Assessing extensive reading through written responses and comprehension tests

by Mark Brierley (Shinshu University)

Abstract

This paper considers the extensive reading construct and examines two kinds of assessment—written responses to books and comprehension tests—with particular reference to the concepts of backwash and formative assessment. The use of written responses must be carefully considered and sparingly applied to ensure that we are not demotivating students. Comprehension tests are based on a slightly different construct and may present a more attractive option, if challenges in their construction and delivery can be overcome.

Keywords: extensive reading, assessment, backwash, learner autonomy, formative assessment, comprehension tests

要旨

本論では多読での構成概念（construct）を考察するとともに、とくにバックウォッシュと形成的評価の概念に言及しながら、筆記での応答（written responses）と理解度テスト（comprehension test）の二つの評価方法を検証した。筆記での応答は綿密に作成されるとともに、学生の学習意欲を下げないように注意深く実施されなければならず、理解度テストはやや異なる評価ターゲットに基づいているため、その作成の労力や実施がより簡便になれば、もっと魅力的な方法が提供できると言えよう。

キーワード 多読、評価、バックウォッシュ、学習者の自律、形成的評価、理解度テスト

Assessing reading has been described as impossible (Brown, 2004) and the assessment of extensive reading (ER) has been deemed undesirable (Day & Bamford, 1998; Sakai, 2008). However, some form of assessment is often essential as ER is implemented by individual practitioners or across institutions (Robb, 2008). While noting that the assessment of ER may be simultaneously impossible, undesirable and essential, this paper considers two kinds of assessment: written responses and comprehension tests, taking into account the fundamental assessment concepts of construct validity, reliability, backwash, and formative versus summative assessment. First, let us briefly highlight those concepts.

Key Concepts

Construct Validity

According to (Biggs, 1999) it is first necessary to clarify exactly what learning outcome we seek to assess in a construct. Construct validity will then inform us whether our assessment tool is actually measuring what we seek to measure. Day and Bamford (1998) suggest that "the goal of extensive reading is not merely reading improvement, but for students to become independent readers" (p. 90). This paper will later elaborate the ER construct, but let us begin with a simpler definition: "students reading a lot of easy, enjoyable books" (Helgesen, 2005, p. 25). A focus on the words "a lot" suggests that we should assess students on how much they read. Bruton (2002) discusses a number of interpretations of this, but let us consider word count as we refer to other assessment concepts.

Reliability

> *"It is common in assessment for a trade-off between validity and reliability rather than the possibility of optimising both"*

If we measure how much students read by asking them, we may wonder whether their answers are accurate. If some students honestly record every book that they have read from cover to cover, while others make lists of any books that they could find the names of, our assessment will not be reliable. It is common in assessment for a trade-off between validity and reliability rather than the possibility of optimising both (Slomp & Fuite, 2004). In our case, it is difficult to reliably assess the construct of students reading "a lot". On the other hand, we can more reliably assess the quality of our students' written book reviews, but this may not tell us how much they have read.

Backwash

Backwash refers to the effect that testing has on learning (Hughes, 1989). For example, if the amount students read is an assessment criterion, students are likely to read more in order to get higher scores. This would be seen as positive backwash. However, students may ask us exactly how much they should read, and then stop when they have reached a perceived "target". Other students may be demoralised at a target that seems unattainable and read nothing, leading to negative backwash. In other words, basing assessment on how much students read may result in mostly positive backwash, but setting an actual number of words may tend to have negative backwash.

Figure 1 shows the reading performance over a semester for a class of 30 first-year university EFL students. Students received credit for each book they reported reading, up to 10 books. In this data, we can see a cluster of students who read exactly 10 books. Mysteriously, no student read nine books.

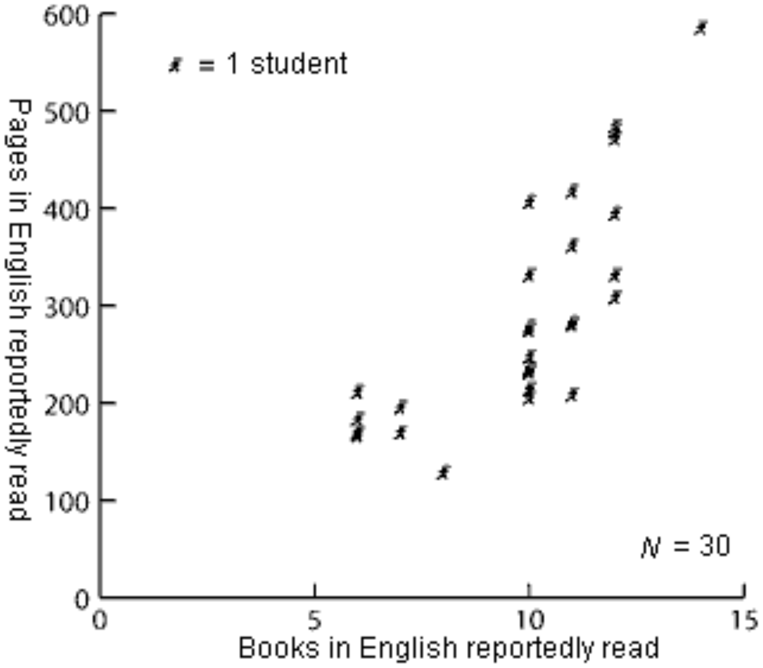


Figure 1. Book and page count for 30 first-year Japanese EFL students in a one semester ER program

Formative vs. Summative Assessment

Traditional, summative testing comes at the end of a course and determines whether a candidate can perform a skill or has attained some knowledge (Bachman, 1990). Formative assessment, on the other hand, evaluates the process of learning and ascertains whether a student is taking the necessary steps to reach the learning objective. Although I do not wish to compare students to production line workers, classrooms can appear to have manufacturing parallels, in which quality control inspectors check products from each production line, rejecting any defects, while quality assurance looks at everything in the process from the acquisition of raw materials to the final operators' performance to ensure that only high-quality products are made. In the same way, formative assessment can identify failing learners during a course rather than at its end when it is too late.

Often, assessment requires teachers to show the results of their teaching, rather than measure the progress of students' learning. Formative assessment relates to ideas of learner autonomy, which are also central to ER (Black & William, 1998; Day & Bamford, 1998). We may at first assume that assessing word count would be summative, as we are adding up a score. However, depending on our construct, it may be seen as formative assessment, if reading is taken to be the methodology and not the ultimate aim of ER.

The ER Construct

Brown (2004) defines ER as anything over one page, characterised by focus on meaning, rather than form, and top-down rather than bottom-up processing, stating, "The purposes of assessment are usually to tap into a learner's global understanding of a text" (p. 190). While this provides part of our construct for summative assessment, "ER" is being used more to define a type of text than a methodology. Alderson's (2000) description is closer to that of many ER practitioners:

> Reading is, for many people, an enjoyable, intense, private activity, from which much pleasure can be derived, and in which one can become totally absorbed. Such reading [...] is difficult if not impossible to replicate within an assessment setting. The intervention of questions, tasks, outcomes, between the reader and the test is likely, for some at least, to be disruptive and to create a self-consciousness which destroys the very nature of the event. We need to acknowledge that in such settings, for some purposes, the assessment of reading may be both difficult and undesirable. (p. 28)

He also warns:

> Certain aspects of reading—like appreciation, enjoyment and individual response—may not be measurable and need to be assessed, or reported, in different ways. This does not mean that they do not belong in our constructs, but that we need to be aware that the tests we produce will inevitably underrepresent those constructs. (p. 123)

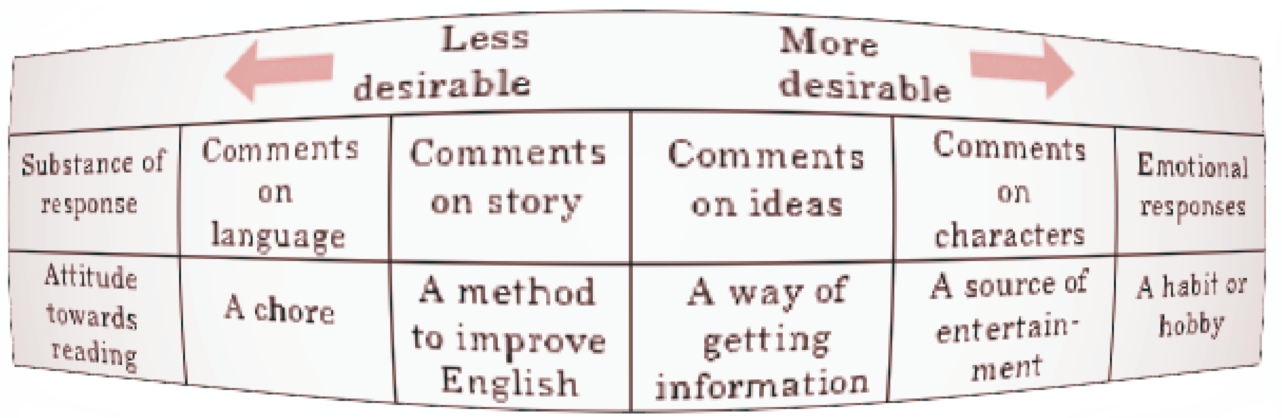
In addition to specifying reading proficiency outcomes, a maximal ER construct could describe students' attitudes and their habits, for example whether they read regularly. Teachers might require students to demonstrate an understanding of the rationale and methodology of ER. In Schmidt (2007), Mason stresses the importance of students understanding why they are reading in order for them to justify investing their valuable time in the practice. According to Black and William (1998), "a student who automatically follows the diagnostic prescription of a teacher without understanding of its purpose will not learn." (p. 54) Students might be required to know:

- to skip words that they don't understand
- to avoid translating
- when to stop reading a book
- which books they enjoy
- their own reading level
- concepts such as "acquisition," "collocation," "context"

> *"the ideal student would see reading as a habit or a hobby rather than a homework assignment"*

We may also expect students to relate to the characters within stories, and engage with the content of their reading rather than the language. In terms of attitude, the ideal student would see reading as a habit or a hobby rather than a homework assignment. The goal, in other words, is for students' activity to become independent of classes and assessment. Mason (in Schmidt, 2007) points to the irony that we must first force students to read in order that they may read voluntarily. Sakai (2008) periodically asks students the vague question of how their books are, and gives lower scores to students with replies such as "I can read it," higher scores to students who report on the content of the story such as "It's about..." and the highest scores to students with emotive responses such as "It's great". Clearly there is a logical problem with giving high assessments to students who are not fulfilling assessment criteria, and such assessment strategies may be at odds with transparent assessment policies.

Table 1. *One way of rating EFL student extensive reading attitudes recommended by Sakai (2008)*



	← Less desirable			More desirable →	
Substance of response	Comments on language	Comments on story	Comments on ideas	Comments on characters	Emotional responses
Attitude towards reading	A chore	A method to improve English	A way of getting information	A source of entertain-ment	A habit or hobby

While it is very easy to ask students whether they enjoyed a book, or ask students about their behaviour, their answers may not be honest, and basing students' grades on this unreliable data may not be appropriate. Alderson (2000) raises other issues related to self-assessment, for example, the amount of training required and the different power relationships that can make students and teachers uncomfortable. Even if they do not show what students are doing, correct or ideal answers to questions about habits, attitudes, motivation, and methodology will at least show that students know what behaviour is expected of them and why.

Written Responses to Books

Students are often required to produce some written response in order to show that they have read a book. This may take a range of forms from short, simple comments to reviews or summaries. It is important to clarify what purpose such writing serves and what exactly we are assessing. For example, Brown (2004) cites Imao's (2001) four criteria for assessing summaries, asking whether students 1) accurately express the main idea and supporting ideas; 2) write in their own words; 3) organise the work logically, and 4) display language facilities enabling clear expression. Brown points out that of these four criteria, only the first should be considered if we are assessing only reading.

In terms of reliability, in order to show that a student has read a book, a written report must contain information that could only have come from reading it. The assessor must be familiar with the book to judge this, which may be difficult where libraries have hundreds of titles. Students may write convincingly having read the book in their L1, seen the movie, looked at the cover or the pictures, read a review from another student, or even read the blurb from the back of the book. To show students had read the whole book, a written response would have to show a holistic knowledge, or contain elements from several places in the book. In addition, students would need to be taught how to write such a review or summary. Skills in such written genres may be valuable to students, and the review or summary may also have a useful function in helping students recall what they have read. This, however, does not prove that they have done the reading in the target language.

It is tempting to see written responses to books as summative assessments, if they are intended to check whether students have read a book and effectively completed a piece of extensive reading. Completing their first book in English may be a great achievement for students, and may fulfil part of Brown's definition of ER. However, depending on the construct we choose, it may be more helpful to see this as formative assessment, showing how students are reading, as an ongoing process, rather than simply what they have finished reading.

If our students in the early stages are reading at a suitable level, the books may only be 500 words long, which, even at a slow speed of 50 words per minute, should take 10 minutes (Brierley, 2007). Any written report in English is likely to take several times this. The satisfaction that should come at the end of a book may be replaced with the dread of having to write about it. It can be assumed that our students can write more quickly in their L1, and students may spend less time writing and more time reading if permitted to write reports in their own language. We may question what place L1 writing has in an English course, and whether reading such reports is a good use of teachers' time. A study by Mason and Krashen (2004) suggested that written activities made no difference to acquisition, and in fact may have been detrimental, as the time was not being spent reading. Any kind of written report is likely to fall somewhere between insubstantially validating that the student has read the book, and overly burdening the student, and there is a high chance that it will be both. Ruzicka and Brierley (2008) found that two teachers had abandoned a book report system introduced in 2006, each for one of these apparently opposite reasons.

On the other hand, written responses may tell us a great deal about students' feelings towards what they are reading, and indicate their level of engagement with texts. Brief comments such as "I cried when I read this book" or "I never knew reading in English could be fun" may tell us more than a detailed summary or a critical review. In addition, if reviews or responses are visible to other students, as is possible in various on-line systems (for example Brierley, Wakasugi & Sato, 2008; Brown, 2009; Sonda, 2009), these responses may provide motivation for other students and a sense of belonging to a larger reading community (Schmidt, 2007).

Tests

There is a wealth of literature on the creation of reading proficiency tests, and tests on ER practice and methodology can play a part in the ER assessment battery. A common desire among practitioners is for a test at the end of each book to show that students have read it, although such tests have been criticised by ER proponents. Putting "no comprehension questions" as his fourth ER principle, Prowse (2002) quotes Widdowson (1979): "Comprehension questions... commonly require the learner to rummage round in the text for information in a totally indiscriminate way, without regard to what purpose might be served in doing so." However, such questions test scanning ability, not comprehension, and the quotation reflects the difficulty of making comprehension questions and the inferior quality of most items that masquerade as such. Consider the first three questions published at the back of *Marcel and the Mona Lisa* (1991), a 20-page Penguin EFL reading text:

```
1. In what month does the story start? (page 1)
2. Where are Marcel and Celine going on their summer holidays? (page 2)
3. The thief is looking for something. What? (page 4)
```

The answers — May, Los Angeles, and his car keys — can all be found (on line 5, line 3 and line 3 of the respective pages) without understanding the story, and are likely to be forgotten if the story has been comprehended, as none has anything to do with any other part of it. Such questions do not reliably tell us whether students have read the book.

Questions should focus on key points of the story. Questions about trivial details are poor indicators of comprehension (Buck, 2001) and in terms of reliability, questions about details may not show that a book has been read at all, particularly if the book can be referred to while the test is being taken. In terms of backwash, such questions may make students concentrate on details rather than the story while they are reading, or may discourage students from reading more books.

Questions should not interfere with each other. The testing literature calls for as many fresh starts as possible, and independent items (for example: Buck, 2001). This often leads to tests with short, de-contextualised items, which stand in sharp contrast to ER texts, in spite of a finding by Engineer (1977, as cited in Alderson 2000) that texts over 1000 words allow different abilities to be measured. Graded readers present much longer texts than usually fit into an exam, with a greater challenge to avoid asking questions that would give away the story or the answers to other questions.

Inferencing is necessary to "get" most stories, although Bloom's taxonomy suggests that asking students about inferred facts will be more reliable than inferred opinions. Better readers are likely to be better at inferencing, and we must clarify whether our tests are testing reading ability or comprehension. We probably wish to test whether students have read a particular book, rather than testing their reading competence.

Other recommendations on constructing questions can be found in Alderson (2000) and Buck (2001).

We must decide whether or not students can look at the book they claim to have read while they are taking the test about it. Clearly, comprehension of a book would be better tested if the student is not looking at a copy of the book (Alderson, 2000). Reed and Goldberg (2008) describe a system that generates paper tests to be administered under test conditions in class each week. On the other hand, many teachers consider the teaching of scanning or skimming to be important pedagogical goals, and require students to refer back to the book, for example, asking on which page a particular event took place, or asking students to write the second word in the third paragraph of page four (Stewart, 2008). A third approach allows students to refer to the books (Robb, 2008); if students can take the test online and unsupervised, they are likely to try to refer to the books anyway.

Next, we have to reckon with the likelihood that "All testing encourages cheating" (Sakai, 2008). This reliability issue can be addressed in online tests by drawing a small number of items from a larger pool for each book and by randomising answers to prevent students from copying answers from each other. However, it may still be possible for students who have read a book to help those who have not.

Alderson (2000) suggests that students often associate formal tests with previous failure. Rather than being demotivating, however, short tests may even provide positive motivation. Reed and Goldberg (2008) found that giving a two-minute, five-question test at the end of each book led to students reading more. This may not just be the extrinsic motivation of students socialised into working towards the test; it may even represent intrinsic motivation as some people like quizzes. It can, however, be argued that these tests reinforce students' negative attitudes towards reading, which we wish to change. "The natural response to a book is emotional or intellectual, and comprehension questions are neither of these." (Prowse, 2002, p. 141)

While a single, correct answer is a basic requirement of a comprehension question, the opportunity arises, particularly in online testing, to introduce items with multiple correct answers, such as: "What would you have done..." or "Who do you like better...". Such questions may be more engaging for the students, encouraging and validating opinions and emotions about the stories they are reading and appearing less test-like. They may, however, make students think that there is only one correct opinion, further expanding the control of the test examiner. Credit for such items could be given for any answer, in other words for having read the item.

Like written responses, these tests appear to be summative. Many students are socialised into completing a test at the end to show that they have completed an assignment, and for the instructor, tests are usually designed to demonstrate that a learning outcome has been met. Unless our goal is simply to read many words, we must again see these tests as formative assessment, indicating whether students are reading and moving towards the broader goals of ER.

> *"Rather than claiming that there is a best way to assess ER, this paper recommends that teachers carefully consider the construct in order to choose well from the range of imperfect assessment strategies and tools"*

Conclusion

Although assessment may appear at odds with ER, and has been deemed impossible and undesirable, this position is not helpful to practitioners. Rather than claiming that there is a best way to assess ER, this paper recommends that teachers carefully consider the construct in order to choose well from the range of imperfect assessment strategies and tools. Choices must also depend on the book, student, and teacher. Heavy emphasis should be placed on backwash and formative assessment. For many practitioners, the goal of ER is to foster readers who will continue reading after we finish teaching them. On top of linguistic goals, we may ultimately wish our students to see themselves as participants in the English language rather than detached observers of it. If we aim to change students' attitudes through ER, then much of our time should be spent assessing attitudes and factors that influence, or are influenced by them.

Written responses play an important part in most English classes, although we must be careful how we apply them to ER. Different kinds of responses are appropriate for different books and, if students are to write lengthy responses, they should be able to choose which book to write about. When possible, their writing should be directed to other students rather than just the teacher. Expecting students to write extensively about everything they read may make their reading much less extensive.

In spite of their poor reputation, well-designed, reliable, short quizzes may provide accountability in reading without being demotivating, although they represent a different ER construct to written responses. With the ever-growing list of extensive readers offered by publishers, the biggest issue is perhaps the sheer number of titles. A great deal of work has already been done on paper and online systems (Robb, 2008; Stewart, 2008; Reed & Goldberg, 2008). The availability of a battery of tests is clearly a major task that will be fulfilled with co-operation, collaboration, and consideration of the issues raised above.

The author is currently researching online systems that can assist the assessment of ER and can qualitatively evaluate both books and students in order to recommend suitable books to each reader.

References

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

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Biggs, J. (1999). *Teaching for Quality Learning at University.* Buckingham, UK: The Society for Research in Higher Education & Open University Press.

Black, P. & William, D. (1998). Assessment and classroom teaching. *Assessment in Education, 5 (1),* 7-74.

Brierley, M. (2007). Extensive reading levels. *JABAET Journal 11*, 135-144.

Brierley, M., Wakasugi, T. & Sato, H. (2009). An online system for assessing extensive reading. In A. Stoke (Ed.), *JALT2008 Conference Proceedings.* Tokyo: JALT. (add pages)

Brown, D. (2009). Using Librarything.com to promote extensive reading. In *add name of editor* (Ed.). *JALT CALL 2008 Conference Proceedings.* Tokyo: JALT. (add pages)

Brown, H. D. (2004). *Language assessment: Principles and classroom practice.* White Plains NY: Longman.

Bruton, A. (2002). Extensive reading is reading extensively, surely? *The Language Teacher, 26 (11),* 23-25.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Day, R. & Bamford, J. (1998). *Extensive reading in the second language classroom.* Cambridge: Cambridge University Press.

Helgesen, M. (2005). Extensive reading reports - Different intelligences, different levels of processing. *Asian EFL Journal 7*(3) 25-33. Retrieved January 23, 2010 from http://www.asian-efl-journal.com/September_05_mh.php

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Mason, B. & Krashen, S. D. (2004). Can we increase the power of reading by adding more output and/or correction? Retrieved January 23, 2010 from http://www.extensivereading.net/er/online.html

Prowse, P. (2002). Top ten principles for teaching extensive reading: A response, *Reading in a Foreign Language. 14* (2) 141-145. Retrieved January 23, 2010 from http://nflrc.hawaii.edu/rfl/October2002/discussion/prowse.html

Rabley, S. (1991). *Marcel and the Mona Lisa.* Harlow: Pearson Education.

Robb, T. (2008). The reader quiz module for extensive reading. Retrieved January 21, 2010 from http://moodlereader.org/moodle/mod/resource/view.php?id=492

Reed, K. & Goldberg, P. Integrating quizzes with extensive reading. Presentation at the 34th Annual JALT International Conference, November 2, 2008. Tokyo: National Olympics Memorial Youth Center.

Ruzicka, D. & Brierley, M. (2008) Selling ER: Investigating factors in classroom management that affect reading performance. *Journal of Humanities, Shinshu University 2*, 223-228.

Sakai, K. Presentation to teachers at Shinshu University, March *, 2008. *ADD CITY*

Schmidt, K. (2007). Five factors to consider in implementing a university extensive reading program. *The Language Teacher 31*(5) 11–14.

Sonda, N. An integrative college reading course: An action research. Presentation at the 2009 JALT Pansig Conference, May *DATE*, 2009. Chiba, Japan: Toyo Gakuen University, Nagareyama Campus.

Slomp, D. H. & Fuite, J. (2004). Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing, 9* (3) 190 - 207. DOI: 10.1016/j.asw.2004.10.001.

Stewart, D. (2008). Did they really read it? A website for checking. *Journal of the JALT Extensive Reading Special Interest Group 1*(2) 17-23.