From Reading Theory to Testing Practice

Carol Chapelle
Iowa State University, carolc@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/engl_pubs

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, and the Educational Assessment, Evaluation, and Research Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/engl_pubs/61. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
From Reading Theory to Testing Practice

Abstract
This chapter acts as a link between the theoretical concerns laid out in the previous section and the computer-adaptive L2 testing issues and practices discussed in this part. It defines and situates critical testing concepts used by authors in this section—test purpose, inference and construct definition—to show potential connections of theory and research in L2 reading to design and development of computer-adaptive reading tests.

Disciplines
Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Assessment, Evaluation, and Research

Comments
This is a chapter from Issues in Computer-Adaptive Testing of Reading Proficiency (1999), 150. Posted with permission.
From reading theory to testing practice

Carol Chapelle
Iowa State University

Introduction

This chapter acts as a link between the theoretical concerns laid out in the previous section and the computer-adaptive L2 testing issues and practices discussed in this part. It defines and situates critical testing concepts used by authors in this section—test purpose, inference and construct definition—to show potential connections of theory and research in L2 reading to design and development of computer-adaptive reading tests.

In the previous section, William Grabe and Elizabeth Bernhardt raised a range of theoretical and research concerns that are relevant for advancing a theory of L2 reading, developing pedagogical approaches and laying the foundation for computer-adaptive reading tests. Despite the fact that theory and research on reading are expected to inform decisions about the design of reading tests, as Grabe pointed out, 'the impact of research on reading assessment does not seem to have been very prominent' (page 35). In other words, as Figure 8.1 illustrates, the link of theory and research on L2 reading to design and development of L2 reading tests appears to be missing. In Figure 8.1, the arrows should be read as 'informs' or 'influences'.

Figure 8.1
The missing link of theory and research on L2 reading to design and development of L2 reading tests
Grabe suggests that the missing link between reading research and assessment practice is due at least in part to the psychometric criteria used to evaluate reading tests. Because reading assessment succeeds by traditional standards, he questions the role that results of reading research can play in the evolution of reading assessment. If research on reading and technology is to improve reading assessment, current methods must first be seen by some criteria to be lacking. Bernhardt also questions the usability of L2 reading research for practical assessment concerns but on different grounds. She suggests that test developers cannot ignore pressing needs for ‘useful, convenient and friendly’ tests despite unresolved theoretical issues in L2 reading.

These two impediments against linking theory and research with practice are strengthened by the fact that issues explored in L2 reading research do not appear to address directly the concerns of test construction. The theoretical papers raise questions about stages of development, strategies for lexical recognition and other areas of interest to researchers, whereas the papers in this section exemplify factors of concern to test developers such as test use, psychometric models and software for test delivery. If theory and research are to be included among those factors, the conceptual gap which exists between theory/research and design/development needs to be filled. The papers in the first section take the first step toward filling the gap by identifying what is known—and not known—about L2 reading. The four papers in this section take the next step by addressing practical issues of design and development of computer-adaptive second language tests. With issues laid out on both sides, we can begin to construct some links.

The focal topic in the papers in this section is the design and development of L2 reading tests as they are influenced by a particular test purpose and by the available resources, as shown in Figure 8.2. The test purpose plays an important role in justifying the choices that are made in test design. For example, Michael Laurier explains how users’ wishes for a single score for making placement decisions in French immersion programs influenced the design of his scoring rubric. Design choices are also influenced by available resources including computer hardware and software, time, personnel and their knowledge. In Laurier’s French project, test developers had both the time and knowledge to write their own software rather than relying on a pre-packaged template system. In the Hausa listening project that Pat Dunkel describes the availability of resources for a two-year period resulted in particular decisions being made about project goals including test design.
The factors affecting test design and development in computer-adaptive language testing projects

The influences on test design decisions described in these papers substantiate Grabe’s and Bernhardt’s observation about the minimal impact of research on assessment. Rather, the primary influences on design and development of L2 tests are factors that are ‘closer to home’ for the test developers. However, in the interest of the evolution of CAT for L2 reading, it seems critical to connect these practical concerns with theory and research in L2 reading. To do so, it is necessary to take a closer look at test purpose.

Filling the gap through ‘test purpose’

Grabe suggests that research on reading might be used to help design tests that could ‘provide more accurate information for the purposes of proficiency measurement, diagnosis and performance skills’ (page 36). To take up this suggestion, we would need to explore more thoroughly what is entailed by each of these specific test purposes. In order to do so, we need clarification on what is meant by ‘test purpose’. Test purpose can be defined through three interrelated concepts:

1) the uses made of test results;
2) the inferences made from test performance; and
3) the intended impacts of the test (Chapelle and Read 1996).

The uses made of language test results include investigation of second language acquisition (e.g. L2 reading development), evaluation of language instruction and decisions about learners in an educational context. All of the testing projects described in this section have educational uses. The Russian and French tests are intended to help with placement of learners into appropriate classes. The Hausa listening test is described as a proficiency test which might be used to assess readiness for exit from a program, for example. A critical test design decision associated with test use is the method of score reporting. Laurier’s decision about score reporting provides a good example. Despite the fact that their placement test at first yields five separate scores...
corresponding the five parts of the test, ‘since the placement decision usually consists in positioning the student on a general competence continuum’, a single score indicating overall ability is desired (page 126). The part scores were therefore combined into a single one to be used for a placement decision. The cases reported in this volume reflect the uses made of other L2 CAT projects to date (e.g. Kaya-Carton et al. 1991; Madsen 1991; Burston and Monville-Burston 1996; Brown and Iwashita 1996; Young et al. 1996) although as Grabe and Larson point out, other uses might include diagnosis of particular strengths and weaknesses in reading.

Inferences refer to the conclusions drawn about language ability or subsequent language performance on the basis of evidence from test performance. For example, an inference is made about test takers ‘reading comprehension’ on the basis of their responses to questions on a reading comprehension test. The term inference is used to indicate that the test result is not itself the object of interest to test users. Instead, test users want to know what a test taker might be expected to be capable of in non-test settings. In my view, inferences are the pivotal point at which theory and research on reading might have an impact on test design, and I will therefore expand on this concept in the next section.

Impact as a component of test purpose refers to the effects that test designers intend for the test to have on individuals (e.g. students and teachers), language classes and programs and on society (Bachman and Palmer 1996). For example, Laurier uses impact on test takers as an argument for the CAT by describing it as ‘a more pleasant experience for the student’ because the adaptive algorithm selects items at an appropriate level. Dunkel (page 91) introduces the listening test in Hausa as a means of remedying the social problem of a lack of externally-prepared standardized tests for the less-commonly taught languages. Part of McNamara’s critical appraisal of CAT is his comment on its impact on the enterprise of language testing: ‘CAT’s very efficiency may act as a conservative force in our field’. These observations illustrate the types of impacts that test designers might consider in constructing their tests (e.g. this test is intended to be a pleasant experience for students).

Figure 8.3 illustrates this elaborated definition of test purpose which one might use for analysis of a particular diagnostic test, for example, by detailing the uses to be made of test results (and therefore the form of the most useful results), the specific inferences to be made on the basis of performance and the intended impacts of test use. As illustrated in Figure 8.3, I suggest that theory and research on L2 reading might play a role—in conjunction with test use and impacts—in defining the inferences to be drawn from test performance. In order for theory and research to help shape the way test inferences are defined for a given test, it is necessary to be more precise about what this term means.
Figure 8.3
The potential role for theory and research on L2 reading in design and development of L2 reading tests

Test purpose

Theory and research on L2 reading

Inferences drawn from test performance

The term ‘inference’ is used to denote the fact that the test user does not observe directly the object of interest, but instead infers it on the basis of test performance. The object of interest is conceived in different ways depending on whether one takes an ‘ability’ or a ‘performance’ orientation to testing. Figure 8.4 illustrates the two different conceptualizations of inferences. Ability testing is based on the point of view that performance on a test is the result of some underlying capacities, which are also responsible for performance in non-test settings. The focal problem in test design is to assess accurately the ability of interest rather than other things. Performance testing, as Figure 8.4 illustrates, aims to make inferences more ‘directly’ about performance in non-test settings on the basis of test performance. The test design problem in performance testing therefore is constructing a test with characteristics as similar as possible to the non-test setting.
These perspectives on inference are important for the role of theory and research in CAT. Computer-adaptive L2 tests tend to work within the ability tradition, defining inferences as abilities such as ‘reading comprehension’ or ‘overall language ability’. Theory might improve ability testing by offering more precise definitions about what comprises ‘reading comprehension’. This is the approach Grabe takes in enumerating the ‘the components of the reading process’ (e.g., word recognition and propositional integration). McNamara’s paper offers a critical perspective on CAT in part by approaching it from a ‘performance’ perspective, describing the inferences to be drawn from test performance as ‘future performances in a non-test setting’. He is then able to question aspects of CAT design on the basis of their lack of ‘authenticity’ relative to reading in non-test settings: ‘In terms of authenticity, computer assisted reading tasks in fact best simulate the reading of computer text, a perfectly authentic skill in its own right, but not the same skill as reading non-computer text’ (page 139). McNamara gives an anecdote to support this observation, but ideally reading theory might improve testing from the performance perspective by better defining the conditions under which similar performances can be expected. For example, one might hope for a theoretical explanation and empirical research results that substantiate or refute the assertion that reading from a computer screen is different from reading from a printed page.

In short, then, the role of theory and research on reading should be to elaborate the inference to be made on the basis of test performance—in other words, to define the construct that the test is intended to measure. The nature of the inference
in turn influences test design decisions. Given the different types of inferences made through the two perspectives on testing, the nature of the construct definitions depends on whether the test developer is working within the ability or performance perspective. Because computer-adaptive reading tests work within the ‘ability’ tradition, I will focus on construct definition from that perspective.

## Construct definition

A construct is a meaningful and useful way of interpreting test performance (Messick 1981). Table 8.1 shows the constructs, or ‘meaningful interpretations’, that the authors in this section have described for the performances on their tests. In each case, the meaningful interpretation is a single trait that the authors see as useful in their particular testing contexts. For example, the test users in the French immersion programs want to use the measure of ‘general ability in French’ to place students into levels.

### Table 8.1

<table>
<thead>
<tr>
<th>Author</th>
<th>Observed performance</th>
<th>Meaningful interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larson</td>
<td>Response to reading comprehension items representing a variety of general content areas, concrete vs. abstract vocabulary and cultural content</td>
<td>general reading proficiency in Russian</td>
</tr>
<tr>
<td>Dunkel</td>
<td>Selection of a segment of text, a graphic or a part within a larger graphic to demonstrate successful comprehension of four ‘listener functions’: identification/recognition, orientation, main idea comprehension and detail comprehension</td>
<td>listening comprehension proficiency in Hausa</td>
</tr>
<tr>
<td>Laurier</td>
<td>Response to items requiring 1) comprehension of approximately 30-word paragraphs, 2) sociolinguistic judgements, 3) filling in a blank with content and function words, 4) comprehension of short aural segments, self-assessment</td>
<td>general ability in French</td>
</tr>
</tbody>
</table>

The constructs in these examples illustrate the trait-oriented constructs endemic to classic ability testing. Trait constructs, which are defined independently of the context of language use, require that test tasks sample across contexts so that the ability can be assumed to be a ‘general’ one. Larson’s explanation of the principle underlying the Russian CAT provides a good example of the influence of the trait construct on test design decisions:

This trait perspective is evident throughout the test confrontation literature. Bernhardt is concerned with the need to correct grammatical errors; because language sentence construction and the grammatical processes of the learner can affect the ‘letter or word’ identification proficiency. The crucial dimension is the construct definition of the text (Grabe’s paper). For the text recognition construct, Grabe’s paper means that the definition begins with the surface, that is, the definition begins with the definition of the surface process.

Ten years ago, a construct definition from the ‘ability perspective’ was specifically described as the ‘application of particular cognitive skills in the specific ‘reading for information’ task, which requires ‘interactional views in applied linguistics on the target language’ and evidences in general terseness’ (Grabe 1996) and 1999 (Grabe 2002). As for words, as Foster (1999) states, it is the narrower scope of meaning that is illustrated in 8.1.


Government agency tests are constructed and administered in order to test consistent and sustained performance in general proficiency. These general proficiency skills are stressed by using general language contexts and topics. To accomplish this purpose, the BYU/LTD reading tests were initially designed to allow three questions in a single content area, then the computer would be instructed to select items from other content areas. (page 86)

This trait perspective is not unique to language testing. It is also evident throughout much of the discussion in the theory papers. For example, Bernhardt identifies two types of on-line syntactic processing, assigning correct grammatical meaning to inflected words and grammatical function to sentence constituents, pointing out that if one can perform these syntactic processes efficiently during reading, ‘[w]ether this happens within a friendly letter or within a magazine report is irrelevant; when it happens—that is the crucial dimension in assigning proficiency levels.’ (page 8). Similarly, Grabe’s paper identifies a number of component reading processes (e.g. word recognition) which he defines in a context-independent manner. Given some consistency of perspective between construct definition in current CAT and definition of reading components by researchers, it seems, at least on the surface, that research might be able to contribute substantively to construct definition by suggesting the components that comprise the L2 reading process.

Ten years ago, I could have stopped there in the discussion of construct definition from the ability perspective. Recently, however, even within the ability perspective, test users are hoping to make inferences about more specifically-defined constructs—constructs that are defined relative to a particular context of language use. Such construct definitions include both a cognitive skill or capacity and a domain where the capacity is relevant, such as ‘reading for academic purposes’. This more complex construct definition, which requires definition of both the trait and the context, is called an ‘interactionalist’ construct definition (Messick 1981, 1989; Zuroff 1986). An interactionalist approach to construct definition is consistent with current views in applied linguistics that suggest language users might be good at using the target language for some purposes but not for others. This approach also evidences influences from work in both performance testing (McNamara 1996) and language for specific purposes (Douglas forthcoming). In other words, as Figure 8.5 illustrates, the interactionalist construct definition has a narrower scope than the trait-type construct definitions exemplified in Table 8.1.
When we look at the evolutionary influence that theory and research on reading might have on CAT in the 1990s, it seems necessary to think in terms of an interactionalist construct definition. The question is the following: what specific contributions can theory and research contribute toward both the 'trait' and the 'context' side of the interactionalist construct definition to improve CAT for L2 reading?

Towards a role for theory and research

A construct was defined as a 'meaningful and useful' way of interpreting test performance. Presumably, each trait-type construct definition illustrated in Table 8.1 holds an important meaning for its respective test use. However, the question is whether or not more meanings can be derived from performance on CAT reading tests if theory and research are consulted for construct definition. I will look at this possibility from the perspective of an interactionalist construct definition. In other words, I will consider the role of theory and research for both the trait and the context aspects of the construct definition.

Moving away from monolithic trait constructs

With respect to the trait aspect of the construct definition, a large gap obviously exists between the multidimensional, process-oriented components Bernhardt and Grabe describe and the monolithic labels that CAT developers use to name their constructs. Henning illustrates a need to move away from monolithic trait aspects of CAT constructs. Henning argues that the construct definition should include an overall label (e.g., reading comprehension or processing) and a set of subcomponent definitions (e.g., word reading, semantic processing, and comprehension). She raises the question of whether CAT developers should use multidimensional, process-oriented, unidimensional, or some other model. She raises the question of whether more theories can be pressed into current CAT designs or whether the models, intended to predict reading performance, should be rethought in light of our increasing knowledge of reading.
use to name the constructs their tests are intended to measure. Figure 8.6 illustrates a continuum of potential approaches to construct definitions for the trait aspect of an interactionalist construct definition. On the left-hand side is the construct definition typically used in CAT (e.g., reading comprehension or overall language proficiency). On the right-hand side is the complex processing model suggested by reading research. Bernhardt sees the definitions at the two ends of the continuum as contradictory, pointing out that her model of L2 reading is a 'multidimensional' and 'multiparameter' one. She raises the concern that has been debated repeatedly (e.g. Canale 1986; Henning 1992; Henning et al. 1985) about L2 CATs: 'At issue is ... how [the multidimensional model] fits with assessment models that assume unidimensionality of the data.' (page 3). However, this concern rests on the assumption that complex models posited by reading theory should necessarily be pressed into service as construct definitions for tests. This assumption bypasses a critical step in test design: defining the construct that the test is intended to measure.

Figure 8.6
The continuum of possibilities for trait-type construct definition

A construct definition requires a 'meaningful and useful' way of interpreting test performance, but what is meaningful and useful, of course, depends on who is doing the interpreting and what they consider useful. In the CAT projects reported in this section, the test users wanted single scores. Even the five component scores that Laurier could have interpreted as meaningful were not seen as useful by test users who wanted simple placement information. As long as test users request a single score representing static models, is there any role for theory and research on CATs?

There are at least two approaches to beginning to move from left to right along the continuum. The first, as Grabe and Larson suggest, is to consider alternative test uses (e.g. diagnosis vs. placement testing). By so doing, test developers are working for a different set of test users (i.e. teachers and learners vs. administrators) who may be prepared to find more detailed construct definitions meaningful and useful because uses might include providing evidence for what needs to be taught and learned. Second, the construct definition should serve not only the audience of the test user, but
also those who design, develop and investigate the validity of the test. If we think of a construct as needing to be useful for constructing item/task specifications for the test, there is a need for greater detail provided by theory and research on reading. If we think of a construct as useful for framing hypotheses about performance across types of items, there is a need for theory and research on reading to inform construct definition. In short, the use of information provided by complex processing models of L2 reading needs to be selectively applied to developing the trait side of a construct definition (Nichols 1994). The application involves purposefully moving from the left toward the right on the continuum. Decisions about how far to move in constructing a more detailed construct definition need to be based on its 'meaningfulness and usefulness' in view of who will interpret the meaning and what use they will make of it.

Including appropriate reading ‘contexts’

The context part of the interactionalist construct definition has a similarly large gap between current testing practice and theory in applied linguistics. In the continuum shown in Figure 8.7, the left side, ‘no substantive theory’, represents the current CAT approaches to construct definition. Influences on test performance believed to come from the testing context are treated as error, called ‘method effects’ as McNamara points out in his discussion of method effects associated with live vs. taped oral interview test (page 144). From the trait perspective, ‘method effects’ are bad because they contaminate observed test performance with influences not associated with the trait construct. From the interactionalist perspective, however, ‘method effects’ can be good if they influence performance in the desired way. For example, an interactionalist construct definition would define ‘live interactive spoken conversation with a person’ as a different construct from ‘spoken “conversation” with a machine’, and, as a consequence, would predict performance in the two settings to be different.

Figure 8.7

The continuum of possibilities for inclusion of ‘context’ in construct definition

The interactionalist construct definition requires that the features of the context be hypothesized as part of the construct definition. This is a problem that perplexes us that there is no complete set of features that would provide a complete context. For example, one would need to specify all of the time, roles and locales involved about the task to define the construct. Perrett (1996) suggests the need to name all of the features of a context as useful if they are meaningful. A context is not useful to the interactionalist if it applies only in some instances. It is useful if there is a complete or at least some concrete specifics.

Text presented on the basis of the same software

Text presented on the basis of the same software

Similarity in the basis of the same software

Similarly in the basis of the same software

Similarly in the basis of the same software

Despite the basis of the same software

A more focused approach to Bachman’s (1990) recognition of test performance depends on the characteristics of a test that can be named. Deville et al (1992) define context and authenticity as constructs that can be compared with the features which...
context be included in the construct definition because the context is hypothesized to influence performance. This requirement, however, raises a problem that is analogous to the one described above. Sociolinguists will tell us that there is a complex set of specific context features which influence performance in any given setting. For example, if one were to define the complete context of the construct measured by the spoken conversation test, one would include the speaker's goals, topics, physical setting, duration of time, roles and relationship relative to the interviewer, interest and knowledge about the topic and task, the role of the language in the event, the channel (oral/written) and the time pressure for production and comprehension (e.g., Perrett 1990). If the test designer takes a sociolinguist's approach, attempting to name all of these, the risk is a construct definition that is too complex to be meaningful and useful to anyone but a sociolinguist! In language testing, it is not useful to assess ability on a construct that is defined so narrowly that it applies only to ability to complete the test. Abilities measured by tests are useful if they are expected to influence performance beyond the test—in at least some other contexts.

In language teaching and testing, the most popular alternative to expressing ‘context’ through excessive numbers of contextual features is to evoke the folk concept of ‘authenticity’. McNamara illustrates the use of the authenticity concept in his discussion of reading tests delivered by computer:

> Text presentation and handling on computer are relatively clumsy and inauthentic, although this is improving rapidly in the area of simulation of layout and the incorporation of graphics. In terms of authenticity, computer-assisted reading tasks in fact best simulate the reading of computer text, a perfectly authentic skill in its own right, but not the same skill as reading non-computer text. (page 139)

Similarly, Bernhardt criticizes materials used in reading research on the basis of the fact that they are ‘researcher generated’ rather than ‘authentic’. Despite the ease with which the term is used, ‘authenticity’ is a relative concept (as McNamara points out), which itself is undefined and therefore does not help in construct definition.

A more revealing approach is that initiated in language testing by Bachman’s identification of ‘test method facets’ that are expected to influence test performance (Bachman 1990). The test method facets, or test task characteristics (Bachman and Palmer 1996), are a list of contextual features that can be used in a framework for test development of CAT (e.g. Chalhoub-Deville et al. 1996). They can also be used for analysis of both the testing context and the non-test setting, thereby providing a basis for defining authenticity of a test relative to a non-test setting because specific features can be compared across settings. Moreover, they serve as hypotheses about features which may influence performance and are therefore relevant to...
Construct definition. For the purposes of construct definition, however, the many features of the test task characteristics should be placed toward the right end of the continuum: they may be too numerous to satisfy the need for a meaningful and useful definition.

To find the middle ground between too much and no context for a construct definition, it is useful to examine the approaches taken by Dunkel in her CAT listening test, and by Kirsch and Mosenthal in the research described by McNamara. In each of these testing projects, the authors attempted to identify contextual factors that they thought would influence performance in a way that was relevant to the construct definition. Dunkel identified—in addition to the ‘listener functions’ that would be considered as part of the trait part of ability—two ‘context’ variables, length of input and type of option. The ‘length of input’ had two levels: word or phrase and longer monologue or dialogue. The type of option could be a text choice, a graphic choice or one part of a graphic within a larger one. These variables of the test were expected to influence performance and therefore predictions of item difficulty were made on the basis of which variables were included in each item. Unfortunately, ‘[t]he analysis of variance showed the absence of a statistical relationship between the full item-writing model (i.e. the dimensions of the conceptual framework used to create the item bank) and the [observed] difficulty of the items ... allowing us to tentatively reject the notion that the conceptual framework is related to item difficulty’ (page 105). One might also request a more explicit statement of how the selected context variables helped to define the construct measured. Nevertheless, this approach provides an example of how one might take a step to the right from the left end of the continuum.

McNamara describes a progressive testing project that steps even further to the right on the continuum. Attempting to assess ‘document literacy’ of native speakers, Kirsch and Mosenthal (1988, 1990) work with a construct defined in part by the contextual features of purpose and document text features. ‘Documents’ refer to written materials such as forms, charts and labels, which one might read in order to subscribe to a magazine, find which bus to take, or determine the appropriate dosage of medicine, respectively. ‘Task demands derive naturally from reading purpose, vary considerably from text to text and may be quite complex, reflecting authentic contexts for reading. Task difficulty is carefully defined by considering the combined characteristics of the stimulus and the task demand’ (page 142). In this research, the defined variables were significant predictors of test difficulty, which means that these factors are the ones responsible for test takers’ performance. Work remains to better understand how the purpose and text variables help to define the construct of document literacy, but this type of work makes important moves toward the right on the continuum.

Notes

1. This observation is developed in class.

2. In the present test purpose, I consider a procedure of analysis.

3. This is a sample of strategy.

4. The fact is important and make inference.

5. The introduction is perspective of theories and context.
Conclusion

Modern approaches to language testing presume that designing L2 tests requires input from L2 theory and research. However, when faced with specific, practical testing projects, we often find it difficult to articulate exactly what kind of input is needed and what effect it should have on test design decisions. The theory and research outlined by Grabe and Bernhardt do not transfer directly to issues in language testing; instead, they lay out the theoretical issues that may inform the definition of inferences made from test performance. The next four chapters illustrate how practical testing issues influence the inferences that they make on the basis of test performance. It is essential to examine both sides of the theory–practice dyad to begin to explore the middle ground, as I have attempted to do.

Notes

1. This observation about the mechanisms by which change can occur is developed thoroughly by Markee (1997) in relation to curricular change. Many of the points about the role of attitudes, perceptions and ideologies in classroom change are equally relevant to testing.

2. In the past, one may not have thought of intended ‘impacts’ as a part of test purpose, but we have included it to signify that test developers should consider impacts as part of test design. An explicit statement of impacts provides a starting point for subsequent validation work that includes analysis of consequences, or ‘washback’.

3. This is an important point which is sometimes lost in part due to the use of the term ‘direct’ for some tests. Because test users are not interested in test takers’ performance on the test alone, all tests are indirect (Bachman 1990). Performance on the test is used as an indicator of ability (which is believed to affect subsequent performance) or of subsequent performance.

4. The fact that constructs can be defined in a number of different ways is important for test design (Chapelle forthcoming). If a test’s purpose is to make inferences about a general construct such as reading comprehension, language proficiency or communicative competence, test design must, for example, specify that tasks be selected systematically from a broad domain of possible tasks. On a test of general reading comprehension, test design might specify that the input for reading tasks be drawn from a variety of sources such as newspapers, magazines, advertisements, letters and stories.

5. The interactionalist construct definition is also consistent with perspectives taken by some SLA researchers who find that cognitive theories of SLA fall short in their treatment of the interplay between contextual factors and ability (e.g. Young 1989; Tarone 1988; Ellis 1989).
6. See Snow and Lohman (1989) for a thorough discussion of the difference between the types of ‘models’ used in educational measurement and in psychology.

Acknowledgement

I am grateful to John Read and Larry Frase for many useful conversations which helped to formulate the conception of test design influenced by test purposes and available resources.

References


Daniel Eichler

**Educational Testing Service**

**Introduction**

While basic tests (CATs) are important, the first being... operationally

Each of these... involved in... Test of English Language...

**Dealing with measurement constructs**

In earlier discussion... the role of test... consideration