versions of the students' narrative writing, student interactions, and interviews. Qualitative analysis of the dataset with categories for each 'sequence' or interaction. These included codes for references to in-progress writing (e.g. characters, action, devices), who initiated the sequence, its purpose (e.g. asking for clarification, borrowing stationery, chatting about personal life), etc. These were tabulated against the corresponding foci and dates of the lesson sequence. Peer feedback in both settings influenced students' revisions at the word, sentence and organizational levels, although students made many more revisions to their writing than could be directly traced to the feedback. The peer talk that influenced the students' revisions served four functions: playing with ideas, clarifying ideas, questioning plausibility, and showing emotional response.
Peterson et al. (2006), Canada	How are the indicated corrections and the foci and modes of sixth-grade teachers' written feedback on student writing	108 Grade 6 teachers from 17 schools in Ontario 2 student narratives (about a dream) – typewritten	Teachers' comment tallied under two categories:

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
	similar and different when directed to two types of writing: narrative and persuasive? What are the similarities and differences in the feedback applied to writing attributed to male and female sixth-grade writers? What are the similarities and differences between male and female teachers' feedback on persuasive and narrative writing attributed to male and female student writers?	2 student persuasive pieces (on value of a First Nations reserve) – typewritten	• 'Focus': Meaning, conventions, organisation, artistic style, effort, process, ideology, and formatting. • 'Mode': correction and criticism, command, closed question, praise, open-ended question, reader response, lesson, explanation, suggestion. Comments also grouped by gender of writer/marker and text genre. Quantitative dataset (text codes established by a previous study). Two-way ANOVA and paired-sample t-tests to compare variables. Significantly higher number of corrections for narrative papers than for persuasive ones. Process, conventions, artistic style, and format were the focus of significantly greater numbers of comments directed to narrative writing. In contrast, meaning, organization, effort, and ideology were emphasized to a greater degree when teachers responded to per- suasive writing. Teachers tended to indicate and make greater numbers of corrections and to provide more criticisms and lessons, explanations, and suggestions when the work was attributed to a male writer. Female teachers generally wrote greater numbers of comments and tended to indicate and make more corrections. Generally, teachers were reluctant to engage with the ideologies in students' writing. There was a correlation between convention errors and the number and types of comments.
Rojas-Drummond et al. (2008), Mexico	How do primary school children 'learn to collaborate' and 'collaborate to learn' on creative writing projects by using diverse cultural artefacts (including oracy, literacy and IT)?	56 fourth-grade children from 1 primary school in Mexico City Multimedia stories produced in groups of 3	Video recordings of 5 sessions, with transcripts and creative products. Two groups selected at random for micro-analysis of the quality and nature of their interactions. Qualitative approach to analysis of videos and transcripts. Asserts value of an ethnographic approach to analysing creative process and identifies the dynamic functioning in educational settings of some central socio- cultural concepts. These include: co-construction; intertextuality and intercontextuality amongst oracy, literacy and uses of ICT; collaborative creativity; development of dialogical and text production strategies and appropriation of diverse cultural artefacts for knowledge construction.
Sandiford and Macken-Horarik (2020), Australia	What kind of framework is needed if we are to take account of positive but embryonic developments in student writing? How can the framework be used by teachers to foster next steps in student writing?	27 primary and secondary teachers 373 narratives by primary and secondary-age students	Semi-structured interviews with teachers and close study of 16 texts. 0–4 scale used to evaluate four dimensions of the texts: • Internal focalisation: resources that allow the reader to experience the world through a character's point of view. • Voicing: resources for creating a narrating voice and for inserting dialogue into the narration. • Attitude: resources for expressing feelings, judgments and appreciation of people, place and things. • Graduation: resources for adjusting the force (volume) and focus (sharpness of softness) of Attitude. Qualitative dataset. Using subset of 64 texts for close analysis, grammatical realisations of these 4 dimensions identified and tabulated.
Yeung et al. (2017), Hong Kong	What is the contribution of oral language and transcription skills to Chinese written composition in the context of other	97 students (57 F; 40 M) in Grade 4 (47) and grade 6 (50).	Written composition scored on 0–5 scale for the following criteria: content, vocabulary, sentence structure and organisation.Measures also gathered for:

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
	cognitive-linguistic skills amongst children in upper elementary grades?	"One day at school a funny/surprising thing happened…" prompt.	• General reasoning ability – Raven's Standard Progressive Matrices • Word spelling – Chinese characters spelling test • Handwriting fluency – Chinese Handwriting Speed Test (adapted) • Working memory – Chinese sentence memory task • Oral narrative skills – same as written composition task • Syntactic skills – Chinese word order knowledge task Quantitative dataset. Descriptive statistics and ANCOVA test, one-sample *t*-test, multivariate analysis conducted.Hierarchical multiple regression results showed that spelling and oral narrative skills were unique predictors of Chinese writing performance. The significant interaction effect of grade and spelling showed that transcription skills played a more important role in Chinese writing performance amongst sixth graders than amongst fourth graders. Together, the present results provide important support for the "simple view of writing" model and underscore the importance of transcription skills and oral narrative skills in children's writing development in Chinese.

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
Ashton and Davies (2015), USA	Does instructional guidance in the use of a rubric help students produce more reliable and valid assessments of creative writing samples? Will the ratings by students who used the guided rubric align more accurately with expert ratings than those by students who used the rubric alone?	7 'experts' (including literary agents, editors, and writers in the field of science fiction and fantasy) 'Rubric-plus' group (with guidance/training) of approximately 320 distance students 'Rubric-only' group (w/o guidance/training) of approximately 230 distance students 3 writing samples (advanced, intermediate and novice) for science fiction/fantasy examples.	Rubric with 5 items: overall quality linked to genre and intended audience; sentence variation; 'hook' (engagement established at the opening); 'blocking' (concerning spatial relationships between characters); and spelling. Quantitative ratings for the experts and the two student groups for comparison. Likert-style 5 point scale for ratings. 2 × 3 split factorial analysis to compare mean scores for three example stories. Mixed statistical results between groups. Ratings were only different for two of the rubric components between the two student groups, with the rubric-plus group slightly more aligned with the experts' ratings on one component. Central tendency rating error evident for both groups in comparison to the experts' ratings. Suggests peer review is more viable with training in using rubrics, ideally with exemplars of varying levels of skills showing differences in writing quality.
Broekkamp, Janssen and Denbergh (2009)), Holland	Does a relationship exist between students' literature reading and creative writing performance?	Independent 'expert' judges ($N = 8$ for writing experts, comprising 4 writers and 4 language teachers; $N = 7$ for reading judges, comprising 4 PhD students and 3 teachers) 19 students in the penultimate year of pre-university education. Mix of good and poor readers (performance and motivations established by teacher identification and/or participation in a previous related study) Short story; haiku poem; poem with 5 given words; Sensory flashback story; recipe for a bad mood.	Rating criteria consisting of 'task appropriateness', 'originality', 'technical quality' and 'personal engagement' amounting to final holistic rating for 'creative writing performance'. Quantitative dataset. Z-scores, Mann-Whitney U tests and effect size calculated. Multilevel estimates of variance and covariance to explore correlation between reading ability and writing quality. The results suggest that a positive relationship exists between literature reading and creative writing ability. Moreover, it shows that both constructs can be measured in reliable ways. Good literary readers outperformed the latter group (poor literary readers) in literature reading and in creative writing. A high correlation between literature reading ability and creative writing ability was found not only for the sample as a whole ($r = 0.82$), but also within the two "extreme groups" ($r = 0.79$).
Cremin et al. (2017), UK	What impact does writers' engagement with teachers have on teachers' identities as writers? What impact does writers' engagement with teachers have on teachers' pedagogic practices in the teaching of writing? What impact does writers' engagement with teachers have on student outcomes in the teachers' classes? What impact does writers' engagement with teachers have on teachers' efficacy in supporting students to develop motivation, confidence and writing skills? What impact does writers' engagement with teachers have on writers' own effectiveness in supporting teachers and young people to develop their motivation, confidence and writing skills?	32 class teachers and their classes (total of 711 students). Each teacher paired with a professional writer. Personal Narrative used for student testing in RCT	On the residential: peer and tutor feedback. For RCT: Test writing blind marked by trained assessment team according to national criteria. Mixed methods combining both quantitative and qualitative data: test data means, gain scores, standard deviations, ANCOVA and effect size calculated. Qualitative data comprised observations at the writers' residential and in classrooms, multiple interviews with writers and teachers, focus groups with students. Comparison group outperformed experimental group with a negative effect size of -0.34. Causal chain discussed and practicalities of writers as a source of authority in the classroom discussed as well as a range of qualitative outcomes, including the teachers' and writers' identities and the range of impacts on the students involved.
Dollinger (2003), USA	What is the relationship between the needs for Uniqueness and Cognition with past creative accomplishments, preference for	150 university students (88 F; 62 M) participating for course credit. Experts for the writing task: 1 creative writing instructor, 3	3 Likert-style questionnaires to provide measures for Uniqueness and Cognition; 7 tests for creativity also conducted and assessed using the consensual assessment technique, including:

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
	complex visual figures, unconventional word associations, vividness of dreams, and creative products?	psychologists Experts for the art task: 1 artist, 3 psychologists Experts for the photo task: 4 psychologists Experts for the dream reports: 3 psychologists	• Creativity behaviour inventory (self-reported accomplishments questionnaire) • Figure complexity preference task involving polygons • Word association responses to stimuli • Art task inviting students to complete an abstract incomplete drawing • Creative writing task responding to an image and initial sentence (rated on a 1–7 'very impoverished'/'very creative' scale) • Photo essay: students took/collected 20 photos that captured their identities • Recent dream recall task (written) Quantitative dataset. Coefficient alphas used for CAT assessed activities. Composite creativity score devised; descriptive statistics and correlation values generated to explore relationships within the data.Concludes that 'need for uniqueness' and 'need for cognition' are important predictors of creativity in young adults; found significant correlation between the predictors.
Form (2019), Germany	Is there a relationship between the popularity of bestsellers and their linguistic originality?	First 1000 words from 92 of the best-selling English books from 1813 (Pride and Prejudice) to 2012 (The Fault in Our Stars) identified from the *List of bestselling books* on Wikipedia. Mixed	Linguistic originality measured through online service MetaMetrics, which uses a logarithm to determine the mean log word frequency of texts with a length up to 1000 words. • Analysis run on first 1000 words of 92 texts (all sentences competed within first 1000 words) • To assess linguistic originality, average word frequency was inverted so that higher scores reflected higher originality. Log-transformation of three indicators of popularity: • Total copies sold (number of copies sold divided by years since publishing to average popularity over time) • Number of characters on text's description on Wikipedia • Number of languages the text was translated into besides English Quantitative dataset.Regression analyses performed to examine influence of linguistic originality on popularity of the best-sellers. Linguistic originality was a significant predictor of a best-seller's popularity ($r^2 = 0.09$, $p = 0.004$, $b = 0.30$). This effect even remained significant when controlling for year of first publishing and gender of author. Furthermore, the effect of originality on popularity was partially moderated by the year of publishing. Recommendation that creative writing teachers stress the features of style, especially original word use and foregrounding.
Furst, Ghisletta, & Lubart, 2017), Switzerland	How do patterns of the writing subprocesses, Generation (which broadly describes ideation, divergent thinking, gestation and experimentation) and Selection (critical evaluation, editing, cutting), vary in intensity and significance throughout the writing of a text? How do these subprocesses affect writing originality and quality?	174 first year undergraduate psychology students (151 F, 23 M). Open 3000-character writing task based on the theme of seasons (spring, summer, fall, winter). No style constraints. Explicit definition and explanation of Generation and Selection as constructs. Completed in 4 phases – preparation, draft writing, clarification and development, and correction. Approximately 6 min for each phase.	10 groups of randomly-assigned participants (9 experimental conditions and a control). Groups approached the writing task with different degrees of Generation and Selection at different phases, as instructed. As well as the writing task, participants completed personality questionnaires, questionnaires on creative writing habits, and vocabulary/verbal ability tasks. Texts rated by 4 groups of evaluators (3 groups of university students/trained research assistants equivalent to "quasi-experts"

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
	What are the optimal patterns to produce maximal creativity during a writing task?		and 1 group of researcher-raters) using 5-point Likert scale based on 6 items (coherence, quality, originality, surprise, creativity, aesthetics) to indicate broad scores for quality and originality. Quantitative dataset. Analysis of variance and multiple regression tests run. Results indicated that a decreasing level of the Generation subprocess and an increasing or constant high level of Selection was the optimal pattern for creativity. Verbal fluency was also found to be positively related to the Quality of the writing and creative writing habits positively related to Originality. No experimental group outperformed the control group.
Gulley (2012), USA	What is the effect of oral feedback delivered via student-teacher conferences on significant revisions to content, structure, grammar, and style for developmental writing students?	70 community college developmental writing students. 2 outside evaluators to rate the original and revised scripts (blinded). Narrative paragraph.	Form including the following categories: thesis statement; content; organisation; sensory details; grammar; and style. Students randomly assigned to one of three groups for three different types of feedback: • Oral feedback only (conference session) • Written feedback only (email) • Both oral and written feedback (conference) Raters used the 2008 Kansas State Assessment scoring guide (with scores for ideas and content; organisation; voice; word choice; sentence fluency and conventions).Quantitative dataset. Descriptive statistics and mixed design ANCOVA for the raters' scores and Editor programme computer scans (for identifying errors).No statistically significant difference amongst treatment groups. Concludes that students improved their drafts regardless of the feedback method.
Kaufman et al. (2009), USA	Who exactly qualifies as an expert to evaluate a creative product such as a short story?	Group 1: 10 published writers (7 F; 3 M. 80% with advanced degrees) all with experience reading the work of student writers. Group 2: 106 university students (81 F; 25 M) participating for course credit. 203 samples of writing from university students (for extra credit) responding to two short story prompt titles within 10 min and submitting online.	Consensual assessment technique: both groups rated the stories for their creativity. Quantitative dataset. Independent means *t*-test to compare average ratings between the two groups; Pearson correlation between the two sets of ratings; Coefficient alpha and Spearman-Brown formula applied to measure the extent of consensus within the groups. Results revealed a small but statistically significant difference and a limited correlation between experts' and novices' mean ratings. Novices also showed low levels of inter-rater reliability and less consistency in their ratings. Expert raters overall tended to rate lower for creativity than the novice raters.
Kelly et al. (2003), USA	Is writing style related to readers' assessments of a story in terms of its interestingness, informativeness, dullness and other story characteristics? Is story subject matter related to readers' assessments of story characteristics?	117 undergraduate students enroled in two introductory communication courses. Mix of academic majors and ranged in age from 18 to 26 (58 F, 57 M) Four identically typeset news stories on crime and on the environment (one 'narrative' and one 'straight-news' story for each).	7-point bipolar rating scales measuring: • 'Story interest' in terms of dull/interesting and unenjoyable/enjoyable. • 'Informativeness' in terms of uninformative/informative and unclear/clear. • 'Credibility' in terms of unbelievable/believable, subjective/objective and inaccurate/accurate. • A 'passive/active' scale also included. Quantitative dataset.16 T-tests; One-way analysis of variance. Those who read the narrative crime story found that version clearer and more active than did those who read the straight-

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			news version. Those who read the environmental narrative story rated their version more informative, clear and believable than did those who read the environmental straight- news story. The subjects rated the environmental stories as less interesting than the crime stories and rated the narrative stories as less interesting than the straight-news stories. However, they rated the environmental stories and the narrative stones as more informative, more accurate and more believable than the crime stories and the straight-news stories. Those who read the straight-news version of the crime story had a higher assessment of their personal risk of homicide than did those who read the narrative version of the crime story.
Ng and Yeung (2011), Hong Kong	Does incorporating a multi-sensory approach to teaching writing improve students' writing scores in three different genres of Chinese essays?	100 first-year associate degree students (80 F, 20 M), all native Cantonese speakers. 1 teacher 600 essays (100 pre-intervention essays in descriptive, narrative and expressive genres; 100 post-intervention essays in the same genres).	1–6 weak/strong scale for narrative criteria, detailing ten features: logical sequencing of time; concrete and clear setting; vivid characters; clearly structured storyline; interesting and appealing ending; appropriate and precise choice of content; well-organised clues; clear and focused central theme; overall appropriateness to the topic; and author's ability to influence the reader's feelings. Separate marking criteria for the descriptive and expressive genres. Quantitative dataset produced from pretest-posttest intervention design. Texts were rated by the student writers themselves, a peer, and the teacher before and after the intervention. A 2 (time: pretest vs. posttest) x 3 (genre: descriptive, narrative, expressive) x 3 (assessor: self, peer, teacher) repeated-measures analysis of variance conducted on SPSS. General pattern of higher posttest than pretest scores, particularly in the scores given by the teacher despite marking both sets of writing in one randomised sequence. Gain scores comparatively smaller for the narrative genre (7%) than for the descriptive (10%) and expressive (18%) writing. For both sets of results, mean scores showed that in general self-ratings tended to be higher than peer ratings, which in turn tended to be higher than teacher ratings, with the notable exception of expressive writing after intervention by the peer student raters. Inconsistency between teachers' and students' own ratings was most marked in the scores for the narrative genre, the genre they should in theory have had most practice with throughout school.
Prescott (2012), UK	What kinds of learning come to the fore when students engage in life writing as part of a tightly-structured programme designed around academic objectives?	7 distance learners based in West Midlands of England 7 pieces of life writing in the form of narrative and poetry, with accompanying 500-word reflections.	Reflective commentaries and questionnaire with one open-ended question for the students to complete. Mixed dataset. Close reading of responses focused on cognitive and affective learning. Overlap between cognitive and affective learning outlined alongside the assertion that only cognitive learning is amenable to assessment.
	What factors do students identify as connected with negative feeling toward instructor feedback or shutdown in their ability	E-participation in the survey by college students (343 total/ 212 fully completed responses).	Self-report survey with 4 open questions and 17 closed responses. Open recruitment using student lists of a number of university

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
Taggart and Laughlin (2017), USA	to use that feedback for revision? Do students feel pressured by instructor feedback to craft texts that do not reflect their values or intentions, but rather conform to meet the instructor's expectations? If so, when? Are there typical types of feedback or response scenarios that students report leading to these negative affective responses more often than others?		and sharing through social media. Mixed dataset. Coding of 3 of the open responses based on affect theory, with attention to trigger phrases and emotion-related words. Descriptive statistics for quantitative dataset. Highlights importance of the instructor's role in affecting revision. 20.7% of responses also stated that self-evaluation played an important role. Outlines significant role of the teacher-student relationship, the negative effects of feedback that focuses on the individual rather than the text produced, and tensions within the feedback-response process, including 'hierarchy responses' (describing conflicting agendas and desires between students and instructors/institutions). 'Disrespect' was the most frequent emotional code, followed by 'frustration'. Impacts on learning discussed.
Vaezi and Rezaei (2018), Iran	How can a generic creative writing more accurately measure the quality of fiction works?	Phase 1: 10 creative writing experts (professors of English/Persian literature, editors of literary journals, creative writers who run workshops). Phase 2: 18 creative writing professors at universities in Iran and USA. Phase 3: 19 samples of university-level creative writing from workshops and competitions to be rated by two professors independently, after some guidance.	5-point Likert scale questionnaire (extremely important/extremely unimportant for different rubric elements); multiple unstructured interviews. 9 elements eventually identified: narrative voice, characterisation, mood and atmosphere, language and writing mechanics, story, setting, image, dialogue, and plot. Ratings collected on samples in Phase 3. Modified Delphi Technique. Largely quantitative ratings, supplemented by interviews and written comments. Correlation coefficient to measure consistency between raters in Phase 3. 88% of sample satisfied with quality of the rubric produced and >70% consistency between the two raters in Phase 3.
Zedelius et al. (2019), USA	If every piece of writing is unique, can we nonetheless compare different pieces and evaluate them according to some objective standard, or is creativity purely subjective?	Study 1: 6 independent raters and 133 undergraduate students (88 F, 44 M, 1NB). All of the participants majored in psychology and participated in the study in exchange for course credit. Study 2: 5 independent raters and 128 undergraduate students (83 F, 44 M, and one who did not indicate gender; mean age = 20.1 years, SD = 3.8). The students participated in exchange for money. The major difference in Study 2 was that we made a specific effort to recruit a more heterogeneous sample of participants, including students with an interest in creative writing. No overlapping participants between studies.	Evaluation rubric correlated with established creativity measures, including Alternate Uses Task (AUT); Compound remote associates (CRA) problems; Creative Behaviour Inventory (CBI) short self-report form. The rubric used by the independent raters detailed aspects and examples for three broad qualities desired in narrative writing: • Image: vivid sensory detail; rich, concrete details; figurative devices/tropes such as metaphor and personification; thinking or speaking voices if appropriate; and lacking elements of 'flat' writing (generalisations and vague abstractions). • Voice: visible and distinctive style including choice of words and application of vocabulary in refreshing ways; mix of sentence structures; use of metaphor or comparison; and use of page layout, punctuation and italics/capitals. Furthermore, command of humour, 'darkness', control over simplicity and/or complexity, intentional choice of narration and different narrative perspectives or tones for a particular effect. • Originality: unusual or unexpected choices of material for the storyline or details within the story (e.g. characters and their attributes, or setting) including conceptual or philosophical departure from conventional approaches to the task. 1–3 scale (poor, fair or very good) based on the combination of these three story attributes.Quantitative dataset.Coh-Metrix to

(*continued on next page*)

(*continued*)

Author, Year, Country	Research questions/focus	Sample, Text, Genre	Method used to assess writing, Type of data produced, Data analysis and Summary of findings
			analyse metrics of text cohesion and readability (predicts story evaluations through computing narrativity, deep cohesion, referential cohesion, syntactic simplicity, and word concreteness); and LIWC (Linguistic Inquiry/Word Count) to identify categories of the text content. Correlation between the features also explored and interrater reliability established.Linguistic features predicted the human ratings of creativity to a significant degree.These results establish the evaluation rubric as a useful tool to assess creative writing provides evidence that at least some aspects of creative writing can be captured by computerized measures.

References

Alhusaini, A. A., & Maker, C. J. (2015). Creativity in students' writing of open-ended stories across ethnic, gender, and grade groups: An extension study from third to fifth grades. *Gifted and Talented International, 30*(1–2), 25–38.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013.

Amabile, T. M. (1983). *The social psychology of creativity*. New York: Springer-Verlag.

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Boulder, CO: Westview.

Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education, 36*(3), 312–334.

Baer, J. (1993). *Creativity and divergent thinking: A task-specific approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal, 10*, 25–31.

Baer, J. (1998). The case for domain specificity in creativity. *Creativity Research Journal, 11*, 173–177.

Baer, J., & Kaufman, J. C. (2019). Assessing creativity with the consensual assessment technique. In I. Lebuda, & V. Glaveanu (Eds.), *The palgrave handbook of social creativity research*. Palgrave Macmillan, Cham: Palgrave Studies in Creativity and Culture.

Baer, J., Kaufman, J. C., & Gentile, C. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*(1), 113–117.

Baer, J., & McKool, S. S. (2014). The gold standard for assessing creativity. *International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE), 3*(1), 81–93.

Beyreli, L., & Ari, G. (2009). The use of analytic rubric in the assessment of writing performance–inter-rater concordance study. *Educational Sciences: Theory and Practice, 9*(1), 105–125.

Bingham, H. (2014). *The writers' & artists' yearbook guide to how to write: The essential guide for authors*. London: Bloomsbury.

Bintz, W., & Shake, M. (2005). From university to classrooms: A preservice teachers' writing portfolio program and its impact on instruction in teaching strategies for writing portfolios in the classroom. *Reading Horizons, 45*(3), 217–233.

Broekkamp, H., Janssen, T., & Denbergh, H. V. (2009). Is there a relationship between literature reading and creative writing? *Journal of Creative Behavior, 43*(4), 281–297.

Carlson, R. K. (1965). An originality story scale. *The Elementary School Journal, 65*, 366–374, 7.

Cheung, W. M., Tse, S. K., & Tsang, W. H. H. (2001). Development and validation of the Chinese creative writing scale for primary school students in Hong Kong. *Journal of Creative Behavior, 35*(4), 249–260.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Craft, A. (2000). *Teaching creativity: Philosophy and practice*. London and New York: Routledge.

Craft, A. (2005). *Creativity in schools: Tensions and dilemmas*. Abingdon: Routledge.

Craft, A., Gardner, H., & Claxton, G. (2008). *Creativity, wisdom and trusteeship*. Thousand Oaks, CA: Corwin Press.

Cremin, D., & Myhill, D. (2012). *Writing voices: Creating communities of writers*. Abingdon: Routledge.

Cremin, T., Myhill, D., Eyres, I., Nash, T., Wilson, A., & Oliver, L. (2017). Teachers as writers. Available at http://www.teachersaswriters.org/wp-content/uploads/2017/12/Teachers-as- Writers- Research-Report-2017-FINAL-.pdf. Accessed 08 January 2018.

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 159–166.

Csikszentmihalyi, M. (1997). *Creativity: The psychology of discovery and invention*. New York: HarperCollins.

deBenedictus, D. C. (2009). What? Write worse: Assessing student writing from both ends of the continuum. *Ohio Journal of English Language Arts, 49*(1), 39–48.

Diercks-Gransee, B., Weissenburger, J. W., Johnson, C. L., & Christensen, P. (2009). Curriculum-based measures of writing for high school students. *Remedial and Special Education, 30*(6), 360–371.

Dockrell, J. E., Connelly, V., Walter, K., & Critten, S. (2015). Assessing children's writing products: The role of curriculum based measures. *British Educational Research Journal, 41*(4), 575–595.

Dollinger, S. J. (2003). Need for uniqueness, need for cognition, and creativity. *Journal of Creative Behavior, 37*(2), 99–116.

Duran, E., & Yilmaz, A. (2019). The evaluation of the fourth grades' writing story skill. *International Online Journal of Educational Sciences, 11*(3), 194–206.

Dymoke, S. (2003). *Drafting and assessing poetry: A guide for teachers*. London: SAGE.

Emig, J. (1988). *The composing processes of twelfth graders*. Urbana, Illinois: National Council for the Teaching of English.

Form, S. (2019). Reaching wuthering heights with brave new words: The influence of originality of words on the success of outstanding best-sellers. *Journal of Creative Behavior, 53*(4), 508–518.

Furst, G., Ghisletta, P., & Lubart, T. (2017). An experimental study of the creative process in writing. *Psychology of Aesthetics Creativity and the Arts, 11*(2), 202–215.

Gardner, P. (2012). Paradigms and pedagogy: Revisiting D`Arcy's critique of the teaching and the assessment of writing. *English in Education, 46*(2), 135–154.

Gowan, J. C. (1965). Book review: Torrance, E. Paul. Rewarding creative behavior. Englewood cliffs, N. J.: prentice hall co. (354 pages), 1965. *Gifted Child Quarterly, 9*(2), 113–113.

Guilford, J. P. (1967). *The nature of human intelligence*. NY: McGraw-Hill.

Gulley, B. (2012). Feedback on developmental writing students' first drafts. *Journal of Developmental Education, 36*(1), 16–36.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Humphry, S., & Heldsinger, S. (2019). Raters' perceptions of assessment criteria relevance. *Assessing Writing, 41*, 1–13.

Johnson, D. (2003). Activity theory, mediated action and literacy: Assessing how children make meaning in multiple modes. *Assessment in Education: Principles, Policy & Practice, 10*(1), 103–129.

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *Journal of Creative Behavior, 43*(4), 223–233.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*(2), 171–178.

Kaufman, J. C., Baer, J., & Gentile, C. (2004). Differences in gender and ethnicity as measured by ratings of three writing tasks. *Journal of Creative Behavior, 38*(1), 56–69.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology, 13*(1), 1–12.

Kaufman, J. C., Gentile, C., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly, 49*(3), 260–265.

Kelly, J., Knight, J., Peck, L. A., & Reel, G. (2003). Straight/narrative? Writing style changes readers' perceptions of story quality. *Newspaper Research Journal, 24*(4), 118–122.

Kettler, T., & Bower, J. (2017). Measuring creative capacity in gifted students: Comparing teacher ratings and student products. *Gifted Child Quarterly, 61*(4), 290–299.

Lamott, A. (1995). *Bird by bird: Instructions on writing and life*. New York: Anchor Books.

Leong, S. (2010). Creativity and assessment in Chinese arts education: Perspectives of Hong Kong students. *Research Studies in Music Education, 32*(1), 75–92.

Leong, S. (2011). Creativity and the arts in Chinese societies'. In J. Sefton-Green, P. Thomson, K. Jones, & L. Bresler (Eds.), *The routledge international handbook of creative learning (1st ed.)* (pp. 54–62). Abingdon: Routledge. https://doi.org/10.4324/9780203817568.

Lucero, M., Fernández, M. J., & Montanero, M. (2018). Teachers' written feedback comments on narrative texts in elementary and secondary education. *Studies in Educational Evaluation, 59*, 158–167.

Magrs, P., Lodge, D., Jones, R. C., Bell, J., Bell, J., & Magrs, P. (2001). Stepping back'. *The creative writing coursebook* (pp. 3–19). London: MacMillan, 233-250.

Marcos, R. I. S., Fernandez, V. L., Gonzalez, M. T. D., & Phillips-Silver, J. (2020). Promoting children's creative thinking through reading and writing in a cooperative learning classroom. *Thinking Skills and Creativity, 6*, 1–13.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), 1–6. Jul 21.

Montanero, M., Lucero, M., & Fernández, M. J. (2014). Iterative co-evaluation with a rubric of narrative texts in primary education /coevaluación iterativa con rúbrica de textos narrativos en la educación primaria. *Infancia y Aprendizaje, 37*(1), 184–220.

Morris, W., Greve, C., Knowles, E., & Huot, B. (2015). An analysis of writing assessment books published before and after the year 2000, Teaching English in the Two Year College. 43:2 118–140.

Mozzafari, H. (2013). An analytical rubric for assessing creativity in creative writing. *Theory and Practice in Language Studies, 3*(12), 2214–2219.

Newman, J. (2007). The evaluation of creative writing at MA level (UK). In S. Earnshaw (Ed.), *The handbook of creative writing.* (pp. 24–36).

Newman, M., & Gough, D. (2020). Systematic reviews in educational research: Methodology, perspectives and application. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic reviews in educational research*. Wiesbaden: Springer VS.

Ng, P. C., & Yeung, A. S. (2011). A multi-sensory approach to enhancing Chinese writing. *International Journal of Pedagogies and Learning, 6*(3), 206–218.

Peterson, S. S. (2003). Peer response and students' revisions of their narrative writing. *L1-Educational Studies in Language & Literature, 3*(3), 239–272.

Peterson, S. S. (2012). An analysis of discourses of writing and writing instruction in curricula across Canada. *Curriculum Inquiry, 42*(2), 260–284.

Peterson, S. S., Childs, R., & Kennedy, K. (2006). Sixth-grade teachers' written comments on student writing. *Written Communication, 23*(1), 36–62.

Prescott, L. (2012). Life writing and life-learning: An analysis of creative writing students' work. *Studies in Continuing Education, 34*(2), 145–157.

Proctor, R. M. J., & Burnett, P. C. (2004). Measuring cogni- tive and dispositional characteristics of creativity in elementary students. *Creativity Research Journal, 16*, 421–429.

Renzulli, J. S., & Reis, S. M. (1991). The assessment of creative prod- ucts in programs for gifted and talented students. *Gifted Child Quarterly, 35*, 128–134. https://doi.org/10.1177/001698629103500304.

Rojas-Drummond, S. M., Albarrán, C. D., & Littleton, K. (2008). Collaboration, creativity and the co-construction of oral and written texts. *Thinking Skills and Creativity, 3*(3), 177–191.

Runco, M. A. (1989). The creativity of children's art. *Child Study Journal, 19*, 177–190.

Runco, M. A. (2014). *Creativity: Theories and themes: Research, development, and practice* (2nd Edition). London: Elsevier.

Sandelowski, M., Voils, C. I., Leeman, J., & Crandell, J. L. (2012). Mapping the mixed methods-mixed research synthesis terrain. *Journal of Mixed Methods Research, 6* (4), 317–331.

Sandiford, C., & Macken-Horarik, M. (2020). Changing stories: Linguistically-informed assessment of development in narrative writing. *Assessing Writing, 45*, 1–12. N.PAG-N.PAG.

Sharples, M. (1999). *How we write: Writing as creative design*. London: Routledge.

Sternberg, R. J. (1988). *The nature of creativity: Contemporary psychological perspectives*. New York: Cambridge University Press.

Sternberg, R. J. (2009). 'Foreword. In S. C. Kaufman, & J. C. Kaufman (Eds.), *The psychology of creative writing* (pp. XV–XVII). New York: Cambridge University Press, 2009.

Sternberg, R. J., & Lubart, T. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York, NY: Free Press.

Taggart, A. R., & Laughlin, M. (2017). Affect matters: When writing feedback leads to negative feeling. *International Journal for the Scholarship of Teaching & Learning, 11*(2), 1–11.

Torrance, E. P. (1965). *Rewarding creative behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Turkman, B., & Runco, M. A. (2019). Discovering the creativity of written works: The keywords study. *Gifted and Talented International, 34*, 1–2, 19-29.

Vaezi, M., & Rezaei, S. (2019). Development of a rubric for evaluating creative writing: A multi-phase research. *New Writing-the International Journal for the Practice and Theory of Creative Writing, 16*(3), 303–317.

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for K-12 classrooms*. West Palm Beach, FL: Learning Sciences International.

Wilson, A. C. (2010). Teachers' conceptualisations of the intuitive and the intentional in poetry composition. *English Teaching: Practice and Critique, 9*(3), 53–74.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 1–14. https://doi.org/10.3389/fpsyg.2019.03087. Article 3087.

Yeung, P.-. S., Ho, C. S., Chan, D. W., & Chung, K. K. (2017). The role of transcription skills and oral language skills in Chinese writing among children in upper elementary grades. *Applied Psycholinguistics, 38*(1), 211–231.

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods, 51*(2), 879–894.