Contents lists available at ScienceDirect

Assessing Writing

An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing

Sandra Dunsmuir*, Maria Kyriacou, Su Batuwitage, Emily Hinson, Victoria Ingram, Siobhan O'Sullivan

Department of Clinical, Educational and Health Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 12 August 2014
Accepted 18 August 2014
Available online 29 September 2014

Keywords:
Writing
Assessment
Written expression
Literacy

ABSTRACT

The study evaluated the reliability and validity of the Writing Assessment Measure (WAM), developed to reflect the skills which children of different abilities are expected to achieve in written expression, as part of the National Curriculum guidelines in England and Wales. The focus was on its potential use in investigations of children's written narrative in order to inform and target related interventions. The study involved 97 children aged 7–11 from one urban primary school in England. Prompt 1 was administered to all the children in their classrooms together with a standardised written expression test. After three weeks, the same procedure was followed and Prompt 2 was administered. Statistical analyses of the reliability and validity of the instrument showed that it is consistent over time and can be scored reliably by different raters. Content validity of the instrument was demonstrated through inspection of item total correlations which were all significant. Analyses for concurrent validity showed that the instrument correlates significantly with the Wechsler Written Expressive Language sub-test. Significant differences between children of different age and writing skill were also found. The findings indicate that the instrument has potential utility to professionals assessing children's writing.

* Corresponding author. Tel.: +44 20 7679 5307.
E-mail address: s.dunsmuir@ucl.ac.uk (S. Dunsmuir).

http://dx.doi.org/10.1016/j.asw.2014.08.001

1. Introduction

Writing is an essential skill that allows people to participate fully in today's society and to contribute to the economy. It is a complex process that is essential for extending learning, thinking and communicating with others (Dunsmuir & Clifford, 2003; Williams, 2000). By the time they reach eight to nine years of age, children in England have had four years of formal instruction in writing in school and many also have experience of writing at home, to variable degrees (Ruttle, 2004).

The assessment of writing is central to the process of effective teaching and learning of writing (Jones, 2002). When done in a robust manner, writing assessment can support teaching, both conceptually and practically (White, 1985). In this way, exploring assessment alternatives that are better suited to the developmental goals set, helps to identify where children are, in terms of their writing development (Herrington & Curtis, 2003). This, in turn, can enable greater understanding of the requirements needed to support their learning.

However, the assessment of writing is problematic and is considered to be the single most significant obstacle to practical progress in writing instruction and research (Cole, Haley, & Muenz, 1997). Researchers have struggled with the development of methods that are able to produce a valid and reliable means of assessing narrative writing (Meier, Rich, & Cady, 2006; Rezaei & Lovorn, 2010). The aim of this study was to develop a valid and reliable writing assessment measure, relevant within the context of the British educational system, for use by professionals to identify students' specific needs in relation to their writing development and contribute to planning subsequent teaching in a targeted way.

2. Literature review

2.1. Approaches to psychological and educational assessment of written expression

The introduction of the National Literacy Strategy in 1998 resulted in dramatic changes to the teaching of written narrative in the English and Welsh Education system. Attainment targets were introduced to set out the 'knowledge, skills and understanding which pupils of different understanding and abilities are expected to have by the end of each key stage' (as defined by the Education Act, 2011). In this context, 'knowledge' refers to the writing strategies/techniques that the child has learned, whilst 'understanding' refers to the awareness of when to use these strategies/techniques in their writing. Attainment targets for writing consisted of level descriptors of increasing difficulty, ranging from Level 1 to Level 8 (see the British Education Act, 2011). Each level descriptor explains the types and range of performance that pupils working at that level should characteristically demonstrate. At the end of Key Stage 2 (Years 3–6, ages 7–11), the majority of pupils are expected to attain Level 4 (Qualifications and Curriculum Development Agency, 2010). As the focus of this study is on the assessment of children's writing from Year 3 to Year 6, the Level 4 descriptor for writing is of interest and is detailed below:

> 'Pupils' writing in a range of forms is lively and thoughtful. Ideas are often sustained and developed in interesting ways and organised appropriately for purpose. Vocabulary choices are often adventurous and words are used for effect. Pupils are beginning to use grammatically complex sentences, extending meaning. Spelling, including that of polysyllabic words that conform to regular patterns, is generally accurate. Full stops, capital letters and question marks are used correctly, and pupils are beginning to use punctuation within sentences. Handwriting style is fluent, joined and legible' (Qualifications and Curriculum Development Agency, 2010: 18).

Therefore, the key domains of writing within the National Curriculum (defined through a government-led consultation with academics and professionals) reflect a focus on ideas development (rhetorical skills), vocabulary, sentence structure and grammar (writing processes), spelling, punctuation and handwriting (mechanics). To determine whether pupils have achieved these skills, writing ability is assessed at the end of Key Stage 2 (at the age of 7) as part of the statutory Standard Assessment Tasks (SATs) in English. Each pupil is awarded an achievement level, based on a holistic assessment system, the most commonly used method to evaluate writing. This involves assignation of a single overall score or rating, based on a set of pre-determined criteria. Holistic methods can therefore be used to compare abilities of groups of children and provide a ranked, overall rating of the quality of a piece of writing. In this respect, National Curriculum descriptors have defined the construct of writing within the WAM.

However, the validity and reliability of holistic scoring procedures have been questioned by a number of researchers (Espin, Weissenburger, & Benson, 2004; Hayes, Hatch, & Silk, 2000; Miller & Crocker, 1990). The major disadvantage of holistic scoring emerges from the limitations of the single general score which can give ranking information but no details. Whilst holistic scoring is considered more economical than analytic scoring (as the assessors are required to give a single score for each writing sample), it can only provide limited diagnostic information about a person's writing ability. This is because a single score does not allow raters to distinguish between various aspects of writing such as syntax, depth of vocabulary, organisation, and so on. Therefore, the same holistic score assigned to two different writing scripts may represent two entirely different sets of characteristics, even if the raters' scores reflect a consistent application of the rubric.

In contrast to a holistic method of scoring, an analytic scoring system is preferred over holistic schemes by many writing specialists for a number of reasons. It has been generally identified as a more reliable approach to writing assessment because of a clearly defined, objective and detailed scoring system (Hayes et al., 2000), and is beneficial in informing interventions (Gregg & Mather, 2002; Hooper et al., 1994). It focuses on identified qualities of good writing and is judged on how many elements of good writing it contains (Huot, 1990), thereby providing more useful diagnostic information about children's written narrative skills. In other words, it provides more insight into the strengths and weaknesses of students, enabling practitioners to tailor instruction more closely to the needs of their students. Additionally, analytic scoring has been shown to be useful for second language learners, who are more likely to convey an uneven profile across various aspects of writing (McNamara, 1996; Shaw & Weir, 2007). For instance, some second language learners may have excellent writing skills in terms of content and organisation, but may have poor grammatical skills; others may have an excellent control of sentence structure, but may not know how to produce a coherent text.

Moreover, because of this method's explicit criteria in separate components, it has been proven easier to train raters to use analytic scales than to train raters to use holistic rubrics (Cohen, 1994; McNamara, 1996). Less experienced raters may find it easier to work with an analytic scale than a holistic rubric because they can assess specific textual features. Finally, the explicitness of analytic scoring guides offers practitioners a potentially valuable tool for providing writers with specific feedback which is associated with improved writing performance in the future (Fathman & Whalley, 1990).

The Office for Standards in Education (OFSTED, 1999, 2011) reported the limitations of SATs writing assessments and from 2012, teacher assessment replaced more formal tests of writing for 11 year old children (Key Stage 2 SATs). This was in response to widespread complaints by teachers about the difficulties with holistic assessment of a writing sample produced in controlled conditions.

To the best of our knowledge, there are a very limited number of tests with norms derived from populations of children residing in the United Kingdom (UK) available for assessing narrative writing. The purchase and use of certain tests that do exist (e.g. Wechsler batteries being the most prevalent ones), is restricted to individuals with specific training and experience, complying with ethical and professional standards of competence defined by the British Psychological Society (BPS, 2009, 2011). Moreover, only one sub-test is designed to assess written expression within the Wechsler tests which are designed to assess general linguistic and/or numeric competencies more broadly.

There have been two editions of the Wechsler tests. The previous version, the Wechsler Objective Language Dimensions (WOLD; Wechsler, 1996) is a UK standardisation of the Wechsler Individual Achievement Test (WIAT-R), developed in the United States (US). This assessment is an individually administered test of expressive language skill in children aged 8–16 years. It includes three separate components; listening comprehension, oral expression and written expression. The assessment was standardised on 400 children throughout the UK on the basis of a stratified sample according to national demographics available from the Government. The updated version, the WIAT-II^{UK} (Wechsler,

2005) was standardised on 892 children aged 4–17 years in the UK, during the Wechsler Intelligence Scale for Children (WISC-IV^{UK}) in 2004 (Wechsler, 2004). The WIAT-II^{UK} is an individually administered test designed to assess the achievement of children and adolescents aged between 4 years and 16 years 11 months. The basic content domains included are reading, mathematics, written language and oral language, with 9 subtests in total. Both the WOLD and the WIAT-II^{UK} use both an analytic scoring system to assess written expression. The WOLD analytic scoring on written expression involves the evaluation of six elements: (1) organisation, (2) unity and coherence, (3) vocabulary, (4) grammar and usage, (5) capitalisation and (6) punctuation. Individual scores for each domain are computed and then collated into a total score. The WIAT-II^{UK} uses a more restricted analytic scoring for narrative pieces like story-writing focusing on mechanics, organisation (sentence structure, paragraphing), and vocabulary (variation of words, any expressions that capture the reader's interest). Due to its more detailed analytic scoring, the WOLD is still widely used as a research tool with normally developed children and children with specific language impairments (Williams & Larkin, 2013; Williams, Larkin, & Blaggan, 2013).

According to the WOLD manual (Wechsler, 1996) the subtests were designed to reflect aims and objectives of the classroom which increases the possibility of matching teaching practices to testing. However, the appropriateness of the Wechsler Written Expression subtests has been questioned, as the content utilises a range of criteria which bear little semblance to National Curriculum attainment targets and descriptors. Furthermore, handwriting is not included within the scoring criteria, despite the fact that surveys estimate that teachers consider 13.9% of pupils to have difficulties with handwriting (Barnett, Stainthorp, Henderson, & Scheib, 2006). Such difficulties may contribute to additional compositional weaknesses that influence the quality of the writing product.

A large literature, which includes both correlational and experimental methods, supports that difficulties with handwriting are associated with higher level aspects of writing such as the quality and fluency of written expression (see Alvès, Castro, de Sousa, & Strömqvist, 2007; Chanquoy & Alamargot, 2002; Connelly, Campbell, MacLean, & Barnes, 2006; Connelly, Dockrell, & Barnett, 2005; Connelly, Dockrell, & Barnett, 2011; Gregg & Mather, 2002; McCutchen, 2006; Olive & Kellogg, 2002; Peverly, 2006; Torrance & Galbraith, 2006). According to this research, an individual who is fluent at handwriting has greater attentional capacity to devote to planning and composing when compared to an individual who has poor handwriting skills and must devote attentional resources to this aspect of writing. Handwriting fluency is therefore related to the compositional aspects of narrative writing and needs to be considered within a comprehensive assessment of written expression to inform intervention planning.

To summarise, writing is a dynamic process of dealing with an excessive number of simultaneous demands or constraints (Flower & Hayes, 1980). Viewed this way, a writer in the process is a thinker on full-time cognitive overload (p. 33, cited by Torrance & Galbraith, 2006). To assess written expression skill, the need for a valid and reliable writing assessment is of paramount importance. This can be provided for by the development of scoring rubrics and methods that define performance criteria on written expression, to better inform practice and learning.

2.2. The development of a writing assessment measure: purpose and context of the assessment

The aim of this study was to develop and evaluate a new, complete test of written expression, which incorporates elements from the National Curriculum attainment targets, to ensure relevance of assessment data to the context and curriculum to which children in the UK are exposed. Hence, the written expression domains targeted are ideas development, vocabulary, sentence structure and grammar, spelling, punctuation and handwriting. As discussed in the previous section, analytic scoring systems have been shown to be more reliable that other procedures and beneficial in informing interventions. This method was therefore selected as a framework to present scoring criteria to assess domains of children's writing within the 'Writing Assessment Measure' (WAM) (see Appendix A for a copy of the elements and scoring criteria).

In the area of writing assessment, researchers have "...struggled with the development of methods able to produce a reliable and valid means of directly assessing writing quality" (Huot, 1990: 237). Evaluating the reliability and validity of any assessment measure is necessary and fundamental component for its future effectiveness. Emphasising the paramount importance of developing and implementing writing assessments that reliably measure written narrative skill Muenz, Ouchi, and Cole (1999) argue that "...examiners should not feel comfortable assessing an individual's writing ability using tests with poor reliability and low validity" (1999: 31). This study therefore sought to establish the reliability and validity of the WAM.

Weigle defines reliability as "...consistency of measurement across different characteristics or facets of a testing situation such as different prompts or raters" (2002: 49). With regard to measuring the reliability of an instrument, the internal consistency can be examined using Cronbach's alpha, and test-retest and inter-rater reliabilities can also be estimated. Much of the early research into writing assessment focused on establishing reliability. Primarily, this research focused on agreement between raters as this was seen as the area with the greatest potential to reduce reliability (Huot, 1990). Inter-rater reliability refers to the degree which independent raters can agree in assessing the quality of a single writing sample and is particularly relevant for tests of written expression which require rater judgement (Huot, 1990). Tindal and Parker (1991) suggest that clear and standardised administration and inter-rater reliability are necessary in order for others to unambiguously interpret test results. Studies have shown that with both training and scoring guides, acceptable levels of inter-rater agreement is largely possible (Penny, Johnson, & Gordon, 2000). Muenz et al. (1999) investigated whether rater judgements in certain domains of the written expression subtest of the Wechsler Individual Achievement Test (WIAT-R; Psychological Corporation, 1992) were more reliable than others. They found rater judgements in the domains of 'Vocabulary', and 'Ideas and Development' were the most reliable (Kendall's $W = .61$ for Vocabulary and $W = .60$ for Ideas, $p < 0.001$). In contrast, Dunsmuir and Blatchford (2004) found higher rates of agreement on structural elements such as 'Organisation' and 'Structure' when they used a series of hierarchically organised statements linked to the National Curriculum to score writing samples produced by 7 year-old children. The first aim of this study was to investigate whether the WAM shows statistically significant inter-rater reliability and whether judgements in certain domains are more reliable than others.

The internal consistency of a test estimates the degree of relatedness of the individual items within a test. Cronbach's alpha is a widely used measure of internal consistency reliability and can be thought of as describing how much each item is associated with each other item and the overall score. In general, a test should have reliability of 0.7 and preferably closer to 0.9 to be considered useful (Aron & Aron, 1999). The standard error of measurement (SEM) is an additional reliability statistic based, in part on the computed reliability coefficient and can be used to estimate how much variability is expected around a particular score measurement error. This study therefore also sought to investigate whether the WAM demonstrates statistically significant internal consistency reliability.

According to Hayes et al. (2000), test-retest reliability measures the degree to which the quality of writer's performance tends to remain the same in successive writing samples. If writing performance varies widely from one occasion to the next, the utility and appropriateness of the writing assessment measure could be considered debatable (Hayes et al., 2000). The second aim of this study was to investigate whether the WAM shows statistically significant test-retest reliability.

Several studies have sought to ascertain the validity of a range of writing assessments (Messick, 1990, 1994; Moss, 1994; Wiggans, 1994). Messick (1994) describes validity as a multifaceted concept. He claims that test validation should take into account all aspects of the assessment situation likely to influence test scores and any factor that effects performance should be controlled (e.g. test content, test administration conditions and scoring criteria). He argues for "validity generalization" which means that test score should be meaningful to those that use them and not misinterpreted to the disadvantage of those that are assessed. Messick contends that construct validity, content validity and criterion validity are types of evidence for the validity of an assessment measure. Construct validity has more recently been defined as being the general, overarching notion of validity, with content and criterion validity being facets of construct validity (Alderson, 2000; Bachman, 2000; Bachman & Palmer, 1996). Within this wider definition, validity refers to the extent to which a given test score can be interpreted as an indicator of the abilities or constructs it is intended to measure. Therefore, the main focus of any test's validity is construct validity, in addition to issues regarding its content and concurrent validity. It may be worth noting, however, that no test is entirely valid because validation is an ongoing process (Weir, 2005).

Kane (2006) claims that validity should begin with an explicit statement of the proposed interpretation of test scores and consider also the rationale for the relevance of the interpretation to the proposed use. In addition, he argues that an important aspect of test validation is to ascertain the degree that the proposed interpretations and uses are appropriate and relevant. This should include examination of the claims that are predicated on a test's scores, i.e. "the network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2013, p. 2), followed by a programme of research that tests those claims.

For the purposes of this study, it was important therefore to define the construct from the outset of the validation process. The direct assessment approach employed by the WAM not only aims to address a child's vocabulary and the application of editing skills (writing processes) and the mechanics of writing, but also the child's skills in formulating an idea and developing that idea into coherent discourse (rhetorical knowledge). It could be argued that these aspects do not adequately capture the construct for writing ability. Indeed, important components of the writing process, identified within theoretical models (e.g. Hayes, 1996) and through research (Rijlaarsdam & van den Bergh, 2006) are not captured by the domains in the WAM. For example, it does not directly assess planning and revision, motivation and affect and cognitive processes such as text interpretation and reflection. However, the WAM does address 'observable attributes' of writing, without making assumptions about underlying traits. Kane (2013) states: "An observable attribute is defined in terms of how well a test taker will perform on average over a target domain of possible observations" (Kane, 2013: 22).

The WAM was also designed to reflect the transition from indirect to direct assessment in the field of education (Huot, 1990). Such a shift has been supported by practitioners for many years, and has been promoted in the publications of professional organisations (Cooper & Odell, 1977). Evaluating construct validity can also confirm that a scoring method measures writing in an instructionally important way through reference to existing groups (Tindal & Parker, 1991). The construct model of validity draws on observations to estimate the construct's value, although it is important to note that the construct is not defined by the observations.

In the current reappraisal of writing in schools, emphasis is given to the need for both effective instruction and for adequate assessment of writing proficiency, since remediation of writing deficiencies and progress monitoring implies accurate assessment (Scardamalia & Bereiter, 1986). It is important therefore that writing assessment provides both evaluation and formative adjustment of instruction (Moran, 1987). Hence, writing tests are needed that are sensitive to increments of skill growth between children of different age and writing proficiency (Tindal & Parker, 1991). In this way, a scoring measure should be sensitive to these differences. According to Tindal and Parker (1991) scaling sensitivity can help target a measure use, for younger versus older students or for low/medium/high achieving students. The third aim of the study was to find out whether the WAM can discriminate significantly between year groups.

Charney points out that ". . .valid writing assessment should be sensitive to a writer's true abilities" (1984: 65). Evidence for content validity is established by investigating whether the test items actually correspond to the content area they are supposed to represent and can be determined using expert judgement and item analysis (Rosnow & Rosenthal, 2002). Muenz et al. (1999) carried out an item analysis of the written expression scoring model of the WIAT-R to assess its validity. They identified 'Ideas and Development', 'Organisation, Unity and Coherence' and 'Grammar and Usage' as being the most valid based on item total correlations. The fourth aim of this study was therefore to explore whether the WAM shows statistically significant content validity and whether specific elements are more valid in terms of their relationship to the total in comparison to the rest of the elements in the scoring model.

Criterion validity is the degree to which the assessment measure correlates (at least low to moderate correlation) to other accepted assessment measures (Tindal & Parker, 1991). According to Charney (1984), if the results of a measure correlate with another measure then the two measures can be considered equally valid. In order for a newly developed writing assessment measure to obtain criterion validity it must demonstrate that results correlate with a 'criterion', a previously validated measure of writing ability. The fifth aim of the study was to ascertain whether or not the WAM shows statistically significant criterion validity.

3. Methodology

3.1. Participants

Information about the 97 participants is shown in Table 1. The participants were recruited from one primary school in a large urban area in the south east of England. The demographic data from the school indicated that the population of students that it serves is representative of the local authority with regard to proportion of individuals with special educational needs (18.3%), students eligible for free school meals, a proxy indicator of low income and economic deprivation (14.6%) and students from ethnic minority groups (20%). This also is representative of national demographic patterns.

3.2. Procedure

Four raters were involved in assessing 97 children's scripts (one per participant). All were psychology graduates who were also qualified and experienced teachers undertaking a post-graduate professional qualification in educational psychology. One rater had substantial experience of teaching and assessing children's writing at the developmental level of children in the study and took the lead in training the other three raters. All scripts were scored over a two-day period using the Writing Assessment Measure (WAM) rubric. This is shown in Appendix A, which details the Elements and Marking Criteria against which the scripts were scored. The Writing Assessment Measure (WAM) rubric contains the following eight domains: Handwriting, Spelling, Punctuation, Sentence Structure and Grammar, Vocabulary, Organisation and Overall Structure, Ideas. Four descriptive statements with defined markers of competence are included within each domain. For example, within the Sentence Structure and Grammar domain, the lowest level of competence is defined as 'Writes simple sentences which include the conjunction 'and" and the highest level of competence is captured by the following descriptive statement: 'Secure control of complex sentences. Understands how clauses can be manipulated for effect. Able to use conditional and passive voice (e.g. having watched him eat a dog biscuit, she felt sick).'

The lead rater coded eight sample scripts and ensured that she was clear about the rationale for decisions about level awarded within each domain. These scripts were then presented to the other three raters, who discussed and agreed the allocation of criteria. Following this, a manual was generated for reference, which included scanned copies of the eight sample scripts and the agreed decisions about application of level criteria, presented by domain. The conferencing between raters ensured that all were familiar with the Writing Assessment Measure (WAM) rubric and could reference the training manual in the event of any uncertainty about the award of scoring criteria.

3.2.1. Measures

3.2.1.1. Writing Assessment Measure (WAM). The WAM is based on the structure and format of the Wechsler Objective Language Dimensions Written Expression subtest (WOLD, Psychological Corporation, 1996), with modified dimensions that incorporate descriptors from the National Curriculum writing attainment targets. It is designed to assess narrative writing in response to a written prompt. Pupils are scored on 7 main domains (or elements) of written expression. Each element is scored

Table 1
Year group, age and gender of participants.

	Number of participants	Gender	
		Male	Female
Year 3 students (7–8 years)	25	14	11
Year 4 students (8–9 years)	25	10	15
Year 5 students (9–10 years)	26	13	13
Year 6 students (10–11 years)	21	11	10
Total	97	48	49

on a 4 point scale, each point having a specific unambiguous description adapted from the National Curriculum Level 1–4 descriptors. Each child's written response is scored for each element, based on which description best suits the sample of writing produced by the child. So both an individual score for each element and an overall score (based on the total of the analytic scores) are recorded.

3.2.1.2. Wechsler Objective Language Dimensions (WOLD) Written Expression subtest. The WOLD is a UK standardisation of the WIAT-R which, at the time of this study, was only available in the USA. It has however been included as a 'written expression' sub-test in the latest version of the WIAT-II^{UK} (Wechsler, 2005). The written expression subtest is designed for ages 8–16 and assesses writing proficiency using either analytic or holistic scoring. For the purposes of this study, only analytic scoring was used. The WOLD is a timed test and it therefore requires students to write for 15 minutes on two occasions, using two writing stimuli. Only one of these stimuli, 'Design their ideal Place to Live', was administered in this study.

Test-retest reliabilities calculated for the Wechsler Objective Language Dimensions (WOLD) Written Expression subtest averaged 0.77 across all age groups (see Wechsler, 1996) which indicates adequate stability across time. The average inter-rater reliability coefficient was 0.89 for the first writing stimulus and 0.79 for the second (see Wechsler, 1996). This indicates that for the Written Expression subtest writing samples which require rater judgement can be scored reliably. However Cole, Haley et al. (1997: 32) comment that the inter-rater coefficients were 'spuriously high' due to the use of an extremely heterogeneous sample. The WOLD Written Expression subtest was correlated with the Woodcock-Johnson Psych-Educational Battery Revised Tests of Achievement Dictation subtest (WJ-R: Woodcock & Johnson, 1989) which required written response and includes questions which assess the children's knowledge of spelling, punctuation, capitalisation and word usage. A correlation of 0.72 was found and was considered to be within the expected range for an indirect measure (Dictation) and a direct measure (Written Expression) of writing achievement (see Wechsler, 1996).

3.2.2. Administration

The Writing Assessment Measure (WAM) was administered to the participants in their class groups following a standardised, scripted introduction. The first writing prompt was presented orally and in written form. This was a timed assessment task and pupils were presented with Prompt 1 and then asked to write for 15 minutes. Later the same day, the WOLD Written Expression subtest was administered to each class group following standardised procedures outlined in the WOLD manual (Wechsler, 1996).

In order to evaluate the stability of the test over time, the Writing Assessment Measure (WAM) was administered to all class groups following a three week interval (test-retest reliability). The second prompt presented in Appendix B was used for this purpose. Both prompts elicit narrative written responses. Standardised administration procedures were once again followed.

3.2.3. Scoring

Students' names were recorded onto the database and randomly assigned a number using a computer application. A 'blind' procedure was followed for the scoring and children's writing scripts were evaluated anonymously. Twenty scripts for the first administration of the Writing Assessment Measure (WAM) were scored by four researchers involved in this study, following the analytic scoring criteria guidelines. The results were used to evaluate inter-rater reliability and the initial content validity of the measure. One rater then scored all 97 samples for prompt 1 (see Appendix B) and the results were used to evaluate internal consistency reliability and the construct validity of the measure. The scripts obtained from the second administration of the WAM (prompt 2) were scored by one rater following the marking criteria outlined and were correlated with the scores from the first administration (stimulus 1) and used to assess test retest reliability. The WOLD scripts were scored by another rater (who had followed the recommended steps outlined by the WOLD manual for mastering the scoring of written responses). Results from the WOLD were used to evaluate criterion validity.

Scores for each script were recorded on a separate sheet and no scoring marks were placed on the scripts. Once raters finished scoring a script they were instructed not to re-adjust the scores assigned. Guidelines on effective scoring procedures were developed, along with annotated scoring of written expression responses to the Writing Assessment Measure (WAM). Together these guidelines form a training manual, which is available from the first author on request.

4. Results

Table 2 provides a summary of the descriptive data (means and standard deviations) for the scores obtained from the administration of the Writing Assessment Measure (WAM) with prompt 1. The distribution of data was considered to be normal. A total mean score of 14.91 was obtained for the sample, with a standard deviation of 5.10. As expected, the mean scores increased according to year group.

4.1. Internal consistency

The data collected was used to investigate the internal consistency of the measure using Cronbach's alpha which estimates the extent to which the items (Handwriting, Spelling, Punctuation, Vocabulary, Sentence Structure and Grammar, Organisation and Planning, and Idea's) appear to be measuring the same concept. Table 2 provides a summary of Alpha model reliability with item total statistics.

A Cronbach's alpha coefficient of 0.87 (N = 97) was obtained which indicates the overall internal consistency reliability of the measure is 'good' based on the number of items and the mean inter-item correlations. In addition, all the 'alpha if item deleted' coefficients are similar, which suggests consistency of the items in relation to the total score. It can be observed from Table 3 that the 'alpha if item deleted' correlations remain lower than the overall alpha calculated. If alpha for an item deleted is greater than the scale's computed alpha level, then that item should be considered for removal. It can be noted however that the 'alpha if item deleted' for the elements 'handwriting' and 'sentence structure' are high and subsequently the item total correlations for these elements are low.

The reliability coefficient provides a relative measure of the accuracy of test scores; however, it does not provide an indication, in absolute terms of how accurate the scores truly are (Murphy & Davidshofer, 2001). To describe the accuracy of the scores concretely the standard error of measurement (SEM) provides an estimate of measurement error. The SEM is inversely related to the reliability coefficient: the greater the reliability, the smaller the error of measurement and the greater the precision of the obtained score. An SEM of 1.82 was calculated using the reliability coefficient (r = 0.87) and the standard deviation (sd = 5.10) of the raw scores. The SEM could be used to build a 95% confidence interval around the true score. In a normal distribution, 95% of all scores fall within 1.96 standard deviations of the mean. Therefore, for a given true score, t, 95% of the obtained scores will fall between $t \pm 1.96$*SEM (i.e. $t \pm 1.96$*1.81).

Table 2
Mean and standard deviations for the year groups and total group.

Total and year group	Number (N)	Mean	Standard deviation
Total	97	14.91	5.10
Year 3	25	10.20	4.32
Year 4	25	14.88	3.38
Year 5	26	16.38	4.52
Year 6	21	18.71	4.17

Table 3
Internal consistency reliability analysis with item total statistics.

Items	Scale mean if item deleted	Scale variance if item deleted	Corrected item total correlation	Alpha if item deleted
Handwriting	12.55	20.85	.49	.87
Spelling	12.47	19.27	.68	.85
Punctuation	12.92	16.99	.73	.85
Sentence structure	12.78	20.77	.59	.86
Vocabulary	12.99	20.34	.67	.85
Organisation and Planning	12.90	20.28	.67	.85
Ideas	12.84	18.14	.77	.84

Table 4
Inter-rater agreement on each writing element.

Element	Kappa means	Kappa ranges
Handwriting	0.80	0.71–0.86
Spelling	0.86	0.75–0.92
Punctuation	0.71	0.61–0.87
Sentence structure and grammar	0.88	0.80–1.00
Organisation and planning	0.78	0.70–0.92
Vocabulary	0.83	0.67–1.00
Ideas	0.62	0.56–0.71

4.2. *Inter-rater reliability*

Four raters scored 20 of the 97 samples using the criteria outlined in the Writing Assessment Measure and levels of inter-rater agreement for each element were computed using Cohen's kappa (Cohen, 1960). Kappa has a range from 0 to 1.00, with larger values indicating better reliability. Generally, a kappa >0.60 is considered satisfactory (Brown, Glasswell, & Harland, 2004). Table 4 represents the Kappa means and kappa ranges between all four raters for each element of the marking criteria for 20 scripts. It can be seen that there was a high level of agreement between the raters for the elements 'Sentence Structure' ($k = 0.88$) and 'Spelling' ($k = 0.86$). Mean rater agreement was least for the element 'Ideas' ($k = 0.62$). Kappa ranges were greatest for the element 'Vocabulary' (0.67–1.00).

Interval data was also provided in the form of total scores for each student. To assess the inter-rater reliability for this interval data, across the four raters, an intraclass correlation coefficient (ICC) was calculated. The ICC assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. It is more sensitive to rater mean differences as decreases in response to lower mean differences. A two way random effects model (absolute agreement definition), which assumes that each subject was rated by two or more raters and that these raters are randomly selected from a larger population of raters, was used to calculate the coefficient (Shrout & Fleiss, 1979). The average rater intraclass correlation was found to be high, ICC = 0.97, $p < 0.01$ (range 0.93–0.99 at 95% confidence interval).

4.3. *Test-retest reliability*

The stability of the scores on the Writing Assessment Measure was assessed across time. The interval between testing was 21 days. For test-retest reliability, a Pearson's r correlation coefficient was calculated to investigate the strength of association between scores on the writing measure at time 1 and time 2. It was found that there was a strong correlation between scores with $r(50) = 0.82$, $p < 0.001$.

4.4. *Content validity*

Item total correlations were inspected for the scores in Prompt 1 ($N = 97$) using Pearson's correlation calculations and any item with an item total coefficient less than 0.2 was considered for revision or replacement. Table 5 presents item total correlations for the elements. All item total correlations were significant at $p < 0.01$ (ranging from $r = 0.63$ for 'Handwriting' to $r = 0.85$ for 'Ideas'). Inter item correlations were also significant with the highest inter item correlation for 'Ideas' and 'Vocabulary' ($r = 0.79$, $p < 0.01$) and the lowest for 'Sentence Structure' and 'Handwriting' ($r = 0.26$, $p < 0.05$).

4.5. *Concurrent validity*

A Pearson's correlation revealed that the Writing Assessment Measure (WAM) correlated significantly with the Wechsler Objective Reading Dimensions (WOLD) Written Expression subtest, with $r(50) = 0.786$, $p < 0.01$. In a future study, concurrent validity will be revised to include correlations between the writing measure and the WIAT II$^{\mathrm{UK}}$.

Table 5
Item to item and item to total correlations for elements using Pearson's *r*.

	Handwriting	Spelling	Punctuation	Sentence structure	Vocabulary	Organisation and planning	Ideas	Total Prompt 1
Handwriting	1	.46**	.51**	.26*	.32**	.35**	.38**	.62**
Spelling	.46**	1	.57**	.50**	.57**	.48**	.55**	.78**
Punctuation	.51**	.57**	1	.46**	.50**	.63**	.62**	.83**
Sentence structure	.26*	.50**	.46**	1	.43**	.53**	.57**	.69**
Vocabulary	.32**	.57**	.50**	.43**	1	.46**	.79**	.76**
Organisation and planning	.35**	.48**	.63**	.53**	.46**	1	.59**	.76**
Ideas	.38**	.55**	.62**	.57**	.80**	.59**	1	.85**
Total Prompt 1	.63**	.78**	.83**	.69**	.76**	.76**	.85**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

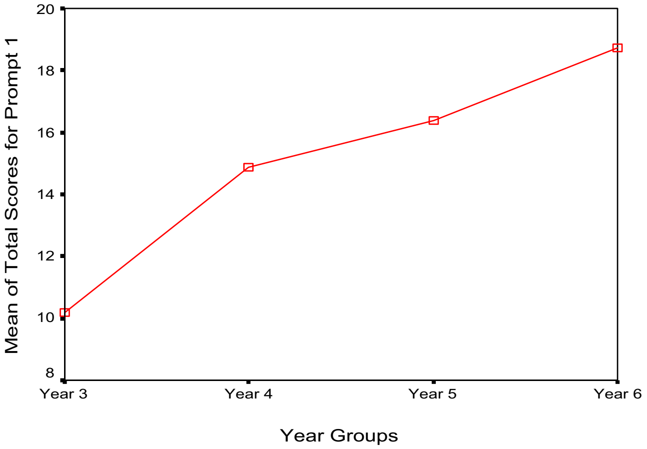

Fig. 1. Mean plots for each year group.

4.6. *Observable attributes*

A one-way ANOVA was used to measure the statistical estimates of variability in test scores associated with differences in year groups. Means and standard deviations across year groups and for the overall sample are reported in Table 1. The model achieved statistical significance with $F(3, 93) = 17.91$, $p < 0.001$. Fig. 1 provides a visual representation of the movement of mean scores for each year group.

Having established from the interpretation of the ANOVA that there was a significant difference overall between the Year groups, post hoc pair wise comparisons, using Tukey test with p set at 0.5, were used to explore differences between the individual Year groups. The Tukey post hoc test revealed that Year 3 students' scores were significantly different from Year 4, Year 5 and Year 6 student scores. Year 4 students' scores were significantly different from Year 6. No other specific post hoc contrasts were significant.

5. Discussion

The focus of the study was to evaluate the reliability and validity of a writing measure that has been developed to reflect the knowledge, skills and understanding that students of different abilities exhibit in written expression.

The internal consistency of the Writing Assessment Measure is considered to be good, indicating that the elements are homogenous and appear to be measuring a unified construct. The estimate of measurement error relative to the standard deviation of observed scores was also small, indicating the measure's accuracy in scoring. There was a high level of agreement between raters on all elements. All kappas were found to be greater than 0.6 which indicates considerable non-chance agreement and provides evidence that the agreements as a whole were reliable (Brown et al., 2004). Muenz et al. (1999) suggest that rater judgements on structural items in written expression measurement criteria tend to be consistently reliable. Cole, Muenz, Ouchi, Kaufman, and Kaufman (1997) define structural items as those that assess the quality of writing in terms of its unity, organisation and development of ideas and refer to items that assess grammar, punctuation and handwriting as 'mechanics items'. Dunsmuir and Blatchford (2004) similarly found higher rater agreement on items that related to organisational criteria. The best levels of agreement between raters were found for the elements 'Sentence Structure and Grammar', 'Spelling', 'Vocabulary' and 'Handwriting' with kappa's greater than 0.8 which includes a combination of both structural and mechanical items. However, kappas for 'Ideas' were less than for the other elements, ranging from 0.56 to 0.71 and differ from the findings by Muenz et al. (1999), where 'development of ideas' is considered to be a structural item. The study by Muenz et al. (1999) found rater judgements in this domain to be amongst the most reliable. In the current study, lower kappas for 'Ideas' may have occurred because some of the criteria were, to some extent, interpreted differently. 'Ideas' as a concept may need to be discriminated from development of ideas', as presented in the study by Muenz et al. (1999). There was a high correlation between this domain and 'Vocabulary' (Pearson's $r = 0.79$). This may be due to the overlap between the descriptive statements for awarding a '2' and a '3' for 'Ideas' in the assessment criteria, as both include a requirement to add detail which may have contributed to the significant relationship between 'Ideas' and 'Vocabulary'.

Intraclass correlation (ICC) reliability for the raters total scores indicated that there was little variance between raters overall. Caution should be taken when interpreting this result however, as ICC is strongly dependent on the trait variance within the population for which it is measured. This can complicate comparisons of ICCs measured in different populations, or in generalising results from a single population (Muller & Buttner, 1997). An extension of the current study, involving a larger and more diverse sample, would make such comparisons possible and the ICC findings more generalisable.

Test-retest reliability was also found to be significant, which indicates that the WAM is able to reliably measure the quality of writing over time. This adds further evidence to the utility and the appropriateness of a measure (Hayes et al., 2000). Hayes et al. (2000) also point out that some variance in the stability of writing is to be expected due to variations in genre, changes in motivation and fatigue, but that these variations shouldn't generally interfere with the overall consistency from one writing assignment to next.

Further analysis showed that the Year 3 student scores were significantly different from Year 4, Year 5 and Year 6 student scores. Year 4 students' scores were also significantly different from those of Year 6. This information indicates that the scoring system is able to discriminate between younger and older students. It would have been preferable if the Writing Assessment Measure had been able to discriminate between all adjacent year groups (e.g. between Year 4 and Year 5, and between Year 5 and Year 6). This may not have occurred due to the limited sample size, and variations in teaching and abilities between year groups. However, the scoring scale does allow for further growth as the maximum score is twenty eight and Year 6 student's scores had a mean of 18.71 (SD = 4.17).

Reviewing key developmental theories that have been adopted by writing development theorists over the past fifty years, Camp (2012) argues that children's developmental pathways are never linear. However, no writing assessment captures all the lines-of-development that constitute growth in writing, and the WAM is certainly no exception. For example, it does not assess a writer's initiative, their fluency and speed, their capacity to critique their own writing. These findings however show that it does capture the mechanics, writing processes and the rhetorical knowledge of children in the primary school years (aged 5–11 years). As the WAM has been developed with reference the National Curriculum attainment targets and uses an analytic scoring system, it can be considered useful in identifying specific relevant targets for children experiencing writing difficulty within this particular age range.

It is worth emphasising the distinction between testing and clinical assessment (Matarazzo, 1990), and the need to collect information beyond the assessment scores and consider writing as a socio-cognitive construct. To provide a complete picture of the child's writing difficulties we recommend that assessment results be integrated with information gained from direct observations, information from school records and background information from the family. Writing difficulties can then be determined and evaluated in the context of the child's instructional environment and history. Using this context (as opposed to viewing the learner in isolation) will help the assessor to gain a deeper understanding of the source of the child's writing challenges.

Content validity of the Writing Assessment Measure was demonstrated through inspection of item total correlations which were all significant. The correlational evidence also suggests that there is a significant association between the element 'Ideas' and the total score achieved on the measure. This is similar to the findings by Muenz et al. (1999) who identified 'Ideas and Development' as being the most valid element in terms of its relationship with the total in comparison to the other elements. However, as mentioned, rater judgement agreements in this domain were deemed the least reliable so further investigation would be needed to confirm the relationship between 'Ideas' and the total score.

Criterion validity was established using concurrent validity by administering the Wechsler Objective Language Dimensions Written Expression subtest (WOLD, Psychological Corporation, 1996) on the same day. Establishing criterion validity provides some assurance about replacing an existing measure with a new scoring method (Tindal & Parker, 1991). The results of the Writing Assessment Measure correlated significantly with the WOLD which is a previously validated and standardised measure of writing ability; therefore the two measures may be considered equally valid (Charney, 1984). The WOLD has also recently been used in research on writing (Kyriacou, 2009; Roberts, 2013; Williams & Larkin, 2013). However, further research on the validity of the instrument will aim to establish the criterion validity between the WAM and the WIAT II^{UK} which was not published at the time of this study.

All in all, based on the findings of this study, the WAM has demonstrated that it is consistent, stable over time and can be scored reliably by different raters. Evidence for validity was demonstrated through construct validity, content validity and criterion validity. As it was developed with reference to the National Curriculum attainment targets and uses an analytic scoring system, it can be considered useful in identifying specific relevant targets for children experiencing writing difficulty. As children are operating at different levels, it is important to be aware of what needs to be embedded in the teaching of written expression, to support and enhance their writing development. This is possible when practitioners are informed by robust assessment results so that they can plan their lessons accordingly.

It is worth noting, however, that validity is a matter of degree, and no instrument is absolutely valid or absolutely invalid. Messick (1994) advocates that, over time, validity evidence will continue to gather, either enhancing or contradicting previous findings. Similarly, the results of the present study provide a stepping stone which future research can build on, and are relevant to the population and the specific context used for the purposes of this study.

This study has focused on the advantages of developing a psychometrically sound instrument. However, critics of psychometric approaches to the assessment of writing argue that in trying to establish fairness and standardise judgements, the assessor's field of vision and interpretation of text can be diminished. There has been a growth of interest in hermeneutic approaches to reliability amongst certain writing researchers. This stance acknowledges and incorporates the individual assessor's interpretation of text and takes account of the preconceptions and existing knowledge they impose in making judgements about writing quality. Lynne (2004) argues against the search for universal, objective criteria to appraise writing as this ignores the context and the social construction of meaning. Likewise, Moss (1994, 2004) states that it is possible for valid inferences to be drawn from information that may not be reliable, a position supported by

Broad (2000), who argued for the incorporation of divergent judgements in the assessment process. Thus hermeneutic approaches to reliability were designed to address the trade-off between reliability and validity and have been demonstrated to be effective in assessing writing in older populations. Portfolio assessment (Hamp-Lyons & Condon, 2000), an assessment method that is becoming more established in the UK with teachers of younger children, is one means of addressing this.

This study needs to be considered with reference to its limitations. As it stands, the WAM provides a 'launching board' to the assessment of narrative writing in the British Educational System, opening the gates for further research and development. This should involve the consideration of adoption of hermeneutic approaches to reliability and evaluation of writing samples over time using wider, more interpretative criteria. In addition, this study did not seek to validate the uses and interpretations of test scores on the WAM, as advocated by Kane (2006, 2013) so in this respect did not fulfil the more complex criteria for validity by evaluating the plausibility of claims made on the basis of assessment results.

The sample size was small and was selected on the basis of convenience. It is probable that the students have been taught writing by the same teachers over the years and therefore the sample could be considered somewhat homogeneous in nature. A larger more diverse sample size would be needed to further validate the reliability and validity of the Writing Assessment Measure. In addition students were exposed to repeat testing, which could pose a threat to the internal validity of the study. Students in each class should have been randomly assigned to two groups, one half could have been administered the Writing Assessment Measure and the other half the WIAT II^{UK}. The situation could have been reversed for the second administration the same day, and in this way practice effects could have been controlled. Also the authors of the study were involved in rating all the scripts. This may have contributed to tester bias as expectations of an outcome by persons running the study may influence the outcome (Murphy & Davidshofer, 2001). Ideally raters should be blind to the purpose of the study; however other research has shown little impact of rater knowledge of the hypotheses (Kent, O'Leary, Diament, & Diez, 1974).

This study overall has contributed to the field by developing and evaluating a writing measure that has potential utility to the professionals assessing children's writing. Jeffery (2009) in a review conducted in the US comparing national and state writing assessments concluded that writing assessment rubrics should align with definitions of writing proficiency – this has been achieved by the WAM, which provides a means of assessment relevant to the context in which children are taught and diagnostic information that is congruent with existing curricular approaches. It also covers the mechanics of writing (handwriting, spelling, punctuation, grammar) as well as narrative skills (vocabulary, organisation and overall structure, ideas), hence providing a more complete picture of students' writing skills. The analytic scoring system provides information about individual pupil profiles with regard to strengths and difficulties with the writing process that can be used to inform and target intervention planning relating to children's written expression. However, it should be noted that it is important that individuals using the WAM are well-trained and focus on children in the process of writing, rather than merely assigning scores to writing products.

This study could be further developed by evaluating the WAM on a larger, more diverse sample. Further consideration also needs to be given to the element 'Ideas' in relation to inter-rater agreement and its relationship with the overall score. The predictive validity of the WAM could also be investigated in relation to students' future performance in writing. However, one of the most important areas for further research reflects the measure's capacity to discriminate between different ability groups (for instance, students with special educational needs, Specific Learning Disabilities, or high achieving students). In this way, further evidence for this element of the measure could be obtained. Using a larger sample size in obtaining quantitative information and collating this information with their scores on the WAM would provide additional information about the relationship between student's approaches to writing and their actual writing.

Appendix A. Writing Assessment Measure (WAM)

Elements and marking criteria

Writing Assessment Measure (WAM)

TIME GUIDELINE: *Prompt 1:* 15 minutes *Prompt 2:* 15 minutes
DISCONTINUE RULE: Stop the child after 15 minutes of writing

Elements and Criteria	Circle Score
Handwriting	
• Writing is consistent, fluent and cursive.	**4**
• Clear, neat and legible and may show evidence of joining handwriting	**3**
• Handwriting may vary in shape and size and is beginning to develop consistency.	**2**
• Handwriting is indecipherable or difficult to read.	**1**
Spelling	
• Evidence of correct spelling of complex words containing prefixes/suffixes or irregular words e.g. souvenir, destruction, and conscious.	**4**
Attempts to spell some complex or polysyllabic words using visual or phonetic strategies, e.g 'safariye' for safari, 'adventerous' for adventurous.	**3**
• Spells the majority of high frequency common words correctly e.g. inside, because, while.	
• Spells some common monosyllabic words correctly (e.g. mum, cat, bird). Uses phonic strategies to attempt to spell high frequency common words e.g. 'grat' for great, 'fhun' for fun.	**2** **1**
Punctuation	
• Uses a range of punctuation to clarify structure and create effect (e.g. speech marks, dashes, brackets, apostrophes, commas to demarcate sentences).	**4**
• Secure use of full stops and capital letters. Uses punctuation in addition to capital letters and full stops, the majority are used correctly (e.g. question marks, exclamations marks, commas in lists).	**3**
• Evidence of accurate use of capital letters and full stops, however few there are. (e.g. Sentence finishes with a full stop and next sentence begins with a capital letter)	**2**
• Shows awareness of how full stops are used in writing.	**1**
Sentence Structure and Grammar	
• Secure control of complex sentences. Understands how clauses can be manipulated for effect. Able to use conditional and passive voice (e.g. having watched him eat a dog biscuit, she felt sick)	**4**
• Beginning to write extended sentences including subordinators (e.g. if, so, while, when, after). The basic grammatical structure of sentences usually correct (e.g. usually consistent and correct use of tenses and nouns and verbs agree).	**3**
• Beginning to use other conjunctions to create compound sentences (e.g. because, but, so, then) and may be using multiple clauses (still mixing up tenses).	**2**
• Writes simple sentences which include the conjunction 'and'.	**1**
Vocabulary	
• Demonstrates use of well-chosen vivid & powerful vocabulary to create effect (e.g. verbs, adjectives, adverbs)	**4**
• Varied use of adjectives, verbs and specific nouns (e.g. delicious for nice/sauntered for went/poodle for dog)	**3**
• Some selection of interesting and varied verbs e.g. jumped, compare, guess	**2**
• Uses simple vocabulary, appropriate to content. Writing is composed of simple nouns and verbs e.g. look, went, go, play, see	**1**
Organisation and Overall Structure	
• Paragraphs are well organised, based on themes and provides a cohesive text for the reader (e.g. paragraphs, subheadings, logically organised events).	**4**
• Uses paragraphs to organise writing, showing an identifiable structure. May be short sections.	**3**
• Themes are expanded upon and linked together in a series of sentences.	**2**
• Communicates meaning but may 'flit' from idea to idea and any themes that are expanded are done so in one sentence.	**1**
Ideas	
• Ideas are creative and interesting in a way that engages the reader. Uses a range of strategies and techniques such as asides, comment, observation, anticipation, suspense, tension.	**4**
• Ideas are imaginative and varied evidence of descriptive detail about characters, settings, feelings, emotions & actions.	**3**
• Ideas are developed to by adding detail (e.g. is beginning to provide additional information or description beyond a simple list).	**2**
• Produces short sections of ideas which may be repetitive and limited in nature.	**1**
Total score	

Appendix B. Writing Assessment Measure (WAM) Prompts

Prompt 1: School trip

Imagine that you could go anywhere you wanted to on a school trip with your class and your teacher. You could go anywhere at all. Write about where you would go and what you would do.

Prompt 2: Ideal teacher

Imagine that you could choose your ideal teacher. What would be your ideal teacher be like? Describe what he or she is like and what they do that makes them your perfect teacher.

Prompt 3: Ideal playground

Imagine that you have been given the important job of re-designing your school playground by your Head Teacher. You can do anything you want to turn it into your perfect playground. Describe what sort of things you would have in it and what it would look like.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alvès, R. A., Castro, S. L., de Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In M. Torrance, D. Galbraith, & L. Van Waes (Eds.), *Recent developments in writing-process research* (Vol. 20) (pp. 55–65). Dordrecht-Boston-London: Kluwer Academic Press.

Aron, A., & Aron, E. N. (1999). *Statistics for psychology* (2nd ed.). London: Prentice Hall.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barnett, A., Stainthorp, R., Henderson, S., & Scheib, B. (2006). *Handwriting policy and practice in English primary schools. An exploratory study*. London: Institute of Education.

British Psychological Society. (2009). *Code of ethics and conduct*. Leicester: BPS.

British Psychological Society. (2011). *Guidance for assessors for the qualification – Test user: Educational, ability/attainment (CCET)*. Leicester: BPS.

Broad, B. (2000). Pulling your hair out: "Crises of standardization in communal writing assessment". *Research in the Teaching of English*, *35*(2), 213–260.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*, 105–121.

Camp, H. (2012). The psychology of writing development. *Assessing Writing*, *17*, 92–105.

Chanquoy, L., & Alamargot, D. (2002). Working memory and writing: Evolution of models and assessment of research. *Annee Psychologique*, *102*(2), 363–398.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*, 65–81.

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cole, J. C., Haley, K. A., & Muenz, T. A. (1997). Written expression reviewed. *Research in the Schools*, *4*(1), 17–34.

Cole, J. C., Muenz, T. A., Ouchi, B. Y., Kaufman, N. L., & Kaufman, A. S. (1997). The impact of the pictorial stimulus on the written expression output. *Psychology in Schools*, *34*(1), 1–9.

Connelly, V., Campbell, S., MacLean, M., & Barnes, J. (2006). Contribution of lower-order letter and word fluency skills to written composition of college students with and without dyslexia. *Developmental Neuropsychology*, *29*(1), 175–196.

Connelly, V., Dockrell, J., & Barnett, J. (2005). The slow handwriting of undergraduate students constrains overall performance in exam essays. *Educational Psychology*, *25*, 99–107.

Connelly, V., Dockrell, J. E., & Barnett, A. (2011). Children challenged by writing due to language and motor difficulties. In V. Berninger (Ed.), *Cognitive psychology of writing handbook: Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 217–245). New York: Psychology Press.

Cooper, C. R., & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.

Dunsmuir, S., & Blatchford, P. (2004). Predictors of writing competence in 4-7 year old children. *British Journal of Educational Psychology*, *74*, 461–483.

Dunsmuir, S., & Clifford, V. (2003). Children's Writing and the use of ICT. *Educational Psychology in Practice*, *19*(3), 171–187.

Espin, C. A., Weissenburger, J. W., & Benson, B. J. (2004). Assessing the writing performance of students in special education. *Exceptionality*, *12*, 55–66.

Fathman, A. K., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–190). Cambridge: University of Cambridge Press.

Flower, L., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg, & E. R. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gregg, N., & Mather, N. (2002). School is fun at recess: Informal analyses of written language with students with learning disabilities. *Journal of Learning Disabilities*, *35*, 7–22.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Practice, theory and research*. Cresskill, NJ: Hampton Press.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, J. R., Hatch, J. A., & Silk, C. M. (2000). Does holistic assessment predict writing performance. *Written Communication*, *17*, 3–26.

Herrington, A., & Curtis, M. (2003). Writing development in the college years: By whose definition? *College Composition and Communication*, *55*(1), 69–90.

Hooper, S. R., Montgomery, J., Swarz, C., Reed, M. S., Sandler, A. D., Levine, M. D., et al. (1994). Measurement of written language expression. In G. R. Lyons (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 375–417). Baltimore: Paul H. Brookes.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–249.

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing*, *14*, 3–24.

Jones, D. (2002). Keeping track: Assessment in writing. In M. Williams (Ed.), *Unlocking writing: A guide for teachers* (pp. 92–105). London: David Fulton Publishers Ltd.

Lynne, P. (2004). *Coming to terms: Theorizing writing assessment in composition studies*. Logan, UT: Utah State University Press.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., Vol. 20, pp. 21–65). New York: American Council on Education and Macmillan.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kent, R. N., O'Leary, K. D., Diament, C., & Diez, A. (1974). Expectation bias in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, *42*, 774–780.

Kyriacou, M. (2009). *The development of narrative writing in primary school children: Designing and evaluating an experimental intervention*. Unpublished doctoral thesis. University of Oxford.

Matarazzo, J. D. (1990). Psychological assessment vs psychological testing: Validation from Binet to the school, clinic and courtroom. *American Psychologist*, *45*, 999–1017.

McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: Gilford Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education: Principles, Policy and Practice*, *13*(1), 69–95.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ: Educational Testing Services.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23.

Miller, H. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education*, *3*, 285–296.

Moran, M. R. (1987). Options for written language assessment. *Focus on Exceptional Children*, *19*(5), 1–10.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*, 5–12.

Moss, P. A. (2004). The meaning and consequences of "reliability". *Journal of Educational and Behavioral Statistics*, *29*(2), 245–249.

Muenz, T. A., Ouchi, B. Y., & Cole, J. C. (1999). Item analysis of written expression scoring systems from the PIAT-R and WIAT. *Psychology in Schools*, *36*, 31–40.

Muller, R., & Buttner, P. (1997). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *16*(7), 821–823.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed.). London: Prentice Hall.

Office of Standard in Education (OFSTED). (1999). *Pupils with specific learning disabilities in mainstream schools*. London: OFSTED Publications Centre.

Office of Standard in Education (OFSTED). (2011). *Removing barriers to literacy*. London: OFSTED Publications Centre.

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory & Cognition*, *30*(4), 594–600.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on interater reliability: An empirical study of a holistic rubric. *Assessing Writing*, *7*, 143–164.

Peverly, S. T. (2006). The importance of handwriting speed in adult writing. *Developmental Neuropsychology*, *29*, 197–216.

Qualifications and Curriculum Development Agency. (2010). *The National Curriculum. Level descriptors for subjects*. London: QCDA & DCSF.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*, 18–39.

Rijlaarsdam, G., & van den Bergh, H. (2006). Writing process theory: A functional dynamic approach. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*. New York: Guilford Press.

Roberts, E. (2013). *The impact of self-regulated strategy development on writing outcomes and affective factors*. Unpublished doctoral thesis. University College London.

Rosnow, R. L., & Rosenthal, R. (2002). *Beginning behavioural research: A conceptual primer* (5th ed.). New Jersey: Prentice Hall.

Ruttle, K. (2004). What goes inside my head when I'm writing? A case study of 8–9 year old boys. *Literacy*, *38*(2), 71–77.

Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 778–803). New York: Macmillan.

Shaw, S., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing. In *Studies in language testing*. Cambridge: Cambridge University Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–427.

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and Practice*, *6*, 211–218.

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. MacArthur, S. Graham, & J. Fitzgerald Jill (Eds.), *Handbook of writing research* (pp. 67–80). New York: The Guilford Press.

Wechsler, D. (1996). *Wechsler Objective Language Dimensions (WOLD)*. London: Harcourt Brace & Company.

Wechsler, D. (2004). *The Wechsler Intelligence Scale for Children—Fourth edition (WISC-IV^{UK})*. London: Pearson Assessment.

Wechsler, D. (2005). *Wechsler Individual Achievement Test, Second UK Edition (WIAT-II UK)*. London: Pearson.

Weigle, S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

Wiggans, G. (1994). The constant danger of sacrificing validity to reliability. Making writing assessment server writers. *Assessing Writing*, *1*, 129–139.

Williams, G. J., & Larkin, R. F. (2013). Narrative writing, reading and cognitive processes in middle childhood: What are the links? *Learning and Individual Differences*, *28*, 142–150.

Williams, G. J., Larkin, R. F., & Blaggan, S. (2013). Written language skills in children with specific language impairment. *International Journal of Language & Communication Disorders*, *48*(2), 160–171.

Williams, M. (2000). The part which metacognition can play in raising standards in English at Key Stage 2. *Reading*, *34*(1), 3–8.

Woodcock, R. W., & Johnson, M. B. (1989). *WJ-R Tests of Cognitive Ability*. Itasca, IL: Riverside Publishing.

Sandra Dunsmuir is Co-Director of the Doctorate in Educational and Child Psychology at University College London. She completed her educational psychology training at UCL in 1986 and her PhD in 2000. Her research interests include early literacy development, and factors that influence children's attainment and progress in writing.

Maria Kyriacou is a Research Fellow at UCL. She completed her doctorate at Oxford University, where she designed and implemented an intervention aiming to improve children's writing. Maria was formerly a consultant at Oxford University, evaluating evidence on children's language and literacy development, as part of a government funded review.

Su Batuwitage is an Educational Psychologist at Hackney Learning Trust. She completed her Educational Psychology training at University College London (UCL) in 2005. She has specialist experience in visual impairment and has a particular interest in literacy assessment and intervention and adult professional development.

Emily Hinson is Acting Principal Educational Psychologist for Bracknell Forest Council. She developed an interest in literacy development during her professional training in Educational Psychology at UCL and now, through her work with schools she works in supporting staff to develop a range of school based literacy interventions.

Victoria Ingram is an Educational Psychologist at Buckinghamshire County Council. She completed her professional training at UCL in 2005. Her special interests lie in literacy assessment and intervention, and VIG (Video Interaction Guidance) to promote positive parent–child interactions to support learning and emotional development.

Siobhan O'Sullivan is an educational psychologist and Programme Leader for the MA in Educational Psychology at Mary Immaculate College, University of Limerick, Ireland. She is completing a professional doctorate in educational psychology at UCL. Her research interests include teacher attitudes towards inclusive education and interventions to develop skills in literacy.