Chapter 7

Mixed Methods: Using a Combination of Techniques to Assess Writing Ability

Hiske Feenstra

Abstract A productive ability such as writing can be assessed only through a candidate's performance on a task, giving rise to concerns about the reliability and validity of writing assessments. In this chapter, it is argued that a combination of different techniques can help improve the quality of an evaluation of writing ability. First, an indirect test can be applied to reliably assess specific components of the writing process (e.g., revision), adding to the validity of the assessment. Furthermore, an analytic rating procedure accompanied by anchor essays allows raters to reliably evaluate the content and overall structure of written pieces. And last, automated scoring techniques can be used to objectively score text features considered important to text quality. Combining these methods provides an evaluation that is solid and informative.

Keywords: writing ability, indirect measurement, anchor essays, automated scoring

Introduction

Measuring a productive language skill such as writing is notoriously complex. Candidates' writing ability is usually assessed through written products demonstrating their performance on a writing task. As illustrated in several studies over time (Breland, Camp, Jones, Morris, & Rock, 1987; Godshalk, Swineford, & Coffman, 1966; Knoch, 2011; Weigle, 2002), the reliability and validity of writing assessments are often questioned. For instance, raters tend to disagree on the quality of the same piece of writing, which impairs reliability, and the discussion of the authenticity of writing assessments is a typical validity issue.

However, techniques such as indirect measurement and the evaluation of essays using an analytic rating procedure accompanied by automated scoring techniques can account for a valid and reliable assessment of specific aspects of the writing process and its end result: a written product. Therefore, a clever mix of assessment techniques can provide for a sound and informative evaluation of writing ability, as is argued in the last paragraph of this chapter.

Indirect Assessment of Writing

To overcome these issues, *indirect writing tests* were developed in the 1960s as an alternative for retrieving information on a candidate's writing ability. These tests are aimed at eliminating rater effects by offering objective tests on components of writing ability, such as spelling or grammatical fluency. Since indirect tests rely on the assumption that writing ability can be deducted via other skills, most research has focused on the correlation between test scores on indirect and direct measures of writing, stating that a high correlation between the two scores validates the use of an indirect instead of a direct measure. Table 1 summarizes a sample of these studies.

Table 1 A Sample of Previous Studies on the Validity of Indirect Writing Assessments

Study (year)	Age of pupils	Number of pupils	Correlation direct and indirect measure
Godshalk et al. (1966)	16–17	646	0.71–0.77
Wesdorp (1974)	12	213	0.67–0.68
Breland and Gaynor (1979)	18	234–926	0.56–.074
Breland et al. (1987)	>18	267	0.56–0.66

Nevertheless, instead of considering indirect tests as substitutes for active writing tasks, perhaps a more valid application for these objective tests is to use them to evaluate different aspects of the writing process. In the 1980s, studies on cognitive writing processes changed the focus for research on writing. Nowadays, writing is no longer considered a single action, but rather as a complicated process in which different components interact. One of the most popular models for the writing process, presupposing interaction among the task environment, long-term memory, and working memory, is shown in Figure 1 (Flower & Hayes, 1981).

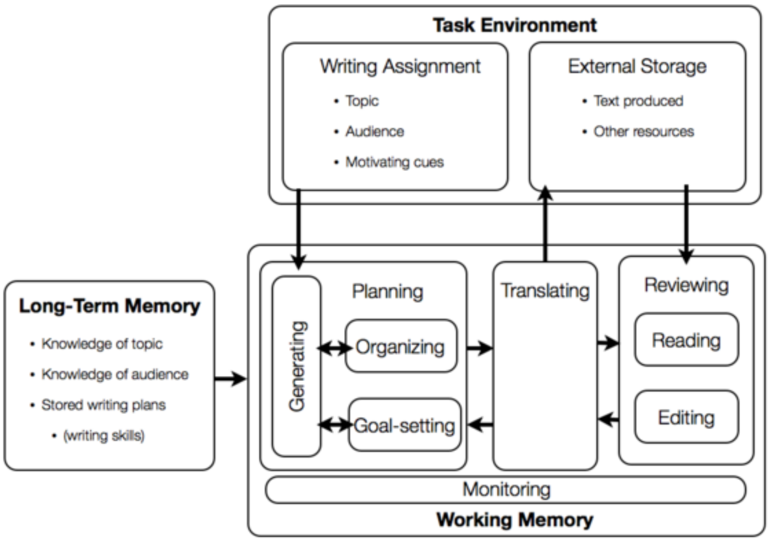

Figure 1 Model of the cognitive writing process by Flower and Hayes (1981)

Evaluation of a New Format for Revision Tests

A popular form of an indirect writing test is the *revision test*, where pupils are asked to correct a text supposedly written by a peer. When mapping indirect writing tests to the model composed by Flower and Hayes, this test assesses the part of the writing process referred to as *reviewing*, where the writer reads and edits his or her text. Feenstra and Heesters (2011a, 2011b) developed a new version of this test as a pilot, changing the multiple-choice format into a semi-open-ended version. In this new format, pupils are asked to actively revise a peer-written text by deleting or adding words, changing tenses, correcting congruence, et cetera. The sentences to be corrected (i.e. containing errors) were indicated by underlining. Table 2 lists the various options for correction. Since its format is more productive and less directive than the original multiple-choice version of the test, the adapted test is thought to be a more natural representation of reviewing a text.

Table 2 Correction Options in Revision Test

Afgelopen zaterdag ging ik naar mijn oma ~~gegaan~~.	**Deleting**
zijn Mijn hobby’s ~~is~~ tekenen, judo en gamen.	**Correcting**
Als ik vrij ben, ik ga graag voetballen.	**Switching**
naar We gingen eerst buiten. Daarna maakten we teams.	**Adding**

To evaluate the new format, both versions of the test (old and new) were incorporated in an incomplete test design. A representative sample of 80 primary schools participated in the study, resulting in a sample of 1,600 pupils. Table 3 reports the results of the pilot study, comparing the test characteristics of the semi-open-ended test version to those of the multiple-choice version.

Table 3 Results on Pilot Study Semi-Open-Ended Writing Test

	Old	New
Reliability	0.80[a]	0.83[a]
Difficulty (*p* value)	0.73	0.62
Discriminating power	3.19	2.53

Note: [a]Estimate for 50-item test using the Spearman-Brown formula.

Given the above, a semi-open-ended indirect writing test appears to be a reliable tool for assessing specific components of the writing process such as reviewing. Except for items on revision skills, it might also be possible to construct item formats with which other aspects of the writing process can be assessed. Paired with a writing assignment, an indirect writing test can therefore be a useful addition to a valid and reliable assessment of writing.

The Use of Anchor Essays

When assessing writing via a writing assignment, several different rating procedures are available to evaluate the essays. The most commonly used procedure in classroom assessment is *holistic scoring*, in which raters assign a score to an essay based on their overall impression of the writing performance (van den Bergh, 1990; Weigle, 2002). A more condensed form of this rating procedure is the *primary trait* approach. The focus in this method is merely the extent to which the essay reaches its communicative goal (Lloyd-Jones, 1977; van Gelderen, Oostdam, & van Schooten, 2010). Since raters are asked only to give one overall evaluation, both methods demand relatively little time and effort. As a result, however, these methods do not provide many details on the ability being measured.

Within an *analytical* rating procedure, different aspects of the writing product are evaluated, enabling a detailed report on writing ability. This analytical method was used in the Dutch National Assessment in Education, where a group of raters used an analytical rating scheme, assessing different aspects of writing (Krom et al., 2004).

One of the objectives of the analytic evaluation is to alleviate the task of raters by having them answer simple yes/no questions on features of the essay. Consequently, raters only have to identify certain features of the text (scoring), while the actual assigning of values (grading) is done within the data analyses. Because of the relatively simple task for the raters, it was believed that this method would provide high rater agreement. However, analyses show that the inter-rater reliability was rather low for some of the aspects (Krom et al., 2004).

Adjusting the Rating Procedure

A writing assessment consists of numerous elements, all of them possible sources of construct-irrelevant variance (Messick, 1989): for example, writing task, rating procedure, and rater characteristics. Although recognized, not all of these sources can be eliminated easily. For example, task effects can be ruled out by dramatically increasing the number of tasks given to a student, and rater effects by increasing the number of raters per essay. However, these methods are generally considered unsuitable, given the extra time and effort they would require of both students and raters. Therefore, most studies focus on altering the rating procedure to improve the reliability and validity of a writing assessment.

To achieve high reliability, raters should agree to a great extent on the scores assigned to essays. Providing the raters with an empirically constructed reference, or benchmark, that they can use to compare the quality of the writing products to be assessed could therefore prove to be beneficial for the agreement between raters, and thus have a positive influence on the rater reliability and validity of the writing scores. An empirical way of providing such a reference is constructing a rating scale illustrated with several examples of writing products, each representing a specific score point. The exemplars are taken from a sample of essays evaluated by multiple assessors and vary from a poor performance on one end of the scale to an excellent performance on the other end of the scale.

Van den Bergh and Rijlaarsdam (1986) developed a method to construct such a rating scale. The authors described all the steps needed to create rating scales for different aspects of writing. According to van den Bergh and Rijlaarsdam, using a rating scale with anchor essays has two main advantages over the use of an analytic rating procedure.

First, the exemplars on the rating scale serve as reference points, supporting the raters in their rating task and reducing instability in their rating. Moreover, using a fixed standard allows scores to be compared between pupils and classes or scores to be monitored over time. In the context of a national assessment, anchor essays can be particularly useful for illustrating different levels of achievement.

In fact, anchor essays were used in earlier cycles of the national assessment for writing (Sijtstra, 1997; Zwarts, 1990), but were eventually replaced in the next cycle owing to their complicated scoring instructions.

Evaluating a Rating Scale with Anchor Essays

Feenstra (2010b) reported on the use of a rating with anchors essays to improve the inter-rater reliability. In this study, three different essay tasks were selected from the pool of tasks in the Dutch national assessment, covering a broad scope of text goals and text genres. Five Dutch primary schools representing different regions, school sizes, and denominations volunteered to participate in the study. A total of 584 pupils, age 8 to 12, participated. In total, 1,476 essays were collected. All essays were digitalized (i.e., retyped, maintaining layout, typos, and punctuation) to facilitate reproduction and distribution. Moreover, handwriting quality can influence the assessment of other aspects of text quality (De Glopper, 1988). Presenting the essays in typescript eliminates this unwanted effect. As in the previous cycles of the national assessment, three aspects of writing were to be rated, as shown in Table 4.

Table 4 Categorization of Writing Aspects Used within the Study

Aspect	Description
Content	Essential content elements, focus on text goal and public
Structure	Composition, layout, coherence, cohesion
Correctness	Syntax, spelling, punctuation

The procedure described by van den Bergh and Rijlaarsdam (1986) was adopted to compose a rating scale with anchor essays for each aspect per task (Feenstra, 2010a), the result being a rating scale with three anchor essays representing specific ability levels (Figure 1).

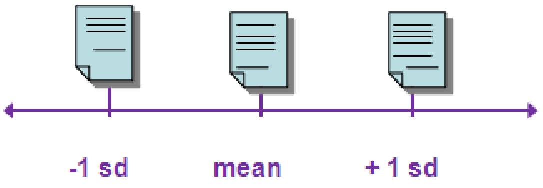

Figure 2 A rating scale with three exemplars

To select the anchor essays, four expert raters first agreed upon the average essay and then evaluated a sample of essays. The anchor essays for each score point were then selected based on their empirically defined value as an exemplar essay: agreement among the four different raters.

A total of 26 raters scored a sample of 150 essays in an incomplete design, to evaluate the quality and usefulness of the new rating procedure compared to the existing method. Each rater was assigned to one of two conditions, where condition 1 represented the existing analytical rating procedure, and condition 2 represented the adjusted version of the original procedure, consisting of the analytical scale *plus* the additional rating scale with anchor essays. Each essay was scored in both conditions, by a minimum of two out of the 13 raters assigned to each condition. In Table 5, reliability scores are presented per aspect.

Table 5 Inter-Rater Agreement (Gower's Coefficient) for All Raters

Aspect	Condition 1 Analytical	Condition 2 analytical + anchors
Content	.85	.84
Structure	.76	.81*
Correctness	.76	.77

*significant (p = 0.008)

As shown in Table 5, the aspects Structure and Correctness seem to generally benefit from the addition of anchor essays to an analytic rating scale. However, the improvement in inter-rater reliability is modest and significant only for the aspect Structure.

On the Use of Anchor Essays for Different Aspects of Writing

Text structure was found to be the only aspect for which the use of anchor essays significantly improved reliability. It might well be that for this aspect in particular having a complete essay as a reference for scoring is beneficial. Apart from the structure within sentences, text structure can be evaluated only by considering the text as a whole, which is encouraged by comparing essays to an anchor. For example, when evaluating a letter, the layout and formal structure of the text are important features that should be present not only in one or two parts of the text. Instead, they should form the basis of the text structure.

An aspect such as Content, however, is more or less locally assessed within a text and less dependent on the overall text quality. Different content elements are detected in the text and scored accordingly: the higher the number of elements that are present, the higher the score. This could be the reason that this aspect did not benefit from the comparison to anchor essays when assessing it. To assess this particular aspect, a detailed analytical procedure seems to be the best option.

In a way, the same holds for the aspect Correctness. This aspect actually requires the impression of the whole text to be taken into account, but as with Content, the elements diminishing the correctness can be more or less counted individually. Although this might sound straightforward, raters tend to disagree relatively strong when scoring this aspect. Apparently, differences in severity still come across, despite the supposed objectivity of the items. These difference can be overcome when automatically scoring specific text features. In the past decades, developments in computational linguistics, artificial intelligence, and psycholinguistics, have enabled the rise of techniques to analyze text features automatically. Several tools for automated essay scoring (AES) have been developed, and many validation studies have been reported. Instead of automatically providing an *essay* score, programs for text analyses could provide a score on different *text features*, thus contributing to a score for the aspect Correctness.

Furthermore, when considering the actual anchor items (i.e., the items where raters were prompted to place an essay on the scale) as individual items, an inter-rater agreement of .82 is achieved for each aspect. However, these figures cannot be interpreted reliably yet, because the raters were led to their final judgment by answering the analytical questions.

Further studies have to be conducted to gain insight into the individual strength of the anchor items. Still, these one-item assessments look promising and might well be developed into useful tools for classroom assessment because of their efficiency (cf. pair-wise comparison: Pollitt, 2004).

Using a Combination of Methods to Assess Writing

As shown in the studies mentioned in this chapter, different aspects of writing ability require different assessment methods. Since a writing product reveals very little about the cognitive processes taking place while producing the text, an objective test on certain components of the writing process (e.g., revision) can be a valuable addition.

Such an indirect writing test can account for a reliable assessment of specific aspects of writing, shifting the focus from solely the product of the writing process to other relevant components and thus adding to the validity of a writing assessment. Furthermore, while the analytic evaluation of text structure benefits from the use of anchor essays, adopting merely the analytic questions is sufficient when assessing the content of a text. Automated text analyses, to conclude, can help in objectively scoring certain text features considered important to text quality, thus contributing to a score for correctness. Hence, assessing writing is not a matter of choice: a decent writing assessment should incorporate a mixture of assessment techniques—and benefit from it.

References

Breland, H., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill*. New York, NY: College Entrance Examination Board.

Breland, H., & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement*, *76*, 119–128.

de Glopper, K. (1988). *Schrijven beschreven. Inhoud, opbrengsten en achtergronden van het schrijfonderwijs in de eerste vier leerjaren van het voortgezet onderwijs* [Writing described. Content, outputs and background of writing education in the first four years of secondary education] (Dissertation UvA). Den Haag, the Netherlands: SVO.

Feenstra, H. (2010a, June). *Opstellen langs de meetlat. De constructie van een beoordelingsschaal voor schrijfproducten* [Constructing a rating scale for writing products]. Poster presentation at the Onderwijs Research Dagen [Educational Research Days] 2010, Enschede.

Feenstra, H. (2010b, November). *Assessing writing ability: Using anchor essays to enhance reliability*. Paper presented at the 11th AEA-Europe Conference, Oslo, Norway.

Feenstra, H., & Heesters, K. (2011a, May). *Assessing writing through objectively scored tests: A study on validity*. Paper presentation at the 8th EALTA Conference, Siena, Italy.

Feenstra, H., & Heesters, K. (2011b, June). *Objectieve schrijfvaardigheidstoetsen: een onderzoek naar validiteit* [Objective writing tests: a study on validity]. Poster presented at the Onderwijs Research Dagen [Educational Research Days] 2011, Maastricht.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York, NY: College Entrance Examination Board.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writin*g, *16*, 81–96.

Krom, R., van de Gein, J., van der Hoeven, J., van der Schoot, F., Verhelst, N., Veldhuijzen, N., & Hemker, B. (2004). *Balans van het schrijfonderwijs op de basisschool. Uitkomsten van de peilingen in 1999: halverwege en einde basisonderwijs en speciaal onderwijs* [Report of the national assessment on writing education in primary schools. Outcomes of the surveys in 1999: Mid and end of primary education and education for special educational needs]. Arnhem, the Netherlands: Cito.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–68). Urbana, IL: National Council of Teachers of English.

Messick, S. (1989). *Validity*. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Pollitt, A. (2004, June). *Let's stop marking exams.* Paper presented at the IAEA Conference, Philadelphia, PA.

Sijtstra, J. (Ed.). (1997). *Balans van het taalonderwijs aan het einde van de basisschool 2. Uitkomsten van de tweede taalpeiling einde basisonderwijs* [Report of the language education at the end of primary education 2. Outcomes of the second language survey end primary education]. Arnhem, the Netherlands: Cito.

Van den Bergh, H. (1990). Schrijfvaardigheid getoetst in het centraal schriftelijk eindexamen [Assessing writing ability within the central written examinations]. *Levende Talen, 451*, 225–229.

Van den Bergh, H., & Rijlaarsdam, G. (1986). Problemen met opstelbeoordeling? Een recept [Issues with essay evaluation? A recipe]. *Levende Talen*, *413*, 448–454.

Van Gelderen, A., Oostdam, R., & van Schooten, E. (2010). Does foreign language writing benefit from increased lexical fluency? Evidence from a classroom experiment. *Language Learnin*g, 61, 281–321.

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

Wesdorp, H. (1974). *Het meten van de produktief-schriftelijke taalvaardigheid. Directe en indirecte methoden: 'opstelbeoordeling' versus 'schrijfvaardigheidstoetsing'* [Measuring productve-written language ability. Direct and indirect methods: 'essay rating' versus 'writing tests']. Purmerend, the Netherlands: Muusses.

Zwarts, M. (Ed.). (1990). *Balans van het taalonderwijs aan het einde van de basisschool. Uitkomsten van de eerste taalpeiling einde basisonderwijs* [Report of the language education at the end of primary education 2]. Arnhem, the Netherlands: Cito.