ASIAN EFL JOURNAL
The EFL Professional's Written Forum

Article Title

Authentic, Performance-Based Assessment in ESL/EFL Reading Instruction

Author

Richard Lynch

Bio Data

Richard Lynch teaches in the graduate program in multimedia-assisted language learning at Woosong University, Korea. He has a Certificate in TESL from Concordia University, Montreal, an MA/TEFL from Southern Illinois University, and a Ph.D. in educational psychology/instructional technology from the University of Southern California. He has taught ESL/EFL in Canada, the U.S., Papua New Guinea, Japan, Thailand, and Korea. His research interests include learner performance and motivation in diverse distance learning contexts including second/foreign language learning.

Abstract

In recent years there has been a reaction against standardized objective testing and a movement toward authentic, performance-based testing in a variety of learning domains, including ESL/EFL reading comprehension. A number of measurement researchers have investigated and developed comprehensive test validity criteria which should apply to all testing, traditional or authentic. The advantage of performance-based testing resides in its potential to engender and sustain positive washback on the teaching and learning process. Such positive impact on the instructional process is not, however, a sine qua non of performance-based testing. It can only derive from a comprehensively valid interaction between the nature of the instruction preceding evaluation and the actual performances being assessed. Validity must be grounded in a range of interrelated factors which can all be subsumed under a general notion of construct validity. In order to maximize the validity of performance-based assessments both test designers and teachers need to be aware of

these factors and their interaction. Construct validity and its interrelated aspects are discussed and applied to the authentic, performance-based testing of reading comprehension. A reading exercise is presented and discussed as a representative instructional model that can be used to inform valid performance-based reading comprehension tests.

Introduction

In recent years, traditional standardized objective achievement tests, consisting primarily of multiple choice and matching items, have been generally criticized as inappropriate, invalid measures of students' academic competencies (O'Malley and Valdez Pierce, 1996; Michael, 1993). Such criticisms have initiated and sustained a movement toward authentic assessments in a wide variety of learning domains. This is as true in English as a second/foreign language (ESL/EFL) assessment as it is in areas such as science or mathematics assessment. In the area of reading comprehension, Bernhardt (1991) argued that " . . .contemporary standardized and classroom testing of reading comprehension, that is, tests in their current conceptualization that place readers on a generic measuring stick, has little, if any link to validity." She broadened her argument by asserting that students are assessed for a variety of purposes but little attention is given to the validity of the assessments in any of these contexts.

Other writers on ESL/EFL assessment have also noted the mismatch between measures of language competence and the actual communicative competence required in real world communicative interaction (Duran, 1988; Kitao & Kitao, 1996; McNamara, 1996; O'Malley and Valdez Pierce, 1996; Spolsky, 1995). The movement toward authentic, performance-based assessment in ESL reading comprehension, therefore, has been an attempt to achieve a more appropriate and valid representation of student communicative reading competencies than that derived from standardized objective tests.

As Michael (1993) pointed out, however, that there are a number of problems with the use

of performance-based assessment measures. He identified three difficulties that need to be considered. First is the problem of obtaining reliable results, i.e., inter-rater reliability. Raters of performance-based assessment tasks need to be trained to conform to an established set of scoring criteria. Such training, however, does not ensure consistency since raters may still interpret the criteria in different ways.

The second difficulty discussed by Michael concerns the adequacy of the sample of the indicators deemed to represent authentic performance. The range of performances in the knowledge domain being tested must be adequately covered in order to ensure validity of assessment. The representativeness of the assessment, in order to adequately capture truly authentic performance, may well require an extended testing period that is not feasible in most educational contexts (see also Messick, 1988). Michael's third difficulty is that of transfer and generalizabilty. This addresses the question of whether the tasks performed in the authentic assessment "carry over to other learning experiences or portray what already has been learned in a different but still important educational context" (Michael, p. 47).

Each of these difficulties with performance-based assessment has been addressed generally by Linn, Baker, & Dunbar, (1991) and, in terms of language testing, by Messick (1996). The former authors argued for an "expanded validity" in considering authentic assessment and propose eight criteria for the validation of performance-based assessments: consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. Messick (1996) discussed six criteria, or aspects, of validation, all linked under an overarching concept of construct validity, that, he argued, apply to all assessments whether standardized and objective or performance-based: the content aspect, the substantive aspect, the structural aspect, the generalizability aspect, the external aspect, and the consequential aspect. The remainder of this paper will apply the Messick validation criteria, as well as those of Linn, Baker, and Dunbar to a consideration of authentic, performance-based ESL/EFL reading comprehension assessment.

Aspects of Validity in Performance-Based Reading Assessment

Messick (1996) argued that validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. . . . Because score meaning is a construction that makes theoretical sense out of both the performance regularities summarized by the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as construct validity. (pp. 245, 246, italics in the original)

Inherent in this definition are Messick's six aspects of construct validity, which need to be distinguished and accounted for if any assessment is to be considered as a valid representation of competence. Each aspect is discussed below.

The Content Aspect

This aspect includes the three criteria Linn, Baker, and Dunbar (1991) referred to as "cognitive complexity", "content quality", and "content coverage". It entails specification of the construct domain to be assessed. In terms of reading assessment, it means, in very simple terms, that reading is the construct being assessed and not speaking, listening, or writing. More importantly, it means that the specific aspects of reading being assessed are clearly delineated. Several writers (Bernhardt, 1991; Kitao & Kitao, 1996; McNamara, 1996; Splosky, 1995) rightly argued that language assessment be firmly grounded in current knowledge in language learning theory. A great deal of research in both first and second language reading processes has distinguished between lower order and higher order reading skills (see for example, Alderson & Lukmani, 1989; Carrell, 1988; Eskey, 1986, 1988; Eskey & Grabe, 1988; Gagne et al. 1993; Grabe, 1991; Horiba, 1996; Hosenfeld, 1995; Tomlin & Gernsbacher, 1994).

Second language reading researchers distinguish between lower order, or bottom up, processing and higher order, or top down, processing (Carrell, 1988; Eskey, 1988). In fluent readers both bottom up and top down processing interact and complement one

another in the attainment of full comprehension. Gagne et al. (1993) subdivided the reading process into four subgroups: decoding, literal comprehension, inferential comprehension, and comprehension monitoring. The first two subgroups involve lower level, bottom up processing skills; the latter two subgroups involve higher order, top down processing skills. Decoding, composed of automated basic skills, is the procedure whereby readers "crack the code" of print and make it meaningful. At the level of literal comprehension the reader uses knowledge attained through the decoding processes to begin full comprehension of a text beyond the word level. Literal comprehension is composed of two processes.

The first is lexical access, which begins once the decoding process has activated the word precept in long term declarative memory. Since words possess both denotative and connotative (or contextual) meanings, it is through the process of lexical access that the reader selects the correct interpretation for the word in the context being read. Selection of the correct interpretation is dependent on the reader's store of declarative vocabulary knowledge. The second element involved in literal comprehension is parsing, the process whereby the reader combines word meanings through the syntactic and the linguistic rules of language to achieve meaningful ideas. Both lexical access and parsing, which are dependent on decoding skills, combine to provide literal comprehension (Gagne et al., pp. 269-275).

The third process involved in reading comprehension is inferential comprehension, a mix of automated skills, conceptual understanding, and strategies. Inferential comprehension is composed of three sub-processes: integration, which produces a coherent representation of a text; summarization, which functions to provide for the reader an overall representation of the writer's meaning which can be stored in declarative memory; elaboration, the process whereby the reader brings prior knowledge to bear upon the writer's meaning (Gagne et al., pp. 275-279).

The fourth and final component of the reading process is comprehension monitoring, a mix of automated skills and strategies. The function of comprehension monitoring in skilled reading is to ensure that reading goals are being accomplished effectively and efficiently. It

is composed of four sub-processes: goal-setting, strategy selection, goal-checking, and remediation (or correction) (Gage et al., pp. 279-280). As Grabe (1991) argued, "a description of reading has to account for the notions that fluent reading is rapid, purposeful, interactive, comprehending, flexible, and gradually developing" (p. 378). Such a description must also be reflected in valid performance-based reading assessment (also O'Malley and Valdez Pierce, 1996).

The problem with standardized objective tests of reading is that all too often they focus only on lower order discrete, non-integrated decoding skills, rather than on an integration of lower and higher order skills of the sort required in authentic reading. Tests which focus on discrete point grammar or decontextualized vocabulary items do not allow for a valid assessment of communicative competence in reading. Duran (1988, p. 109) identified four dimensions of communicative competence:

- grammatical competence - which includes mastery of vocabulary, word formation, sentence formation, and spelling all of which may be subsumed in reading comprehension under Gagne et al.'s decoding and literal comprehension processes,
- sociolinguistic competence - which includes mastery of the appropriate use of language in different contexts and the appropriateness of meanings and topics depending on context,
- discourse competence - which includes mastery of the cohesion and coherence devices employed to achieve unity in a text or piece of productive discourse,
- strategic competence - which includes the metacognitive strategies used to compensate for communication breakdowns and to ensure effective meaning construction.

Each of these four dimensions of communicative competence needs to be accounted for in performance-based reading comprehension assessment if such an assessment is to attain content validity.

Authentic, performance-based tests of reading comprehension must also ensure that the interaction of lower order automated processing skills and higher order interpretive skills is in fact operative in the respondents. This will involve selecting appropriate reading

assessment tasks which are appropriate in terms of learners' prior topic and linguistic knowledge. Messick (1996) pointed out that assessment tasks should not only be relevant to the construct domain, but should also be representative of the construct domain. That is, the tasks selected for the assessment should cover all the important areas of the specific knowledge domain being assessed. Construct relevance and construct representativeness can be attained by having subject matter experts review the tasks selected for testing. This raises a final aspect of domain that requires consideration. Not only should the cognitive domain, e.g., reading comprehension, be delineated, but also the authentic context in which the behavior is to be executed. A key to authentic assessment is that it simulate, as far as possible, the authentic behavior which learners will need to enact in real situations. Messick (1996) refers to the use of job analysis, task analysis, and curriculum analysis as means to delineate the boundaries and structure of the construct domain to be assessed.

The Substantive Aspect

This aspect relates to Linn, Baker, and Dunbar's (1991) cognitive complexity criteria.

Messick divided it into two points. First is the need for test tasks to not only sample domain content but also to adequately sample domain processes. Second is the need to ensure through empirical evidence that respondents actually engage in the sampled processes during task performance. Linn, Baker, and Dunbar (1991) argued that a potential advantage of performance-based assessments is that they emphasize higher order thinking skills as well as cognitive and metacognitive learning strategies. They go on to note, however, that it should not be assumed that complex cognitive tasks will necessarily require the use of complex cognitive processes by students. By way of an example, they noted that the construction of an open-ended proof of a theorem in geometry can be a cognitively complex task or simply the display of a memorized sequence of responses to a particular problem, depending on the novelty of the task and the prior experience of the learner. (p. 19)

A reading comprehension test needs to be assessed in terms of its difficulty, or novelty,

level for the students taking it. For example, students with a great deal of prior topic knowledge may well be able to perform reading-based tasks simply because of that prior knowledge and not because they have fully comprehended the specific text used in the assessment, thus invalidating the assessment.

The Structural Aspect

This aspect relates scoring of the assessment tasks to "what is known about the structural relations inherent in behavioral manifestations of the construct in question" (Messick, 1996, pp. 249-250). In terms of reading comprehension, therefore, task scoring should take into account the interactive relationships between lower order decoding skills and higher order interpretive skills (O'Malley and Valdez Pierce, 1996). Performance-based assessment scores should also reflect not merely the cognitive processes involved in reader-text interaction but also the task relevant behaviors required by the assessment. As Bernhardt (1991) pointed out, people read for a purpose. She noted that readers use information for a variety of purposes and that therefore reading tasks ought to be purpose-oriented. Performance-based reading assessment tasks should involve reading for an authentic purpose, or at least as authentic a purpose as any testing situation can muster.

The Generalizability Aspect

Here the concern is that a performance assessment should "provide a representative coverage of the content and processes of the construct domain" (Messick, 1996, p. 250). As Messick noted, "the issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning" (p. 250). Linn, Baker, and Dunbar (1991, p. 19) commented on the highly task dependent nature of performance. They continued by observing that "the limited generalizability from task to task is consistent with research in learning and cognition that emphasizes the situation and context-specific nature of thinking." Linn et al. suggested that this limited generalizability across tasks be addressed by increasing the amount of performance assessments for each student. However, as Messick pointed out, generalizability should be considered in terms of both reliability and

of transfer. Reliability refers to "consistency of performance across the tasks, occasions, and raters of a particular assessment, which might be quite limited in scope" (Messick, p. 250). An assessment with only a few tasks may attain high reliability in cross-task complexity, but will diminish construct validity because of minimal construct domain coverage. On the other hand, transfer refers to consistency of performance across tasks that are representative of the broader construct domain. That is, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of. (Messick, 1996, p. 250).

As a result, given the time required for performance tasks, there is, as Messick pointed out, "a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation" (p. 250). This is a major problem with validity in performance-based assessments and one that needs to be carefully considered by teachers and administrators involved in designing and implementing such assessments..

The External Aspect

This refers to the criterion validity of the assessment or "the extent to which the assessment scores' relationships with other measures and non-assessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed" (Messick, 1996, p. 251). Messick grounded this external aspect of construct validity in the relationships between the performance assessment scores and the actual uses of, or decisions made on the basis of, such scores. The focus here is on how useful the scores are in providing the specific information that the assessment was designed to acquire. In ESL/EFL reading assessment, such information may range from placement decisions, achievement decisions, and/or course/program evaluation decisions.

The Consequential Aspect

Linn, Baker, and Dunbar (1991) stressed the importance of the consequences of assessment and defined consequential validity as both the intended and unintended effects of assessment on teaching and learning. Messick broadened the definition by noting that the consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness and distributive justice . . . as well as to washback. (1996, p. 249; also Messick, 1989, 1988)

The concept of "washback" referred to by Messick is one frequently used in language teaching to refer to the effects of testing on teaching and learning, i.e., "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning" (Messick, 1996, p. 241). Language teaching professionals who favor performance-based testing argue that standardized objective tests promote classroom activities which stress inauthentic language performance, an example of negative washback. On the other hand, they argue, authentic, performance-based assessment promotes the use of communicative tasks in the classroom that better approximate genuine communication, an example of positive washback (Bernhardt, 1991; Kitao & Kitao, 1996; O'Malley & Valdez Pierce, 1996). Messick (1996) pointed out that the link between testing and washback is not so neat.

He noted that poor tests may have good effects and that good tests may have poor effects due to other factors within the educational system. As well, a test may influence instructional content but not instructional practice. An example would be that of a teacher who is committed to a communicative approach to language instruction who finds herself in an educational context committed to a more traditional grammar-based approach to instruction. The assessment measures may be comprised of discrete point grammatical items but the teacher may well teach the targeted grammar in a communicative, context-performance-based manner. Of course, the opposite may well occur, with a teacher teaching grammar in a traditional way in a communicative curriculum. The point is that the test itself does not necessarily washback on teaching method.

Positive washback, however, ought to be a goal of test designers and teachers. Language teaching and language testing ought to be grounded in current knowledge of the cognitive and metacognitive processes involved in language learning. If such grounding is achieved, then both classroom practice and testing can be linked in appropriate and valid ways. Messick (1996) argued that the possibility of negative washback, or any adverse consequences on individuals or groups, can be diminished if construct under-representation and construct irrelevant variance are minimized. That is, test developers need to ensure that the reading processes discussed above are adequately covered in the assessment and that there is nothing irrelevant in both the assessment content and procedure that might interfere with a test taker's demonstration of reading competence. If important aspects of the construct are under-represented on an assessment, teachers may overemphasize in their teaching those aspects which are well-represented, thus resulting in negative washback. As well, discrete tests of vocabulary, for insistence, since they stress decontextualized vocabulary knowledge over communicative competence, may cause teachers to focus on irrelevant difficulty as opposed to focusing on the development of communicative competence.

Linn, Baker, and Dunbar (1991) argued that the fairness of the assessment must also be considered. The sue of fairness is pertinent not only in the selection of assessment tasks but also in how the responses are scored. In terms of reading assessment, students' prior knowledge of topics, sensitivity to topics, and familiarity with and prior exposure to reading task types all figure in fairness of selection. Low prior knowledge or familiarity with task type will introduce construct irrelevant difficulty. Topics which may be offensive to some students may raise anxiety and lower motivation to complete the tasks effectively. In terms of scoring, it is critical that performance ratings reflect the students' true competence and are not a result of the raters' perceptions and biases of the persons taking the test.

Summary

As Messick (1996) noted, assessments adhering to all six aspects of construct validity should adequately represent authentic reading comprehension processes and minimize irrelevant elements. As such, positive washback on instruction should be maximized. Grounding performance-based assessments and specifically reading assessments in all six aspects of construct validity is no easy task. However, as Messick (1988) pointed out, validation procedures are an eternal and ongoing process. Test developers and teachers need to attend to issues of construct validity by ensuring as far as possible the content, substantive, structural, generalizability, external, and consequential aspects of validity. These issues are as pertinent to performance-based tests as they are to standardized objective tests. However, as Linn et al. (1991, p. 20) pointed out, "the fundamental purpose of measurement [is] the improvement of instruction and learning." If one is committed to such a purpose, then positive washback in teaching and learning ought to be a primary goal of assessment.

Practical Implications for Reading Comprehension Assessment

What follows is a holistic intervention model, aimed at addressing deficiencies in any and all skills involved in the reading process. It is an instructional and not a testing model. However, the reading comprehension tasks involved in the model can be adapted for the performance-based testing of ESL/EFL reading comprehension.

While specific interventions for specific reading processes are necessary, there is a danger of fragmenting the interactive dynamism of the reading process by focusing too exclusively on individual component processes. As Grabe (1991) stated the truism, "readers learn to read by reading" (p. 396). Automaticity of the component reading processes results from constant, extensive practice (Bamford & Day, 1997). Hence the value of extensive reading which is achieved in the classroom by sustained silent reading exercises and beyond it by having the students read a variety of texts on their own (on the value of extensive reading in ESL instruction see Hosenfeld, 1995; Krashen, 1993; Nation, 1997; Pitt, White & Krashen, 1989; Robb & Susser, 1989). The instructional model proposed here can accommodate sustained silent reading and intensive language practice

in the classroom as well as extensive reading beyond the classroom. It can also accommodate interventions at the level of specific reading processes as necessary and it can also be adapted for authentic testing purposes.

Standard Reading Exercise

The model is based upon Stiefenhofer's (1996) "Standard Reading Exercise." Stiefenhofer's stated objective in constructing the exercise is that "by repeatedly working their way through the various tasks of this "standard exercise" (either in class or in guided self-study) students are expected to improve their reading speed with nonfictional [sic] English texts and to learn how to process the information in those texts more effectively" (p. 1). Stiefenhofer's exercise is presented below as he wrote it (pp. 1-2, italics in original). Comments are then offered on each of the tasks.

Standard Reading Exercise

(Please work through all tasks in the sequence given!)

Read only the title of the chosen text. What do you already know about the topic? In note form write down pieces of information you expect to find in the text.
(Mother tongue or English).

Write down at least 5 (key) words you expect to find in the text.
Is the text
a) an extract from a book?

b) a newspaper article?

c) a magazine article?

d) a scientific article?

e) a _____

4. When was it published?

5. Read through the whole text as quickly as possible. Don't worry about the words you don't understand. Now write down, in not more than 15 words, the main theme of the text.

6. Ask yourself if the text may suit the needs that made you choose it as a source of information.

7. Read through the text again trying to understand as much as you can. When you come across a word which you don't know and which you think is important for the text write it down and beside it write your idea of what it probably means. (Mother tongue or English). Use the dictionary only if absolutely necessary!

8. Divide the text into sections. Name the sections according to their function (e.g. introduction, main part(s), conclusion, etc.) and give one content-related keyword for each.

9. Write down the main idea of each paragraph using one sentence only.

10. Draw a diagram or flowchart to show how the information in the text is organized.

11. Which of the expectations/anticipations you listed in task 1 does the text meet?

12. On a separate sheet write a summary of the text. Not more than 100 words!

Make use of the results of tasks 8, 9, and 10.

13. What do you think of the text? Evaluate in the light of your reading purpose. Give

reasons for your evaluation.

Clearly, the exercise, on its face, is heavily biased in favor of higher-level skills, both for teaching and testing purposes. It does, however, possess potential for development of lower-level skills, again both for teaching and testing purposes. Before commenting on the individual tasks in the exercise, a word on text selection is in order. Reading teachers need to be aware of what Carrell (1987) called "readability." This relates to the point made by Robinson, McKenna and Wedman, in reference to beginning first language readers, that texts ought to be selected with "not too many new words" (1996, pp. 86-87). The point is made for second language readers by Koda's (1992) caution that teachers should avoid texts with too many new words in order to avoid "cognitive overload" on the part of the students. Text suitability is also connected with the notion that ESL/EFL reading ought to be taught as part of a content-centered, integrated skills curriculum (Grabe, 1991, p. 396). All of these aspects, in terms of testing, relate to issues of content and consequential validity. Instructional (and testing) content provides motivation and purposeful reading. Skills integration reinforces learning.

Each of Stiefenhofer's tasks has a purpose related to the reading process discussed above:

Task 1: The purpose is to activate the learner's prior knowledge to the topic. Doing so sets the stage for the higher-level skills of integration, summarization and elaboration. It also prompts the student to evaluate his or her choice of the specific text in terms of her specific reading goal.

Task 2: Whereas task 1 activates prior content knowledge, this task activates the reader's linguistic knowledge in the form of domain specific lexicon and may also activate specific syntactic registers.

Task 3: This will enable the reader to activate any linguistic knowledge she or he possesses of discourse structures particular to particular text types.

Task 4: Being aware of the publication date of a text is important for university level readers since it indicates the currency of the information.

Each the first four tasks are "pre-reading" tasks that, with sufficient practice, assist students to overcome the common urge to begin reading a text closely right away. They also prepare the student for actual reading by activating both linguistic and content schemata. In instruction, teacher intervention may be necessary at any of these stages to build background knowledge that students lack by teaching it (see Alvarez & Risko, 1989; Christen & Murphy, 1991). These tasks may also have the benefit of showing students that they actually possess prior knowledge that they did not know they had (Anderson, 1994).

Task 5: This practices the reading strategy of skimming and enables the reader to test his or her initial predictions made in the first four tasks.

Task 6: This encourages the student to evaluate the value and relevance of the text in relation to his or her purpose in reading it, a valuable study skill especially applicable to university level students.

Task 7: This task is reading for full comprehension and involves all four component processes of the reading process model discussed earlier in this paper.

Task 8: This practices the cognitive processes of integration and summarization involved in inferential comprehension. It is also a metacognitive strategy in that it enables the reader to check his or her comprehension.

Task 9: This task encourages the reader to make conscious connections between the information found in individual paragraphs with overall text meaning. Lower-level skills of decoding and literal comprehension are combined with higher-level skills of inferential comprehension and comprehension monitoring in order to complete this task successfully. It also begins the reading-writing connection, an important element in developing and reinforcing reading comprehension.

Task 10: This continues the process of text integration and summarization in a different format. It is also a very useful study skill for university students. It provides for transfer of the textual information to graphic format. Teachers will likely have to teach and demonstrate various diagramming techniques.

Task 11: This functions to confirm the student's appropriate selection of the text in terms of his or her reading goal.

Task 12: This continues the reading-writing connection as well as providing a further comprehension check for the reader. Several writers on ESL/EFL reading have emphasized a positive correlation between reading and writing. Grabe referred to both skills as "mutually reinforcing interactive processes" (1991, p. 394; see also Hosenfeld, 1995; Kennedy, 1994; Koda, 1992; Reid, 1992). It also has value for the university level student in that it results in a written summary of the text for future reference, perhaps for examination study. When used for a performance-based test of reading comprehension, this activity closely resembles what Bernhardt (1991) termed a "recall protocol." In this case, students are assessed not on their writing proficiency but on their comprehension of the text as revealed in the written recall, or summary of the text. Test designers need to develop a recall protocol scoring system in order to assess the students' level of comprehension (several examples are provided in Bernhardt, 1991, pp. 201-216).

Task 13: This promotes the metacognitive skill of reflecting on what has been read in terms of the reader's purpose.

The exercise, as Stiefenhofer pointed out, needs to be introduced and worked through intensively in class several times before students will be able to employ it in self-study. It will be necessary for the teacher to provide, in some cases, declarative knowledge (linguistic or topic) that the students may lack in relation to specific texts. By working through the exercise several times in class, the teacher can demonstrate and model each of the tasks (on teacher modeling using "think alouds," see Anderson, 1994; Davey, 1983;

Maria & Hathaway, 1993; on student use of "think alouds," see Davis & Bistodeau, 1993). The teacher can also, as necessary, intervene during any of the tasks to provide extra modeling as well as feedback. If lower-level skill deficits inhibit reading for full comprehension during task 7, the teacher can intervene with specific lower-level skills exercises. Thornbury's (1997) notion of "noticing," i.e., awareness of what one does not know, if acquired by learners, will create cognitive consciousness of L2 linguistic deficits and, as Collins (1994) pointed out, "knowledge precedes control" (p.3). Several writers have stressed the importance of vocabulary development and word recognition skills (Eskey & Grabe, 1988). Ooi & Lee Kim-Seoh (1996) proposed that vocabulary should be taught in context, i.e., through reading and not in isolated word lists. They also suggested that once new words have been identified through context, they should be taught through lexical sets and collocations, which increases learners' declarative knowledge of vocabulary. Carrell (1988) pointed out that learning vocabulary is "learning the conceptual knowledge associated with the word" (p. 242). On word recognition, a procedural skill, both Paran (1996) and Anderson (1994) provided examples of word recognition exercises and suggest a systematic approach whereby such exercises are practiced for a few minutes every reading class. Such repeated practice is necessary for automaticity to occur.

Summary

Stiefenhofer's exercise provides a reading comprehension instructional model that can and should inform authentic, performance-based reading instruction and assessment. The entire model need not be adapted for testing purposes. Which aspects of the model are adapted for such purposes will depend on which aspects of the reading process are being assessed. However, for broad, representative coverage of both the content and cognitive complexity of reading, the entire model should be adapted. Selection of texts will depend on the students being assessed (their prior knowledge and level of current proficiency) as well as the specific domain characteristics of the context within which they are expected to perform, e.g., academic or work. Scoring protocols need to be developed and raters trained in their use. The advantage of the model is that it treats reading as a purposeful whole with complex authentic tasks derived from what we know about what it is that individual

readers actually do when they read, thus promoting construct validity (see Wiggins, 1990; Dutcher, 1990; Elliot, 1995). It also conforms to Grabe's (1991) multidimensional description of reading (cited above) and also to Linn et al.'s stress on the multidimensional potential of performance-based assessment (cited above). Finally, it conforms to Messick's (1996) argument that valid performance assessments will strive to increase construct representation and relevance.

However, as Linn et al. (1991) and Michael (1993) pointed out, issues of cost and efficiency may well inhibit implementation of such performance-based assessment. Michael noted that "it is possible to construct objective test items that do indeed tap higher level cognitive processes and meaningful tasks" and that "traditional modes of testing afford a greater sampling of activities, a higher degree of reliability, and probably a more cost effective approach than that provided by authentic assessment" (p. 49). The model discussed above has the potential of resolving Michael's first two issues, sampling size of activities and reliability. The issue of cost and efficiency is one that individual programs need to address based on their own resources.

Conclusion

Michael (1993) noted the advantages of both standardized testing and performance-based testing, while at the same time implying the drawbacks of each. He pointed out that standardized testing tied to universally agreed objectives have advantages in large-scale nationally-based testing scenarios. However, such testing lacks the motivational impetus for learners and the positive impacts upon instruction characteristic of local district or school-based performance testing. There is a place, therefore, for both types of testing.

In a very real sense, though, all testing is local in that what is measured resides in the competencies of the individual learner. And what that learner learns will derive from the instruction provided. The link, therefore, between instruction and assessment is a critical one. The potential for positive washback on instruction and learning is perhaps the greatest strength of valid performance-based reading comprehension assessment.

References

Abu-Akel, A. (1996). The role of schemata in ESL reading comprehension. The Language Teacher Online, 20(9). [on-line]. Retrieved August 24, 2003, from http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/96/sept/schemata.html

Alderson, C. & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. Reading in a Foreign Language, 5(2), 253-270.

Alvarez, M. C. & Risko, V. J. (1989). Schema activation, construction and application. ERIC Digest, [on-line]. Retrieved 12 June, 2001, fromhttp://www.indiana.edu/~eric_rec/ieo/digests/d46.html

Amer, A. A. (1992). The effect of story grammar instruction on EFL students' comprehension of narrative text. Reading in a Foreign Language, 8(2), 711-720.

Amer, A. A. (1997). The effect of the teacher's reading aloud on the reading comprehension of EFL students . ELT Journal, 51(1), 43-47.

Anderson, N. J. (1994). Developing active readers: a pedagogical framework for the second language reading class. System, 22, 177-194.

Bamford, J. & Day, R. R. (1997). Extensive reading: What is it? Why bother? The Language Teacher Online, 21(5), [on-line]. Retrieved May 17, 2002, from http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97may/extensive.html

Bernhardt, E. B. (1991). Reading development in a second language: theoretical, empirical, & classroom perspectives. Norwood, New Jersey: Ablex Publishing Corporation.

Blakey, E. & Spence, S. (1990). Developing metacognition. ERIC Digest, [on-line]. Retrieved Nov 25, 2002, from http://www.valadosta.peachnet.edu:80/~whuitt/psy702/digests/metacogn.dig

Block, D. (1990). Seeking new bases for SLA research: looking to cognitive science. System, 18(2), 167-176.

Burger, S. E. & Burger, D. L. (1994). Determining the validity of performance-based assessment. Educational Measurement: Issues and Practice, 13(1), 9-14.

Carrell, P. L. (1987). Readability in ESL. Reading in Foreign Language, 4(1), 21-40.

Carrell, P. L. (1988). Interactive text processing: implications for ESL/second language reading classrooms. In P. Carrell, J. Devine, & D. Eskey (Eds.), Interactive approaches to second language reading (pp. 239-259). Cambridge: Cambridge University Press.

Carrell, P. L. (1989). Metacognitive awareness and second language reading. The Modern Language Journal, 73(2), 121-134..

Chikalanga, I. (1992). A suggested taxonomy for inferences for the reading teacher. Reading in a Foreign Language, 8(2), 697-709.

Chikamatsu, N. (1996). The effects of L1 orthography on L2 word recognition: A study of American and Chinese learners of Japanese. Studies in Second Language Acquisition, 18(4), 403-432.

Christen, W. L. & Murphy, T. J. (1991). Increasing comprehension by activating prior knowledge. ERIC Digest, [on-line]. Retrieved March 12, 2002, from http://www.indiana.edu/~eric_rec/ieo/digests/d61.html

Clarke, M. A, (1988). The short circuit hypothesis of ESL reading - or when language

competence interferes with reading performance. In P. L. Carrell, J. Devine, & D. Eskey (Eds.),Interactive approaches to second language reading (pp. 114-124). Cambridge: CambridgeUniversity Press.

Collins, N. D. (1994). Metacognition and reading to learn. ERIC Digest, [on-line]. Retrieved February 18, 2002, from http://www.ed.gov/databases/ERIC_Digests/ed376427.html

Cornish, F. (1992). Foreign language reading comprehension as "externally-guided thinking." Reading in Foreign Language, 8(2), 721-752.

Davis, J. & Bistodeau, L. (1993). How do L1 and L2 reading differ? Evidence from think aloud protocols. The Modern Language Journal, 77(4), 459-472.

Davey, B. (1983). Think aloud - Modeling the cognitive processes of reading comprehension. Journal of Reading, 44-47.

Dekeyser, R. M. (1997). Beyond explicit rule learning: automatizing second language morphosyntax. Studies in Second Language Acquisition, 19(2), 195-221.

Duran, R. P. (1988). Validity and language skills assessment: non-English background students. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 105-128). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dutcher, P. (1990). Authentic reading assessment. Eric Digest, No ED328607.

Elliot, S. N. (1995). Creating meaningful performance assessments. ERIC Digest, No ED381985.

Eskey, D. (1986). Theoretical foundations. In F. Dubin, D. Eskey, & W. Grabe (Eds.), Teaching second language reading for academic purposes (pp. 3-24). Reading, MA:

Addison-Wesley.

Eskey, D. (1988). Holding in the bottom: an interactive approach to the language problems of second language readers. In P. L. Carrell, J. Devine, & D. Eskey (Eds.), Interactive approaches to second language reading (pp. 93-100). Cambridge: Cambridge University Press.

Eskey, D. & Grabe, W. (1988). Interactive models for second language reading: perspectives on instruction. In P. L. Carrell, J. Devine, & D. Eskey (Eds.), Interactive approaches to second language reading (pp. 221-238). Cambridge: Cambridge University Press.

Foorrnan, B. Fletcher, J., & Francis, D. (1997). A scientific approach to reading instruction [WWW document]. http://www.ldonline.org/ld_indepth/reading/cars.html

Gagne, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). The cognitive psychology of school learning (2nd ed.). New York: HarperCollins.

Geisinger, K. F. & Carlson, J. F. (1992). Assessing language minority students. ERIC Digest, No ED356232.

Grabe, W. (1986). The transition from theory to practice in teaching reading. In F. Dubin, D. Eskey, & W. Grabe (Eds.), Teaching second language reading for academic purposes (pp. 25-48). Reading, MA: Addison-Wesley.

Grabe, W. (1991). Current developments in second language reading research. TESOL Quarterly, 25(3), 375-396.

Hammadou, J. (1991). Interrelationships among prior knowledge, inference and language proficiency in foreign language reading . The Modern Language Journal, 75(1), 27-38.

Hammerly, H. (1992). The need for directed learning in the FL classroom: a response to Collier. Studies in Second Language Acquisition, 14(2), 215-216.

Hirsch, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? Reading in a Foreign Language, 8(2), 689-696.

Horiba, Y. (1996). Comprehension processes in L2 reading, language competence, textual coherence, and inferences. Studies in Second Language Acquisition, 18(4), 433-473.

Hosenfeld, C. (1995). Reading in English as a second language: some current issues, applications, and directions for future research. College ESL, 5(2), 21-34.

Jacobs, B. & Schumann, J. (1992). Language acquisition and the neurosciences: towards a more integrative perspective. Applied Linguistics, 13(3), 282-301.

Johnson, P. (1982). Effects on reading comprehension of building background knowledge. TESOL Quarterly, 16(4), 503-528.

Kennedy, B. L. (1994). The role of topic and the reading/writing connection. TESL-EJ, 1(1), [on-line]. Retrieved July 19, 2001, from http://violet.berkeley.edu/~cwp/TESL-EJ/ej01/a3.html

Kitao, S. K. & Kitao, K. (1996). Testing communicative competence. The Internet TESL Journal, 2(5), [On-Line]. Retrieved August 17, 2001, from http://aitech.ac.jp/~iteslj

Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: implications for instruction. The Modern Language Journal, 76(4), 502-512.

Krashen, S. (1993). The power of reading: insights from the research. Englewood, CO: Libraries Unlimited.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validity criteria. Educational Researcher, 20(8), 15-21.

Livingstone, T. A. (1997). Metacognition: an overview. [On-Line]. Retrieved September 14, 2001, from http://www.gse.buffalo.edu/FAS/TShuell/CEP564/cep564/Metacog.htm

Maloy, K. (1993). Toward a new science of instruction. Washington, D. C.: Office of Educational Research and Improvement, U.S. Department of Education. [on-line]. Retrieved September 19, 2002, from http://www.ed.gov/pubs/InstScience/title.html

Maria, K. & Hathaway, K. (1993). Using think alouds with teachers to develop awareness of reading strategies. Journal of Reading, 37(1), 12-18.

Mason, D. (1992). The role of schemata and scripts in language learning. System, 20(1), 45-50.

McClain, V. (1991). Metacognition in reading comprehension: what is it and strategies for instruction. Reading Improvement, 28(3), 169-172.

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

Messick, S. (1996). Validity and washback in language testing. Language Testing, 13(3), 241- 256.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 33-46). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Michael, W. B. (1993). Performance-based (authentic) assessment. In W. B. Michael (Ed.), EDPT 655 Research Design Course Reader (pp. 46-51). University of Southern California.

Nation, P. (1997). The language learning benefits of extensive reading. The Language Teacher Online, 21(5), [on-line]. Retrieved July 16, 2002, from: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/may/language.html

Nolan, T. E. (1991). Self-questioning and prediction: combining metacognitive strategies. Journal of Reading, 35(2), 132-138.

O'Malley, J. M. & Valdez Pierce, L. (1996). Authentic Assessment for English Language Learners. New York: Addison Wesley.

Ooi, D. & Lee Kim-Seoh, J. (1996). Vocabulary Teaching: looking behind the word. ELT Journal, 50(1), 52-58.

Paran, A. (1996). Reading in EFL: facts and fictions. ELT Journal, 50(1), 25-34.

Perkins, D. N. & Grotzer, T. A. (1997). Teaching Intelligence. American Psychologist, 52(10), 1125-1133.

Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: a replication of the Clockwork Orange Study using second language acquirers. Reading in a Foreign Language, 5(2), 271-275.

Reid, J. (1992). The writing-reading connection in the ESL composition classroom. Journal of Intensive English Studies, 6, 27-50.

Robb, T. N. & Susser, B. (1989). Extensive reading vs. skills building in an ESL context. Reading in a Foreign Language, 5(3), 239-251.

Robinson, R. D., McKenna, M. C., & Wedman, J. M. (1996). Issues and Trends in Literacy Education. Boston: Allyn and Bacon.

Samuels, J. S. & Kamil, M. L. (1988). Models of the reading process. In P. L. Carrell, J. Devine, & D. Eskey (Eds.), Interactive approaches to second language reading (pp. 22-36). Cambridge: Cambridge University Press.

Schulz, R. A. (1991). Second language acquisition theories and teaching practice: how do they fit? The Modern Language Journal, 75(1), 17-26.

Schumann, J. (1990). Extending the scope of the acculturation/pidginization model to include cognition. TESOL Quarterly, 24(4), 667-684.

Sensenbaugh, R. (1996). Phonemic awareness: an important early step in learning to read. ERIC Digest, [on-line]. Retrieved April 17, 2001, from http://www.indiana.edu/~eric_rec/ieo/digests/d119.html

Spolsky, B. (1995). Measured words: the development of objective language testing. Oxford: Oxford University Press.

Stiefenhofer, H. (1996). How to read nonfictional English texts faster and more effectively: a "Standard Reading Exercise" for ESL students. The Internet TESL Journal, 2(6), [on-line]. Retrieved October 6, 2003, from http://aitech.ac.jp/~iteslj

Strodt-Lopez, B. (1996). Using stories to develop interpretive processes. ELT Journal,

50(1), 35- 42.

Sweet, A. P. (1993). State of the art: transforming for teaching and learning to read. Washington, D.C.: Office of Educational Research and Improvement, U.S. Department of Education. [on-line]. Retrieved May 12, 2002, from http://www.ed.gov/pubs/StateArt/Read/ideal.html

Tomlin, R. S. & Gernsbacher, M. A. (1994). Cognitive foundations of second language acquisition. Studies in Second Language Acquisition, 16(2), 129-133.

Thornbury, S. (1997). Reformulation and reconstruction: tasks that promote "noticing". ELT Journal, 51(4), 326-335.

Weaver, C. (1994). Reading process and practice: from socio-psycholinguistics to whole language (2nd ed.). Portsmouth, NH: Heinemann.

Wiggins, G. (1990). The case for authentic assessment. ERIC Digest, No ED328611.