The Test Usefulness Feedback Framework: Developing and evaluating a procedure for describing measures of student achievement employed in web-based language learning environments

Gerry Lassche

In this paper, I explain a framework for the evaluation of testing procedures in web-based language learning programs containing four dimensions: construct validity, reliability, authenticity and impact. Qualitative, open-ended criteria are proposed and applied in an n=1 case study format. The evaluated program shows little validity with its stated objectives, tending towards informal, pre-communicative assessment that reduces the status of teachers and frames student performance in terms of discrete language functions. The framework itself, while informing the analysis, is still rather awkward to use due to the complexity of the subject.

1.0 Background context

Network technologies, information systems, the internet – all these technological affordances are impacting the teaching and learning of educational domains, including foreign languages, to a greater and greater degree. For instance, such technologies can provide more opportunities for learners to obtain otherwise inaccessible information, such as access to knowledge experts and vast quantities of multimedia documents. To that end, there have been many efforts to develop interactive multimedia programs (IMM), and advances in computer technology (multi-media and video-conferencing) are making those efforts more tangible day by day (Brett, 1998). In the realm of web-based language learning, such programs allow students worldwide, who might not otherwise have the access or wherewithal to attend a classroom, to learn a language without ever visiting the country or meeting a native-speaking resident face-to-face (Brett, 1998, 81).

Correspondingly, the push for incorporating these technological affordances into the classroom has been very strong, finding advocates among esteemed educational experts worldwide (see Appendix 1: Brief annotated bibliography of international IT integration into education). The RAND report (Glennan and Melmed, 1996), has been seminal in marshalling federal efforts to wire the classroom in the United States, and the KERI report (KERI, 1998) has served a similar function in Korea.

This integration has not been occurring without some debate concerning the degree of substantiating research (US White House, 1997; Daugherty and Funke, 1998), which asks the question "Is IT more conducive to learning than traditional classroom-based methods?" Hara and Kling (2000) suggest that the vast majority of research that is being conducted and published tends to highlight only positive aspects of the use of network technology. In Lassche (2000), I have described some of the outstanding pedagogical issues. In that paper, I proposed a simplified approach to language teaching, and from that platform criticized IT efforts in terms of cognitive and social drawbacks in the web-based learning context. Busbee (2000) agrees with this position, but approaches the subject in more historical and socio-political terms.

Hara and Kling (2000) complain that their investigation of research which could inform their web-based evaluations was made more difficult due to inherent cross-disciplinary requirements. Such a research effort requires, for example, knowledge of curricular design, information technology and systems management, TESOL, distance education, instructional assessment and evaluation, as well other areas. Chapelle (1997) weighs in with computational linguistics, instructional design and second language acquisition (SLA). With such a broad range of issues to deal with, it is not surprising that little synthesis of research has occurred, resulting in an apparent lack of a guiding research agenda, which would inform future empirical practice (Warschauer and Healey, 1998). Harper et al (2000), for example, allege the situation is due to the newness of the domain, and with time pertinent research will emerge. Chapelle (1997) suggests that the failure of SLA to produce a viable, all-encompassing theory of language learning has led to the distance between the field of computer-assisted language learning (CALL) and general foreign language pedagogy. An implication is that time is not necessarily going to change the research landscape unless there is an accompanying development in SLA theory or convergence of disciplines.

1.1 Parameters of research focus

It is critical to note that computer usage for learning a language exists along a continuum of integration, from occasional adjunct use in a classroom setting or the use of CD-ROMs in the class with the teacher as monitor (examples of CALL), to the full-blown web-based package, which allows students from all over the world to participate in cyber-space classrooms in real (synchronous) or delayed (asynchronous) time. Salaberry (1999) notes that the literature also uses the terms open systems to refer to programs which allow for flexible student use not foreseen by the developer, and closed systems to refer to programs like CD-ROMs, where the only choices available to users are within the defined parameters of the program itself. Hedberg et al (1997) offer the terms bounded versus unbounded information source as other terms. Some of the issues arising from such a distinction between partial integration and fully-online programs are discussed in Johnson (1999). Such distinctions are important because

> as Levy (1997) argues, "evaluation studies to date have shown a tendency to view CALL materials as tutorial in nature" (p. 203). In contrast, when computers are conceptualized as tools, the focus is on "how well the tool helps the user accomplish the task, not how well the computer can teach" (p. 204). In this regard, the analysis of computer mediated communication, including the analysis of learners' use of technical components that render CMC possible, deserves to be at the forefront of future research agendas. (Salaberry, 1999, 105)

The classroom-based use of computer technology I shall define as CALL, in which the word "assist" denotes a supplementary status of the technology and a primary status for the teacher in a classroom. The issues of CALL in the classroom context have already been given extensive coverage in the literature, as I will demonstrate in section 2.1 CALL research. Instead, the focus of the present inquiry will be on the comparatively lightly-researched area of unbounded, open system programs which involve language learning at a distance online, or as mediated by web-based technology (or web-based language learning). Such an approach is not irrelevant, since there are many pedagogical issues common to both fields.

1.2 Purpose of the inquiry

The websites applicable to the present study are more or less characterized by (or at least aspire to) the features described by Curtis et al (1999):

> **distance** – students who do not have access or time to attend campus classes are able to enroll;
> **flexible** – students can study at the time and pace that best suits them;
> **multi-media** – video and audio clips and CMC text from a variety of sources provide students with examples of authentic material;
> **interactive** – course material is interactive and students also have the opportunity to interact with peers, tutors and special guest tutors;
> **personalised** – students will be in constant touch with their tutor via e-mail and the feedback will therefore be prompt (subject to time-zone constraints and other

administrative gremlins);
integrative – each unit builds on the previously learned language and communication skills.

As noted, Lassche (2000) presents a summary of the current literature about web-based language learning. The paper was an effort to guide future research agenda about such learning contexts toward discussions based on pedagogical, rather than technical, issues. The present paper represents a continuing effort to address one aspect raised by that paper - the assessment of achievement outcomes in web-based language learning contexts. As noted previously, Harper et al's (2000) review of the IT literature reports that little if any work had been done with regard to assessment outcomes of web-based language learning. This finding motivates the present study, but also makes the attempt to frame such a study within relevant research more difficult.

Chapelle (1997) notes that what is needed is a perspective on web-based language learning which provides appropriate empirical research methods for investigating the critical questions about how such technology can be used to improve instructed SLA. To that end, this study seeks to use and specify language testing constructs as applied to IMM learning contexts in a evaluative device called the Test Usefulness Feedback Framework (TUFF). It is hoped that this study will provide a platform for further research into the assessment of achievement outcomes in the new web-based industry.

TUFF will be used in a case study conducted with one web-based company, whose testing protocol will be evaluated. The investigation will explore what assumptions underly testing procedures and how these are realized in practice, in an attempt to draw out possible generalizations of current industry standards of outcome testing.

1.3 Limitations of the inquiry

The results of this study should be viewed in light of the following limitations. First, the study is limited to the participation of only 1 company. While the description of this company's testing practices was comprehensive and detailed, the degree to which the characterizations can be generalized is certainly questionable due to the restricted sample size. Second, the structure of the TUFF measurement device was developed emically, in the sense that it underwent modification and adaptation as the inquiry progressed in consensus-building exchanges among pertinent stakeholders. Although such practice is within the scope of qualitative research paradigms, there is a danger that generalizations were made post hoc and less objectively than initially intended by the researcher.

2.0 Review of the Literature

2.1 CALL Research

There has been much research that demonstrates the benefits of computers used in the language classroom, (ie CALL). Noting Salaberry's (1999) recommendation for defining the context of use, the following examples would fall within the CALL category. Brett (1998), for example, cites several, where teachers in classrooms have used computers and multi-media software to enhance student learning and increase achievement gains. Warschauer and Meskill (2000) cite several case studies of integrated multi-media and internet technologies, where case evaluation has been framed in terms of teacher or student attitudes toward computer use and perceived student gains by teachers or by students themselves. Trokeloshvilli and Jost (1997) describe a program that targets the development of writing skills while paying attention to teacher goals and student needs. Fleta et al (1999) also found a positive effect for IMM technology on the acquisition of vocabulary, and improvements on listening comprehension.

Although there are still calls for further research, (Chappelle, 1997; Salaberry, 1999; Levy, 1999; Lassche, 2000), the findings seem to have reached a status of consensus: CALL is useful in a limited capacity. In fact, Gatton (1999), written by the president of the Japan branch of Dyn-Ed, a multimedia language learning program, specifically endorses a limited view of CALL that integrates computer use within classroom-based contexts. The work of Seedhouse (1995) also is of importance here, which contends that the appropriate measure of evaluation of CALL resides in the interaction between the students in classroom contexts.

> …the software is not communicative in itself, but the teacher can use the software as a tool to produce a particular type of interaction. (Seedhouse, 1995, 26)

The issue that remains is whether the student assessments that were conducted in these contexts could have benefited from greater attention to construct validity: Bearing in mind Weir's (1993) contention that testing is central to learning, would more detailed assessments have, in fact, improved the learning environment? Could tasks related to real-life communicative demands be developed? If they existed, did such tests represent real language in use for accomplishing communicative ends? These questions will be dealt with in section 2.5, which discusses language testing in more detail.

2.2 Web-based learning: general recommendations

The literature relating to web-based learning, as differentiated from CALL but not specific to language learning, suggests that web-based materials and contexts should be characterized by, among other things, an attention to the learner, the context, and the program itself. In terms of the learner, it is important for materials and activities to be interactive, encouraging communication between the learner, the instructor, and other participants (Oliver et al, 2000).

As well, the materials should be engaging, involving social activities such as online group collaboration on tasks, reflection on task applicability, and articulation of problems and experiences (Oliver et al, 2000). The creative processes that these kinds of activities facilitate are part of the constructivist approach to web-based learning (Hedberg et al, 1997). These group-centered activities are thought to create opportunities for comprehension of the materials through interaction (Chapelle, 1997). This is important to language learning as the act of negotiating meaning is thought to increase the saliency of new information, which in turn encourages its teachability and acquisition (Long and Robinson, 1996). If information is presented in such a way that students can make multiple linguistic associations with it through personally relevant experiences, the information has a greater chance of being stored and retrieved for later application (Wilson and Cole, 1996). Such authenticity increases the likelihood that students demonstrate greater motivation to learn as well (Reeves and Reeves, 1998).

Materials and tasks should be authentic, in the sense that they reflect skills and tasks demanded outside of the learning context, in real life domains (Hegelheimer and Chapelle, 2000). Decontextualized materials present information in way that reduces the appearance of cohesion, a phenomenon that can be characteristic of hypermedia, web-based environments (Oliver et al, 2000). Oliver et al (2000) suggest this sense of fragmentation can be mitigated through the demonstration of the link between learning modules by presenting summaries of, for example, learning objectives.

The more explicit the learning objectives and outcomes are, the more easily the program can be grafted into demands of the curriculum (Kennedy and McKnaught, 1997), and the more easily students can assess their own achievement or be given feedback, and the greater the ease of assessing impact (Hedberg et al, 1997). By considering the impact they have on the learning context, programs and web study guide designers show they are trying to provide enough return on investment in terms of time and financial cost (Phillips et al, 2000). Thus, great care needs to be taken that designs follow established pedagogical principles, repaying the users for the faith they put in the product (see Lassche, 2000 for a discussion of this issue).

In terms of the program itself, the web-based materials should aim for aesthetic ideals of simplicity and clarity. If a page is uncluttered, and the design is relatively straightforward, the more likely the user will (1) have confidence in the product, (2) enjoy the learning experience; and (3) demonstrate real learning gains. The learner feels more comfortable and relaxed when the design creates a sense of structured, yet open space, and finds the text more readable and therefore comprehensible (Hedberg et al, 1997).

2.3 Web-based Interactive Multi-media (IMM) research: Limitations

Where CALL is used as a supplement or stimulus in classroom contexts, web-based programs are stand-alone programs, not used in the classroom, and thus cannot avail themselves of any F2F interactions. Bain (1999) notes that most educational multi-media programs funded by IT initiatives pay little attention to either the processes or outcomes of learning, as called for earlier by Shirley and Alexander (1994). Jones and Paolucci (1999) similarly note a paucity of empirical research (less than 5% of currently published studies) that ties IT intervention and program objectives (input) to learning outcomes (output) and describes learning in terms of process. They note that much current research simply alleges "untested educational quality resulting from relatively unproven paradigms involving technology."

Several examples of allegorical research in the language learning domain are given by Zhao et al (2000), Yoon and Kwon (1998), and Curtis, Duchastel, and Radic (1999). Zhao et al (2000) describe their literacy multimedia program TELE-Web, which was developed through a research grant from the US Department of Education. Zhao and his colleagues give detailed information about the processes students are hypothesized to use when approaching various program tasks, but these learning processes are intuitively derived; as Zhao and his colleagues do not have any empirical evidence (such as can be provided by think-aloud protocols) to substantiate that learners actually engage the material in the way they suggest. The paper does not give any indication whether their program development involved any outcome validation. Another example is Yoon and Kwon (1998), who provide a description of their web-based program Voila Web. They also allege various learning processes involving material engagement without providing substantiating evidence, and significantly do not mention a need for continued program validation in terms of learning outcomes either. Curtis et al (1999) provide a detailed proposal for developing a web-based language program in terms of course description, topics, epistemology, design, delivery methods, and student support services. They go further than either of the two proposals just presented above by providing some information about course objectives as well. However, the objectives have not been rendered to a degree of clarity that would allow some of idea of assessment procedures. In fact, assessment was not directly dealt with.

The research papers that do provide information about language learning outcomes tend to frame program assessment in terms of subjective teacher or student perception or gauges of student interest and motivation (Gunn, 1999). One example of this tendency is the McMeniman and Evans (1998) survey of 4 case studies of multi-media programs designed for the learning of Asian languages. The first is an interview of teacher beliefs and self-described practices. The next two are surveys of student approval ratings for Chinese and Korean language learning programs. The fourth study, learning Japanese through email, cited proficiency gains as perceived by student self-assessments, and was limited to a sample of only 4 students. Another example is Wideman et al's (1999) analysis of the Networked English Language Learning (NELL) Project in conjunction with York University in Toronto, Canada. They conducted their study with a very small number of students, and as a result were not able to make statistical comparisons between control and experimental groups on the language achievement scores. Yet, the paper concludes that significant improvements in language development occurred. Their conclusions are based instead on the students' subjective self-assessments of language improvement, but the validity of these findings were not tested empirically through a comparison with actual test scores.

One study that has been done in an attempt to make a generalized evaluaion of the web-based industry is Smith and Salam (2000). Disguised as students, they evaluate some syllabus issues (teaching approach, range and level, material design), and some administrative issues (teacher access, help available, cost, and scheduling). Although their review covered 35 sites, many of these sites demanded credit payment up front, and thus in-depth analysis was restricted to less than 10 free sites due to concerns of fraud: would these sites take their money and then renege on the access? Their conclusion was that the practices of these web-sites did not seem to be characterized by techniques for assessing student outcomes, opportunities for encouraging student collaboration, and the use of authentic materials.

However, Smith and Salam (2000) do not describe in detail the pedagogical framework under which they made evaluations. Second, the criteria used for evaluating programs were intuitive and anecdotal, and thus do not lend themselves easily to replication and verification. Lassche's (2001) article describes an evaluative framework in which the pedagogical assumptions are made explicit, but whose outcomes are similarly subjective. Third, due to the concern noted above with regard to possible fraud, site selection was restricted to those that were offered free of charge to prospective students. It is possible that such programs, not being financially competitive, would not be characterized by use of sophisticated IMM technology, and thus would not be representative of the industry. As a result, their research is interesting for its description of some of the design and administrative shortcomings of language learning websites, but does not provide the straightforward summary of industry standards that it claims.

2.4 Web-based IMM Program research: Reasons

One reason for this one-sided treatment of the outcome issue in web-based contexts may be a lack of resources to develop pedagogically-sound programs and procedures to evaluate them in terms of student performance. The literature is replete with models to holistically evaluate multi-media programs, such as Alexander and Hedberg (1994), Taylor's (1994) NOVEX analysis, Laurillard et al's (2000) MENO framework, to name just a few. Merrill (2000) provides a very readable overview of many of the approaches used in program evaluation. One limitation of these models is that they do not adequately transfer abstract notions of construct validity, practicality, etc. into operationalized constructs for practical manipulation specific to the field of foreign language learning. That is, they are not particularly suited, especially for classroom teachers who are not assessment specialists, to (1) inform the design of language learning environments; (2) inform the design and construction of associated language tests; and third, (3) inform the validity evaluation of testing outcomes. Such a procedure necessarily involves the synthesis of language learning testing theory into multi-media program outcome analysis.

A response to the lack of empirically-derived data about achievement outcomes in literature would need to address 2 obstacles: (1) a model of language learning testing theory which could evaluate the usefulness of tests, but (2) was able to operationalise its constructs which led to clear interpretations in IMM environments. Thus, the next section evaluates some of current models of language testing theory in an effort to identify some of these salient features.

2.5 Assessment models: Competing approaches of language test validity assessment

In the language learning literature, there are several concepts that emerge as essential when evaluating language tests. The nature of the present inquiry into web-based program learning environments requires an exploration of achievement versus proficiency testing, and will be discussed first. Next, language testing depends on the use of an interactionalist paradigm for interpreting test usefulness (Chapelle, 1998), and this paradigm will be discussed. Finally, the relative merits and applicability of the six dimensions identified by Bachman and Palmer (1996, 18) as construct validity, reliability, authenticity, interactiveness, impact and practicality, will follow.

2.5.1 Achievement versus Proficiency Testing

Achievement usually refers to the attainment of specific course or syllabus/program objectives. The target of achievement test evaluation is related to the particular job or curriculum objectives it is intended to measure within a particular context of use.

Brindley (1989) describes several types of achievement tests. A level 1 test corresponds to a summative proficiency test. That is, how much improvement in the student's general language ability has resulted from the program? Prior to recent changes in the Australian Migrant English Program (AMEP), Brindley notes there was a practice-wide disuse of level 1 assessment because short-term programs rarely produce noticeable gains in general ability (Brindley, 1989, 31). Instead, AMEP teachers preferred level 3 testing. This type of testing is an informal assessment of the "enabling skills and knowledge which are part of a particular course of study." Focusing on these sub-skills renders the assessments generally unrelated to communicative performance (Brindley, 1989, 17).

Brindley recommends the use of a Level 2 test, which is concerned with the functional gains in student abilities - that is, what new skills can the student use as a result of program intervention - as a compromise between the necessities of administrative compromise and classroom-level relevancy. This satisfies the desire of the teacher to provide tailored feedback, and the administration need to provide credible standards of performance to stakeholders in the larger community.

Measuring what a student knows and what he can do as a result of teaching intervention is a legitimate endeavour that needs to be concerned with curricular relevance, such that test item characteristics correspond to the various program/curriculum objectives (Messick, 1988, 66). Bachman (1990, 71) also describes the relationship between achievement tests and proficiency tests (see Fig 1). He sees language ability as the degree of language facility in real-life contexts or the "real-life domain" - language in real use. Proficiency tests try to measure this overall language ability directly. On the other hand, programs may try to train students in particular skills or components of language ability.

Achievement tests assess the students' successful use of these skills. The degree to which both tests correspond to tasks in a real life domain produces their degree of similarity. In essence, a series of achievement tests would approximate what is covered in a proficiency test.

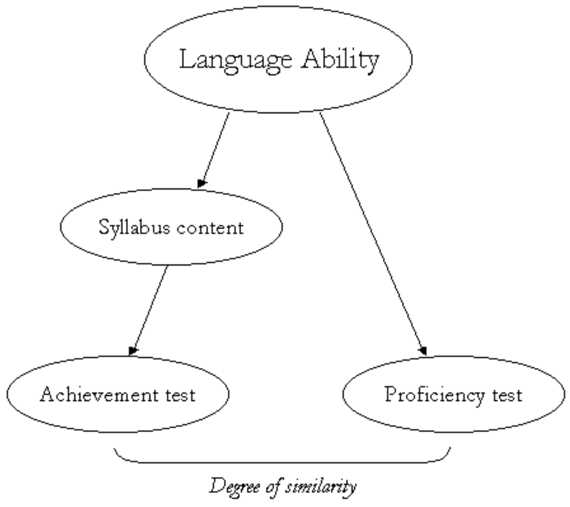


Fig. 1. *The relationship between achievement and proficiency*

2.5.2 Interactionalist paradigm for test interpretation

Chapelle (1998) notes that an interactionalist position, in contradistinction to trait or behaviorist positions, involves the recognition of relevant aspects of both trait and context influence performance and thus interpretation. Classical testing theory treated individual variation as test error (eg Bachman, 1990, 309), where Chapelle sees such variation as expected due to contextual differentiation. Generally, trait theorists are said to hold individual variations as a reflection of test error, are more interested in finding evidence for the performance of unchanging, underlying abilities (Chapelle, 1998), and view situational specificity as a statistical artifact which contaminates the interpretation of a true score. Messick (1988, 55) disagrees with the trait position also, saying that "particularities of score interpretation must be determined on the basis of local information about specific individuals in the specific setting." Chapelle (1998) in turn recommends that individual performance be viewed as "a sign of underlying traits influenced by the context in which it occurs, and is therefore a sample performance in similar contexts" (p 43).

On a related note, Messick (1988, 14) observes that validity is not concerned with the test itself, per se, but with the response set it elicits. This is because test scores are a function of items, the test-taker, and the testing context. Interestingly, this is a distinction that Chapelle fails to mention in her description of the interactionalist perspective (1998, 64).

Bachman and Palmer (1996) similarly elevate situational specificity to a prominent place among the test characteristics, notably setting, test rubric, input, expected response, relationship between input and response (p47). With respect to these facets of test items, the authors state:

> Language use involves complex and multiple interactions among the various individual characteristics of language users, on the one hand, and between these characteristics and the characteristics of the language use or testing situation, on the other. Because of the complexity of these interactions, we believe that language ability must be considered within an interactional framework of language use. (p 62)

2.5.3 Construct validity: Specification of the learning objectives

Messick's seminal work on the construct validity was an important source for Bachman's (1990) and Bachman and Palmer's (1996) works about the evaluation of test usefulness. Messick sees construct validity as "an integration of any evidence that bears on the interpretation or meaning of the test scores." Predictive evidence from test scores should correlate with the achievement construct being measured, or the extent to which test scores generalize to a specific domain of language performance (Bachman and Palmer, 1996). Hegelheimer and Chapelle (2000, 52), referring to the Bachman and Palmer framework, suggest that such interpretations depend crucially on

1. The specificity of the generalization and inferences;
2. The relevance of the assessment tasks and the scoring procedure to the inference;
3. The specificity of the test participant (ie knowledge);
4. The specificity of test tasks and scoring procedures and their degree of correspondence with test usage.

.

Messick (1988) sees construct validity as involving both an interpretation of the construct being measured, as well as the introduction of social implications (p20), or impact in Bachman and Palmer's testing paradigm. Construct validity, then, is thought to subsume content relevance, representativeness, and criterion-relatedness (p 17), and is reproduced here, adapting Bachman's (1990, 242) concept:

	Function of Outcome Testing	
Justification	Construct interpretation	Test use and interpretation
Evidence	*Construct validity*	*Construct validity+ Authenticity*
Consequence	*Construct Validity + Reliability + Interactivity*	*Construct validity+ Authentcity + Social consequences (Practicality + Impact)*

Fig 2. *Validity as a progressive matrix*

Validity is seen here as involving two aspects: First, we need to define the language construct itself. What aspect of language are we actually trying to measure? To answer this question, aspects of the test task and the test taker (interactivity), and all protocol with regard to choosing and assessing test objectives (reliability) need to be specified.

Second, the definition of validity needs to be operationalized. What remains to be specified are, first, the degree to which test items correspond to language use in real-life domains (authenticity), and second the consequences this test use has on developmental and testing constraints (practicality) and on society at large (impact).

The idea in figure 2 is that evaluation of language tests cannot proceed in isolation of the other components, such that all elements must be progressively examined in order to determine if test use is validated. The fewer elements involved in the examination, the lesser the confidence test users can place in alleged claims of validity. The following questions, from Bachman and Palmer (1996, 140ff), modified for an achievement context due to the focus of this project, need to be asked when evaluating construct validity in terms of extent to which it is satisfied and the explanation of how this was done:

1. Is the language achievement construct for this test clearly and unambiguously defined?
2. Is the language achievement construct for the test relevant to the purpose of the test?
3. To what extent does the test task reflect the construct definition?
4. To what extent do the scoring procedures reflect the construct definition?
5. Will the scores obtained from the test help us make the desired interpretations about the test takers' language achievement?

From the discussion above, it would appear that construct validity is the single, multi-faceted concept identified by Messick (1988, 1996), to which the other five dimensions contribute information for the interpretation of test usefulness. Messick (1996, 248) notes that an evaluation of construct validity involves an "integration of multiple complementary forms of convergent and discriminant evidence." The framework of Bachman and Palmer (1996, 150ff) containing the six dimensions makes this evaluation more explicit by drawing attention to aspects of the testing context which might otherwise be overlooked, as Messick (1996) suggests. Once assessments of the five other dimensions are obtained, an overall statement with regard to construct validity can be given.

Construct validity	criteria			Comments
	A	B	C	
1.Construct definition	Construct clearly specified?	Stakeholders in agreement?	Stakeholder use is consistent?	
2.Construct interpretation	Construct clearly specified?	Stakeholders in agreement?	Stakeholder use is consistent?	
3.Construct realization	Test tasks/items reflect above?	Interaction of 3A & 2 ABC correspond to 1 ABC?		
	Comments			
4.Overall assessment of reliability				
Objectives and purpose defined?				
5.Overall assessment of interactivity				
Test user competency defined				
6.Overall assessment of authenticity				
Test language construct and context defined? Interpretations predictive?				
7.Overall assessment of impact				
Consequences of test defined?				
8.Overall assessment of practicality?				
Interpretations feasible?				
9.Construct validity **Test useful?**				

Construct validity Checklist

In an effort to conceptualize what is involved with evaluating construct validity, the construct validity checklist (shown above) has two main parts. The first part is concerned with evaluating the language construct itself, in three components:

Construct definition, in terms of clarity, agreement among stakeholders, and usefulness in developing test tasks/items.

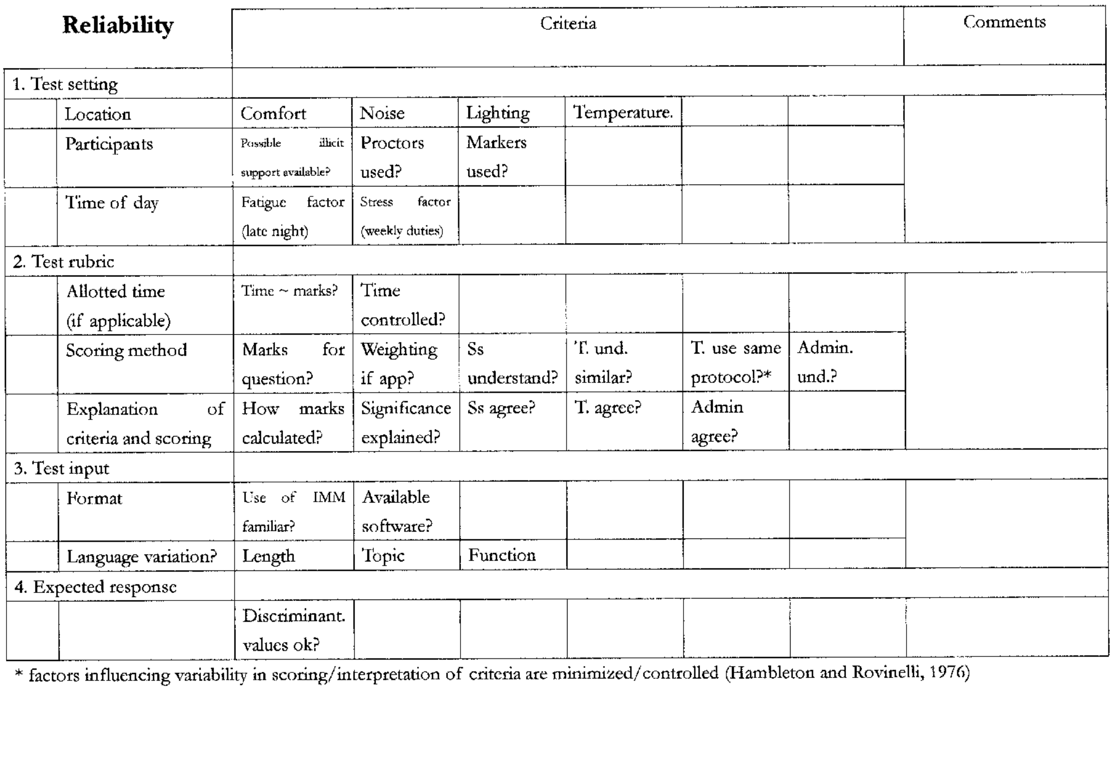
Construct interpretation: the way that the construct, once defined, is perceived tangibly by the stakeholders; dimensions of clarity, stakeholder agreement, and usefulness for scoring, feedback, and decision-making.

Construct realization: the way that characteristics of the test tasks/items reflect construct definition

The second component of construct validity evaluation involves a holistic treatment of the five dimensions, and is shown in sections 4 through 9 in the checklist

2.5.4 Reliability: Specification of test tasks, test rubric, and test setting

That the test obtains samples of consistent individual peformance while minimizing irrelevant variation is a measure of reliability (Hegelheimer and Chapelle, 2000). This is done through the use of reliable instrumentation, which is essential for claiming valid test use (Lynch, 1996, 44). Assuming that an interactionalist paradigm is essential for interpreting such variation, factors such as task characteristics (ie input such as a text to be manipulated in some way), the instructional rubric, and characteristics of the test setting need to be specified in order to ensure reliability. It is important to distinguish between the reliability and authenticity of test task input (ie language and length) characteristics.



Reliability		Criteria						Comments
1. Test setting								
	Location	Comfort	Noise	Lighting	Temperature.			
	Participants	Possible illicit support available?	Proctors used?	Markers used?				
	Time of day	Fatigue factor (late night)	Stress factor (weekly duties)					
2. Test rubric								
	Allotted time (if applicable)	Time ~ marks?	Time controlled?					
	Scoring method	Marks for question?	Weighting if app?	Ss understand?	T. und. similar?	T. use same protocol?*	Admin. und.?	
	Explanation of criteria and scoring	How marks calculated?	Significance explained?	Ss agree?	T. agree?	Admin agree?		
3. Test input								
	Format	Use of IMM familiar?	Available software?					
	Language variation?	Length	Topic	Function				
4. Expected response								
		Discriminant. values ok?						

* factors influencing variability in scoring/interpretation of criteria are minimized/controlled (Hambleton and Rovinelli, 1976)

Reliability Checklist

Reliability is concerned here with the degree to which the expected responses vary, or reactivity (Bachman and Palmer, 1996, 55ff), as a function of language, length and format, independent of authentic purpose.

Bachman and Palmer (1996, 150ff) suggest the following questions for checking the reliability of test use:

1. To what extent do characteristics of the test setting vary from one administration of the test to another?
2. To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?
3. To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, from one task to another, or on different forms of the test?
4. To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?
5. To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?

2.5.5 Authenticity: Specification of real life domain

While the constructs of validity and reliability have a long history of debate in the literature, with some measure of consensus apparent, authenticity seems to be the subject of current disagreement. In light of this, I discuss some of the issues in order to propose my criteria at greater length.

What tests *should* measure by virtue of their construct validity is of great importance, an issue Messick (1988, 66) refers to as "ultimate relevancy." Bachman and Palmer (1996, 105) refer to this situation as the degree of match between test task characteristics and an authentic, real-life task. This is a very important characteristic of construct validity. Just because a test measures what it says, questions still need to be asked about the appropriacy of what the test is measuring: "it is no longer acceptable to have language tests which are statistically valid and reliable but inauthentic in their tasks" (Hoekje and Linnell, 1994, 122). This idea is dealt with more explicitly in Bachman and

Palmer's (1996, 151 - 152) notion of authenticity, who suggest the following two questions:

1. To what extent does the description of tasks in the real-life target language domain include information about the setting, input, expected response, and relationship between input and response?
2. To what extent do the characteristics of the test task correspond to those in a real-life domain?

The problem lies in the definition of authentic. In terms of testing, how would one know whether a given task is authentic or not? In order to answer the two questions suggested above, criteria for defining what "real-life" constitutes would be necessary.

A basic consequence of the Bachman and Palmer (1996) approach to construct validity is that test developers need to specify their view of language in order for authenticity to be determinable: What kind of language does the learner engage in during the test (after Chapelle, 1997)? Bachman and Palmer (1996, 62) imply that the use of their framework necessitates a communicative approach, ie language is used in discourse, for making meaning between individuals. Test items by definition need to involve all the various language competences that individuals are endowed with, and these interact with the real-life context in which the discourse arises. Thus, test items need to be contextualized as much as possible in order for the testing of these individual competences to be interpreted in ways that satisfy the criteria of authenticity (see Fig. 2), ie "the extent to which the characteristics of the assessment reflect activities that the learner would engage in beyond the assessment setting" (Hegelheimer and Chapelle, 2000).

Gatton (1999) and Smith and Salam (2000) both suggest that web-based programs rely most heavily on texts to be read rather than listened to (audio texts) or watched (audio-visuals). The reason for this is probably due to a band-width problem, associated with the larger memory requirements required for IMM clips (audio and video) than html texts (Brett, 1998). Thus, the present discussion will focus on that medium as an example.

With regard to reading, Widdowson (1979, 165) noted long ago that authenticity lies in the process of using a text, and thus is a quality ascribed to the text by the user, rather than an intrinsic quality residing in the text itself. Widdowson's notion of a pre-communicative task is also relevant here. Such activities are specially geared to the low-level learner in order to scaffold their progress. Full-length, unmodified texts are difficult to teach to such students due to an obvious gap between current ability and the level of language present in the texts (described in Seedhouse, 1994). The way forward follows Widdowson (1979, 257ff) recommendation for the use of interlanguage texts, described as such using Selinker's (1972) term for the language system that L2 learners use as they progress toward native speaker competence in the target language. Widdowson maintains that task authenticity is not compromised as long as the manipulation of such texts reflects purposes and activities that real-world users would normally engage in.

This position is opposed by the idea that texts produced specifically for the classroom are automatically inauthentic and pedagogically contrived (eg Nunan, 1988, 1989; Clarke, 1989), where distinctions are made between material (or input) authenticity and activity authenticity (Nunan, 1989, 60). Clarke (1989, 74) asserts that Widdowson (1979) rejects the notion of simplifying texts by altering content, thereby rendering them inauthentic. Interestingly, just one paragraph later, Widdowson (1979, 165) comments: "I am not sure that it is meaningful to talk about authentic language as such at all", since authenticity is found in how the text is used. Although Clarke (1989) later acknowledges this statement (p 78), his original distinction between materials and their use channels his later arguments: use only genuine materials, and use them authentically. In contrast, Widdowson (1979, 252) argues, "[even though learners] will have to be able to cope with genuine language use at the end of the course [,] it does not follow that the only way of achieving this is to expose [them] to genuine language use at the beginning."

In any case, many authors in the current literature on literacy lean towards a critical literacy approach (Martin, 2000), where the act of reading and writing is a social semiotic process of making choices about meanings (eg after Eggins, 1994), in which the distinction between the producer of the text, the audience, and the text itself become inextricably entwined, in a sense collapsing the distinction between the text and the user. The upshot is that the text, its context, and the student as user cannot be evaluated in isolation of each other (Martin, 2000), giving support to the position adopted by Widdowson. This also recalls the holistic principle noted by Van Lier (1988, 47ff), who recommends paying close attention to the interactional, intersubjective features of context in order to determine significant factors influencing language learning and its assessment.As Lewkowicz (2001) notes, Widdowson's distinction views the interaction between the text and its audience as paramount for determining authenticity. Applied to the Bachman and Palmer (1995) model, in order for characteristics of the test task to be determined as authentic or not, the use made of the text by test developers needs to be evaluated. Alderson (2000, 250) evaluates techniques for assessing reading along these lines, suggesting the importance of matching text use to the real-life purposes of users in real-life domains: what purpose would a real user of the text have for the text? what kinds of activities would a real user do with the text? what information would a real user want to gain from the text?

A system for describing contextual aspects of language with regard to test tasks can be accounted for in the functional linguistic system proposed by Halliday (Chapelle, 1997), in terms of register (field, tenor and mode) and genre (cohesion, rhetorical staging). Tasks which foreground the significance these contextual characteristics, in terms of task response and its assessment, can be said to be more authentically representative of communicative language functions.

Widdowson (1979, 257) observes that language is characterized by interaction, meaning that it depends upon the interaction of individuals to give rise to a negotiation of meaning in the social semiotic sense realized by the Hallidayan system of functional linguistics. Thus, tasks which are characterized by such interactions can be said to be more authentic as well (Widdowson, 1979, 257). This gives rise to the notion of "interactional authenticity" (Hoekje and Linnell, 1994, 111), where the degree to which the communicative competencies of test takers are measured by test tasks is itself a measure of authentic practice.

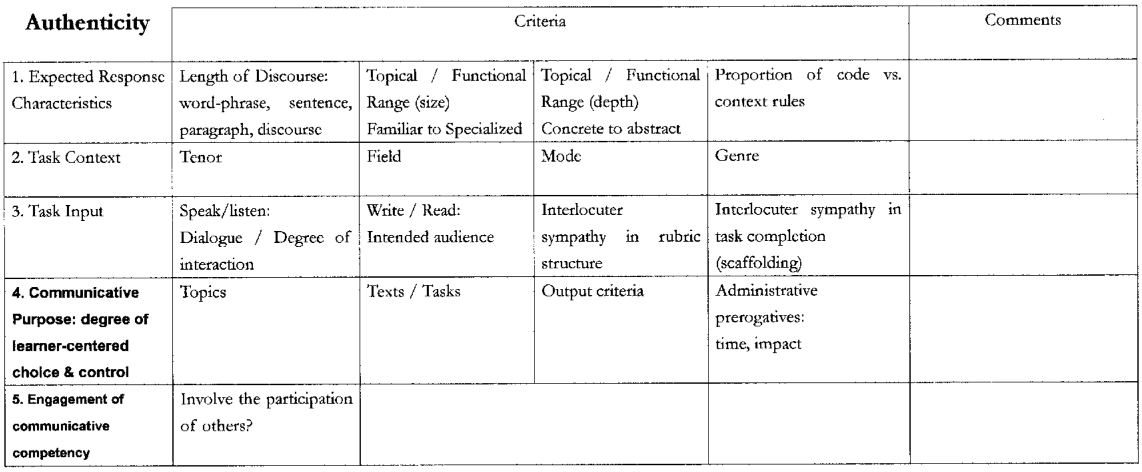
As well, Widdowson (1979, 194ff) suggests that, in communicative settings, learners resort more to the use of rules of context than to rules of code. Code rules refers to knowledge that learners have with regard to the target language, but which learners fail to apply consistently in their interlanguage performance. Context rules refers to the system learners actually use in trying to accomplish various communicative ends. Widdowson (1979, 195) suggests that errors occur in learner performance, that is, "learners do match what they know with what they do", because "normal communicative activity does not require them to do so." Thus, tasks that are characterized by a greater frequency of code rule errors are, by implication, more communicative in nature.

The constructivist approach to web-based learning (section 2.3) would recommend the use of a learner-centered techniques. Nunan (1988, 102) states that learners are engaged by an environment that relates to their own personal needs and interests, by giving students a purpose for undertaking the activity (Clarke, 1989, 83). In real-world domains, people exert exclusive influence over the communicative choices they make, the information they come into contact with, and the activities they engage in. Ostensibly, by abdicating as much of the control over syllabus content and methodology to these learners as possible, the degree of control that learners have in their real-world domains is replicated. When learners choose their own tasks, presumably they will select preferentially in ways that relate most closely with self-perceived needs and wants, and adopt a learner-as-producer orientation (Hedberg et al, 1997). In contrast, when these processes are controlled by teachers and educators (the practice of instructivist learning domains, described in Hedberg et al, 1997), such that input is limited or otherwise modified for pedagogical purposes, the learner is reduced to consumer status, and a corresponding loss of authenticity results.

This kind of approach is implied in the administration of such tests as the Interactive Performance test (Hoekje and Linnell, 1994, 121). The authors feel that the IP test is more authentic than, for example, ACTFL's Oral Proficiency Index (OPI), because the interviewee has greater control in setting the topical agenda and initiating turn-taking. In other words, the degree of control possessed by the test-taker is a proximate standard of authenticity in test task characteristics.

Thus, several criteria for authenticity are evident:

1. Expected response characteristics: whole text focus; based on Breiner-Sanders et al (2000) and Widdowson (1979);
2. Task context: described and evaluated in terms of field, tenor, mode, genre; based on Chapelle (1997);
3. Task input: referring to the degree of rubric scaffolding (flipside of point 4)
4. Communicative purpose: degree of learner control (ie test taker context)
5. Engagement of communicative competencies: test-taker / input interaction



Authenticity	Criteria				Comments
1. Expected Response Characteristics	Length of Discourse: word-phrase, sentence, paragraph, discourse	Topical / Functional Range (size) Familiar to Specialized	Topical / Functional Range (depth) Concrete to abstract	Proportion of code vs. context rules	
2. Task Context	Tenor	Field	Mode	Genre	
3. Task Input	Speak/listen: Dialogue / Degree of interaction	Write / Read: Intended audience	Interlocuter sympathy in rubric structure	Interlocuter sympathy in task completion (scaffolding)	
4. Communicative Purpose: degree of learner-centered choice & control	Topics	Texts / Tasks	Output criteria	Administrative prerogatives: time, impact	
5. Engagement of communicative competency	Involve the participation of others?				

Authenticity Checklist

2.5.6 Impact

Messick (1988, 20) urges test developers to formally consider value implications and social consequences of test use. As demonstrated earlier in fig 2, a progressive matrix conceptualization of construct validity sees test use binded to interpretation through construct validity, such that test use analysis is a consideration of the social consequences of interpretations and decisions made on the basis of test results in terms of relevance and appropriacy (Messick, 1988, 21). Bachman and Palmer (1996, 29ff) suggest the dimension "impact" as embodying this analysis, compartmentalizing the variable into micro (test-taker and teacher) and macro (social) components.

Shohamy et al (1996) describe test-taker impact in terms of their experience of taking the test and the feedback they receive on their performance. In the checklist below, I have included three possible aspects of perceptions they may develop as a result of taking the test: their perception of target language use, of the target language culture, and their own abilities with respect to the target language. These aspects are implied by Bachman and Palmer's (1996, 153) framework, in the following question:

> "To what extent might the experience of taking the test or the feedback received affect characteristics of test takers that pertain to language use (such as topical knowledge, perception of the target language use situation, areas of language knowledge, and use of strategies)?"

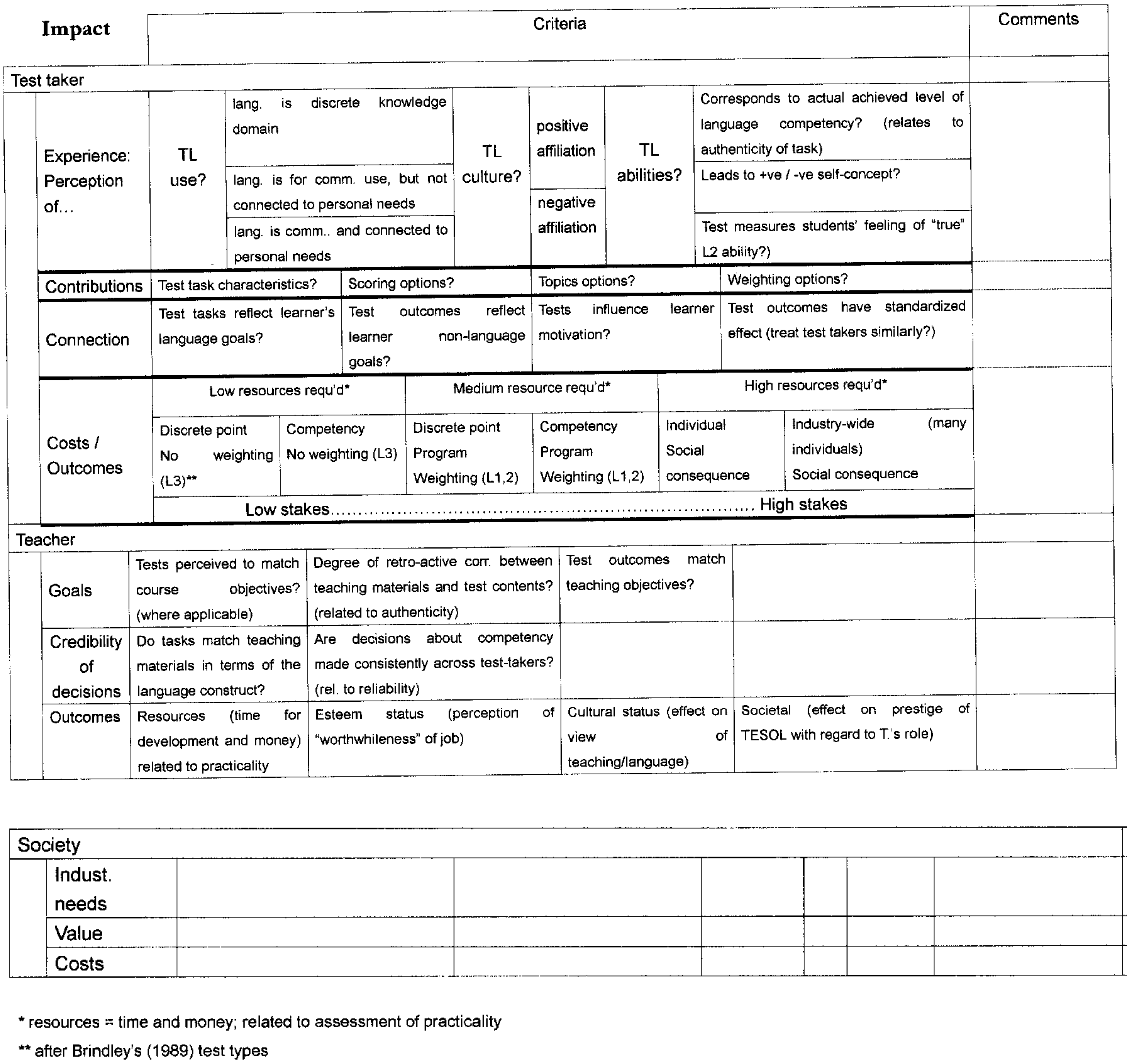
Shohamy et al (1996) also suggest that learners have the ability and desire to see a connection between their learning goals and the testing objectives, and that such an interaction is similar among their peers. This aspect is realized under the "connection" section of the checklist.

The third component of impact relating to the test-taker is an attempt to define what low stakes and high stakes tests actually are. Low stakes are presumed to have the least influence on test-taker experience, with a corresponding low degree of resources required to create such assessments. At the other end, high stakes have real, social consequences on the test-taker's quality of life, and are usually characterized by a high degree of resource investment in development.

A test-taker's affective response to any particular task/test item is probably more significant to the test developer whose aim is an informal test. Developers of standardized, formal tests cannot foresee and take into account the varying needs of every possible test taker. On the other hand, teachers and practitioners who give ad hoc, informal testing tend to do so because it affords them an opportunity to tailor feedback to specific needs (Brindley, 1989). Thus, the discussion above with regard to the applicability of the profiling of the test taker's affective characteristics is pertinent to informal, "in-classroom" diagnostic assessments and self-assessment, so-called alternative assessments (Type 3 achievement tests) rather than placement and proficiency tests (Type 1 achievement tests).

Test impact on teachers are divided into three components: goals, credibility of decisions, and outcomes, following the recommendations of Bachman and Palmer (1996, 153):

1. How consistent are the areas of language achievement to be measured with those that are included in teaching materials?
2. How consistent are the characteristics of the test and test tasks with the characteristics of teaching and learning activities?
3. How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?



Impact	Criteria						Comments
Test taker							
Experience: Perception of...	TL use?	lang. is discrete knowledge domain lang. is for comm. use, but not connected to personal needs lang. is comm.. and connected to personal needs	TL culture?	positive affiliation negative affiliation	TL abilities?	Corresponds to actual achieved level of language competency? (relates to authenticity of task) Leads to +ve / -ve self-concept? Test measures students' feeling of "true" L2 ability?)	
Contributions	Test task characteristics?		Scoring options?	Topics options?	Weighting options?		
Connection	Test tasks reflect learner's language goals?		Test outcomes reflect learner non-language goals?	Tests influence learner motivation?	Test outcomes have standardized effect (treat test takers similarly?)		
Costs / Outcomes	Low resources requ'd*		Medium resource requ'd*		High resources requ'd*		
	Discrete point No weighting (L3)**	Competency No weighting (L3)	Discrete point Program Weighting (L1,2)	Competency Program Weighting (L1,2)	Individual Social consequence	Industry-wide (many individuals) Social consequence	
	Low stakes.. High stakes						
Teacher							
Goals	Tests perceived to match course objectives? (where applicable)	Degree of retro-active corr. between teaching materials and test contents? (related to authenticity)	Test outcomes match teaching objectives?				
Credibility of decisions	Do tasks match teaching materials in terms of the language construct?	Are decisions about competency made consistently across test-takers? (rel. to reliability)					
Outcomes	Resources (time for development and money) related to practicality	Esteem status (perception of "worthwhileness" of job)	Cultural status (effect on view of teaching/language)	Societal (effect on prestige of TESOL with regard to T.'s role)			

Society						
Indust. needs						
Value						
Costs						

* resources = time and money; related to assessment of practicality

** after Brindley's (1989) test types

Impact Checklist

Criteria for the impact on society have not been developed. The web-based program under analysis in the present study was not widely used in the industry.

Certification of student work did not appear to have credibility beyond the program setting, and thus discussion of societal consequences did not appear to be relevant. The items that do appear (industrial needs, correspondence with societal values, societal costs), taken from Bachman and Palmer's (1996, 154) list of questions:

1. Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?
2. To what extent do the values and goals of the test developer coincide or conflict with the those of society and the education system?
3. What are the potential consequences, both positive and negative, for society and the education system, of using the test in this particular way?

2.5.7 Interactiveness: Specification of test user competences

The term "interactive" in many IMM discussions refers to the ability of technology to be reciprocally active in responding to the user's input (Canning-Wilson and Wallace, 2001; Oliver et al, 2000). In the Bachman and Palmer (1996) framework, however, it refers to the extent which the assessment "the assessment engages participants' knowledge, communicative language strategies, and interest in the tasks" (Hegelheimer and Chapelle, 2000, 53). The Bachman and Palmer (1996, 68ff) framework necessitates a four-fold specification of the test-taker's characteristics: (1) language ability in terms of pre-existing language competencies (the knowledge of the language) and performance (how this knowledge is applied in context); (2) affective factors, which mediate performance; (3) topical knowledge, and (4) personal characteristics.

While the Bachman and Palmer (1996) framework describes these areas of knowledge competency comprehensively in Chapter 4 of their book, operationalising and applying the information in meaningful ways to test development is beyond the resources of most web-based companies (Gatton, 1999), much less classroom teachers (see, for example, Gunn, 1995). In addition, their characterization of some aspects of language knowledge leads to some uncomfortable semantic conflicts.

For example, there is a distinction between functional (how communication is related to the goals of users), and socio-linguistic (how communication is related to setting) knowledge areas. Noting Widdowson's idea that authenticity arises from the interaction between the speaker and the text, and recalling Halliday's notion from functional linguistics that communication is a process of social semiotics (Martin, 2000), does it make sense to isolate these components? It is perhaps due to problems like these that gave rise to the observation by Halliday that such distinctions are unnecessary, as noted by Canale and Swain (1980, 3).

Byram's (1997) approach to assessment, on the other hand, involves evaluating performance solely in terms of 7 competences:

Linguistic	Produce and interpret meaning; corresponds to ideational and manipulative
Socio-linguistic	Interpret contextual cues to make linguistic choices
Discourse	Construct and interpret whole texts (interaction of organizational + manipulative)
Strategic	Manage communicative breakdown (interaction of instrumental knowledge + assessment)
Socio-cultural	Interpret contexts outside of L1 use
Social	Maintain an attitude of willingness to interact
Inter-cultural	Adaptability in foreign, unfamiliar contexts in order to maintain communication

Byram can be seen as making a case for extending the traditional communicative competences originated by Canale and Swain (1980) and elaborated and expanded by Bachman and Palmer (1996) to include new dimensions concerned with socio-cultural interactions. Byram notes that the assessment of FL language learning should be more concerned with how learners "manage communication in inter-cultural interactions" (Byram, 1997, 29). This leads to a definition of competency that involves more than just strategic competence. Learners also need skills of interpretation and establishing relationships between aspects of the two cultures, and skills of discovery and interaction (Byram, 1997, 33).

Although Byram (1997) dismisses the use of native speaker targets as assessment criteria, it is hard to conceive of a low-level learner having performances which register on these competencies. The Breiner-Sanders et al (2000) framework developed for the OPI assumes that such competences are only marginally characteristic of mid-intermediate speakers (ie, performance rises to this level of competence intermittently). How does one develop inter-language criteria for these competences? The approach taken by van Ek (as cited in Byram, 1997) and used by the Council of Europe for informing classroom assessment, describes language in terms of competences, graded along a continuum of presumed difficulty.

Alderson (2000) notes that many such efforts to grade skills and competences is fraught with difficulty to a lack of precision in specifying constructs, and a lack of research to support intuitively-based notions of what constitutes sequencing of inter-language competency and performance. Admittedly, faced with such a host of issues surrounded by conflicting evidence, I decided to omit interactivity as a dimension in the TUFF. The criteria became increasingly complex, and in effect, unwieldy for evaluation.

Although the degree and characterization of this interaction would have been facilitated by the interactivity dimension, the competences that the interactivity dimension is thought to comprise is implied in the authenticity dimension (ie "the engagement of communicative competence"). In effect, I felt that the coverage given to this interaction between test task and test taker by the authenticity dimension alone was sufficient for determining ultimate relevancy and usefulness of test tasks without reference to particular subsets of test-taker characteristics.

2.4.8 Practicality: Specification of required resources

Hegelheimer and Chapelle (2000) state that practicality is an assessment of the available resources required for developing and carrying out test development and use. Bachman and Palmer (1996) simplify the analysis by suggesting that if available resources are greater than required resources, than a given test is practical. As well, Bachman and Palmer (1996, 155) offer the following two questions for guiding such inquiry:

1. What type and relative amounts of resources are required for (a) the design stage; (b) the operationalization stage; (c) the administration stage?
2. What resources will be available for carrying out (a), (b), and (c) above?

In the present project, however, the testing procedures of the web-based program under study has already been developed, making concerns for resources etc moot. Thus, the practicality dimension will not be a component of the TUFF investigation. Instead, the information presented in this section has been included for interest sake.

3.0 Analysis design

3.1 Mode of Enquiry

The present study is, as far as the researcher is aware, one of the first attempts at characterizing holistically the testing methods of web-based language learning programs. The study and data obtained is meant to be more exploratory in nature, and offer testing devices (ie TUFF) that can inform and guide the research agenda in the field.

A naturalistic approach, being descriptive and emic in nature (Van Lier, 1988), will be used here. That is, I am interested in developing a perspective on the nature of test design and assessment, from which to generate hypotheses about, for example, the enhancement of design and testing protocol. The criteria developed for the TUFF framework are resultant of extensive consultation with and observation of stakeholders at work in their natural setting: the functioning web-based program.

3.2 Stakeholders

Language learning web-based programs from around the world will be solicited for their participation by use of a joint venture research proposal, using a text similar to the one appearing in Appendix 2. The idea is that, in exchange for free access to the site for research purposes, the researcher will provide detailed feedback on the usefulness of test procedures. The selection of programs would be restricted to those offering distance language education online, which required tuition, and involving students who will never have F2F meetings with either fellow students or their instructors. Once consent has been obtained from the appropriate authorities, online examination of the programs will ensue, and interviews will be conducted via email or telephone contacts. A script of interview questions appears in Appendix 3.

As noted section 2.2, restricting attention to those programs which are commercially-driven heightens the prospect of more sophisticated IMM technology use, and thus gives a more representative view of the industry than that offered by Smith and Urai (2000). Potential candidates could be the following:

Company name	Location
Baeoom.com	Korea
Campclick.com	Korea
Englishtown.com	USA
Netlearnlanguages.com	UK
Netlanguages.com	Spain
Edusaonline.com	USA

Some of these programs are among the most well-known programs in the industry. The two Korean companies have been cited for design awards within Korea (Korea Herald, 2000). Edusaonline.com is the largest commercial developer of online programs for educational institutions in the US, with over 1500 clients. The other companies were discovered by searches conducted on the Internet, were chosen to give a more international perspective, and responded to inquiries. Many programs that were contacted did not respond to repeated calls and emails, suggesting that the sites were no longer commercially viable, a finding echoed in the Smith and Salam (2000) study.

3.3 Data collection procedures

Gathering data. To collect data, the checklists about the four dimensions of construct validity, reliability, impact, and authenticity based on Bachman and Palmer (1996) and others in section 2.4 will be used as the investigative framework. It is hoped that such a procedure would yield greater interpretative clarity and facility of use. This framework I have referred to as the Test Usefulness Feedback Framework (TUFF).

Program instructors were questioned concerning their scoring procedures and other feedback techniques which assess student performance. These will be conducted via email and telephone informal interviews upon approval from administrative officials. The data collected here will be qualitative in nature, and coded according to the TUFF components.

Artifact evidence, including samples of teacher feedback, student questions, discussion logs, and actual syllabus and test items from website pages, were obtained over the course of the observation period. Extracts appear in the paper.

Recording data. Observation by the researcher of the specifics of each program will be recorded by means of the TUFF checklist in the form of detailed notes. Relevant extracts of web-pages will be downloaded and reproduced within the feedback given to companies as part of the venture agreement. Such extracts would not be available in the published document, as they contain specific copyrighted materials developed by the individual companies. Thus, in the published article, the names of the companies will be protected. Summative analyses will be offered in terms of the TUFF dimensions, ie qualitative summaries of site contents

3.4 Data analysis procedures

Using such criterion-based data will result in coded qualitative analyses. It is hoped that such procedures will allow underlying assumptions of how language should be taught and assessed will emerge from the analysis. Interviewing of instructors and program designers would allow for the triangulation of observer findings: do the teachers and program developers view language testing and learning in ways that correspond to current views of testing practice as embodied by the TUFF measure? These underlying epistemologies would then be presented as being representative for the web-based language learning industry, as justified in section 4.2.

In order to ensure greater validity of the findings, the researcher will use the following techniques as recommended by Lynch (1996):

1. Persistent observation, referring to the researcher's attempt to identify the most relevant aspects of testing procedures, examining in detail the entire breadth of the site;
2. Peer debriefing, referring to the researcher's intention to discuss findings with disinterested parties, such as the thesis advisor, and experts with the field of language assessment;
3. Negative case analysis, referring to the researcher's intention to observe all web-site data, and re-construct working hypothesis about the testing epistemologies as the observation unfolds;
4. Member checks, referring to the practice of triangulating ongoing observations with instructors and program designers;
5. Thick description, referring to the detailed and comprehensive description of testing procedures, in terms of time, place, and context as suggested by the modified Bachman and Palmer (1996) framework;
6. Multiple perspective negotiation, referring to an attempt on the part of the researcher to develop consensus among stakeholders concerning the validity and applicability of findings, such that the reactions and perspectives of all program design stakeholders (administrators and instructors) is included in the final analysis;
7. Utility criterion, referring to the applicability component mentioned in point 6 above, and concerned with the degree of "usefulness" the TUFF measure provides to program developers for informing their test protocol; that is, it is expected that TUFF will undergo revision and modification as the analyses proceed in an effort to provide greater applicability for program development.

4.0 Findings

4.1 General characteristics

Categories of assessment		Purpose	Cumulative. Weighting
Entrance test		Placement Level 1: Proficiency-based	0%
Within-course exercises, activities & mid-level test		Level 3: Diagnostic	0%
Tutorials			
	Email	Level 2: Progress achievement	20%
	Writing	Level 2: Progress achievement	25%
	Speaking	Level 2: Progress achievement	25%
Exit test		Level 2: Syllabus achievement	30%

Table 1: Assessment categories

As indicated in the table 1, the program has four types of assessment. As I was only given access to the first five units of the level, and the exit test is part of the last five units, I was unfortunately not able to inspect it. In addition, I also was not able to obtain criteria or archived scripts of the telephone interviews Each of the remaining assessment categories will be dealt with in turn, in terms of construct validity, reliability, authenticity, and impact. Several constructs are used in this assessment: overall construct of language proficiency (which informs program content and tests), and the construct for evaluating written exercises.

Construct validity. My informant at the site stated that his working definition of mid-intermediate proficiency was based on the Common European Framework of Reference for Languages (CUP 2001) descriptor for the B2 band (independent user):

> Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

Three aspects of the above construct definition are important for the present analysis: (1) the emphasis on whole text; (2) the range of topics from concrete to abstract; and (3) the emphasis on fluency and spontaneity. The construct definition suggested above associates fluency with the ability of native speaker interlocuters to understand the L2 learner without strain. Cucchiarini et al (2000) note that the literature tends to frame these fluency characteristics in terms of oral productivity only, summarized below:

a) the ability to talk at length with few pauses
b) the ability to talk in coherent, reasoned, and "semantically dense" sentences;
c) the ability to have appropriate things to say in a wide range of contexts;
d) the ability to be creative and imaginative in language use.
(Cucchiarini et al, 2000)

Finally, teachers use a scoring protocol for the written tutorials which typically follow five dimensions:

1. Task achievement - does the content match the specifications? (4 marks)
2. Range, complexity and accuracy of grammar (4 marks)
3. Vocabulary range and accuracy (4 marks)
4. Style and layout (3 marks)
5. Overall effect on the reader - is it coherent and communicative? (5 marks)

There is a high degree of correspondence between these characteristics and the construct definition. Both value a range of grammatical and topical L2 facility. Whole texts are characterized by an attention to coherency. Both scoring systems recognize the importance of texts being written for an audience for communicative purposes.

The website does not present this information on any of its pages. As will be shown in discussion of the findings, there was no indication that this construct guided the development of any of the tests and feedback mechanisms in the course proper. the nature of the test items, tasks, and feedback did not refer to this construct in any tangible (ie explicit statement) or theoretical way (ie the nature of test items).

The within-course exercises, the mid-level tests, and exit tests tend to deal with the texts (reading and listening) at a "face level", through extracting specific details and facts, without requiring the student to consider the socio-linguistic features of context.

Fluency. The program under study in this research does not facilitate the oral communication of students, except in the assessment of speaking during the oral interviews by telephone with the tutors, information I was unable to gain access to. In general, due to the asynchronous nature of this web-based learning program, fluency in terms of temporally unbroken speech could not be assessed. As described in the whole-text discussion above, the nature of the within course exercises do not tap creativity or imagination. The choices that users are presented with in the program are limited to the menu of exercises designed by the program developer (ie Which word is correct? Which facts are mentioned?). This drawback of web-based learning, that students are exposed to only the learning options created by the developer, was identified almost ten years ago by Mena (1993).

The validity of the language proficiency construct identified by the research participant and presumed by the entry test developers is questionable. The construct has not been specified by any of the stakeholders on the online site, nor is referred to in correspondence between the users and teachers. Although the within-course exercises rise to a higher level of correspondence with the construct, the sequencing apparent in the course materials beg the legitimacy of the proficiency rating in the first place: if students already possess the ability to process whole texts, why begin with decontextualized phrase and sentence-length items? It would appear as though the caution from Ellis (1997, 221), against the use of non-useful tasks is in order here:

> … no attempt [has been] made to ask whether the nature of the response required by the learner is a useful one (ie in the sense that it promotes language); the evaluator simply assume[d] that the response required is useful and tries to establish whether the learners' actual response matches the response intended by the task. (Ellis, 1997, 221)

To complete the evaluation of usefulness, however, requires the theoretical investigation of the five other dimensions of the model proposed by Bachman and Palmer (1996), to which I now turn.

Reliability. Evaluation of reliability involved an investigation of at least four general areas: test setting, test rubric, test input, and expected response.

The test setting occurs in an online, asynchronous mode for all 4 categories of assessment (see table 1 above), with the exception of a speaking interview. This interview is conducted via telephone, between 1 teacher and 1 student. Being a web-based format, the location (other than being in front of a computer) and time allotments for task completion are beyond the control of the teacher. Since these elements are chosen at the convenience and discretion of the learner, they can be assumed to confer optimal arrangements for eliciting best test performance. As well, the online learning experience allows for anonymity (Freeman and Capper, 1999). In a testing context, this means that the test tasks may be completed by someone other than the alleged user. In the program, the user's identity is verified through the provision of a credit card number to initiate access. Once a password is issued, the program proceeds assuming that the user and the credit card holder are one and the same. Those categories of assessment that carry higher weighting for final grading, implying higher stakes in terms of program value, in turn portend a greater threat of an imposter user.

As will be discussed later in the section relating to impact, many web-based companies offer programs that have not received recognized accreditation, and thus any testing will be low-stakes. With test setting being out of the control of the test developer, and with identity of user being difficult and expensive to verify, the evaluation of reliability with regard to test setting is relatively insignificant in the final analysis of test usefulness.

The test rubric does not involve a time factor, as noted above. If any additional time beyond the 6 months is needed to complete any task, students pay more money. With regard to scoring procedures, feedback is generally provided and tends to be formalized and dichotomous (ie true or false, right or wrong).

The third element involved with evaluating reliability is gauging test input in terms of format and language used. Describing task input in these terms is a very complicated and drawn-out process. The extent to which language can be described according to functional, illocutionary, etc characteristics seems to be arbitrary. That is, the test user's interpretation of the language used is the critical feature in determining variation. A post hoc categorization of task input only demonstrates the test developer's pre-conceptions. In which case, such descriptions are better dealt with in the area of construct validity.

Authenticity. The assessment tasks in the site seemed to suffer the greatest deficiency in the dimension of authenticity. Despite intentions to provide interaction, as shown by chat rooms and discussion lists, tasks were not utilized by students. Perhaps a willingness to communicate with strangers is presumed upon to greatly by web-based environments. For example, although Freeman and Capper (1999) note the enhancement of anonymity for communication, their baseline was prior classroom interaction between students. In other words, F2F interaction may have been a crucial precursor to online discussion, creating a sense of reality to the context, and perhaps an element of mystery (ie "I wonder which student this is…").

Another factor to the lack of authentic practice may have been due to the nature of the program itself, which dealt with discrete level knowledge domains as described above in construct validity. Being exposed to such content may have reinforced in the students' minds the lack of value attached the idea of language for meaningful use.

Interactivity. The tasks and activities in this program seem to be characterized by the observation made by Gatton (1999), who suggested that most web-based programs aim at a mid-level approach, and offer generalized content that can reach the largest possible market. Without having access to the language learning users of the program, I found it very difficult to estimate the various knowledge areas that were being tapped. As a result, developing criteria for this particular dimension proved to be quite fruitless.

What I have included instead is a generalized description of student demographics which were provided to me by the informant. Most students tended to come from Spain, where the site was located. At times, use of Spanish was used in tutor feedback to scaffold understanding. Other language users would not receive this benefit, suggesting that the preponderance of students shared a Spanish background.

While the topics in the proficiency test were generally concerned daily events and routines, the content seems to have western topics, using such names as Tony, Barbara, Jenny, Jack, etc., as opposed to Julio, Franchesca, Igor, and Wahid. Do English speakers have only Western names? As well, sports and recreation (shopping, reading), free time activities, school/work, family, clothing are topics that may not be common to war-torn, third-world countries that nonetheless speak English in daily use. Thus, students from Serbia, Vietnam, India, many African countries may not find these topics familiar or comfortable. Presumably such places would not have access to IMM equipment either, ad thus access and socio-economic status may pre-determine the usefulness of tasks with such content.

While the program may at present be content to target such a restricted market (ie Spain), it may in the future wish to expand in use with, for example, UN-sponsored literacy programs. Before that could happen, issues truly globalizing the material content would need to be resolved.

Impact. I observed that different aspects of the program have different types of impact. The four assessment components could have an impact on the test user and teacher. The program as a whole could have an impact on social status, or the way language learning is perceived. However, it seems to me that it does not really make sense to talk of the social impact of single items or tasks of an assessment setting. Even if it did, it would be difficult to gauge the differential impact of particular tasks on particular perceptions, especially when such extrapolations are speculative due to a lack of access to the language learners in the program.

The participant has developed the program and licensed its use to various agencies, from local to international (ie the local education authority in charge of secondary schools (approximately 90 000 students) to franchises in neighboring countries). In that respect, its influence on the perception of the way EFL is taught and assessed could be far-reaching.

The program has a connection with an internationally recognized distributor of EFL training materials, and according to the participant is theoretically endorsing the certification of students. However, this relationship has not been put to the test (personal communication).

The five teachers associated with the program are hired on a part-time basis, and do not meet with each other face to face. Their input is solely in terms of assessing student performance with in-course exercises. Since the program is pre-packaged, their input, like that of the students, is not used for informing design and learning objectives. The lack of significant input coupled with the lack of access to fellow colleagues reduces the status of professional status of the teacher. This type of teaching role reduces the teacher to one of technician, a consumer or end-user of materials developed by large corporate entities.

What follows is a description of the four types of assessment categories, in terms of the three categories that lent themselves to task-specific an application: construct validity; reliability, authenticity, and impact.

4.2 Findings: Entrance test

Construct validity. While the definition suggests an emphasis on whole text, none of the test components described in table (x) tap the skills necessary for constructing whole texts. The first 70 questions on the entry test purport to measure proficiencies up to and including advanced levels, yet the test items themselves are discrete sentences. The test taker needs to insert the word into the correct place (click on the word and drag it into place):

21.	have	If I'd had the money I would gone to Italy.
31.	was	The plane crashed and everyone on it killed.
41.	by	The problem was caused the computer breaking down.

There is no feedback on these exercises, and the criteria required for a given level of performance not stated or explained. Whole text features, such as mood, transitivity, theme, conjunctive relations, reference, lexical relations, cohesion, and schematic structure (see Eggins, 1994), are not dealt with. The 4 written response questions also require a two or three sentence response for each. Examples are shown below:

1. How do you spend your free time?
4. What did you do last weekend?
7. What are your plans for the next few months?
10. How important is the Internet to you?

The entry test focuses on topics of general, daily events. No specific genres are readily apparent in the examples. The written component also focus on themes of a concrete nature (schedule, family, education, job), but two imply a more abstract discussion ("ideal journey" and "importance of the internet").

Reliability. Students are simply provided a rating global proficiency rating on their performance. As noted in the construct validity section, the written section requires the students to provide 2 or 3 sentence responses to questions. No criteria for the response is specified.

Question 5:	*Have you been learning English long?*
Mid-intermediate Rating	I started learning English when I was very young. I'm a half French-Malaysian girl and when I went to Malaysia, my family always spoke to me in English, because this country was an English colony. I use to travel in London and areas there. I was in a host-family and continue practicing my English with them .
Upper-intermediate Rating	I had been in the United States for three years to study English and the computer. After I graduated from the two-year college, I came back to Japan one and half years ago.
Advanced rating	I have been learning English for eight years. I attend private lessons because this is the most popular way of learning foreign languages. I've realized that it's time to start looking for other way of learning.

Proficiency level with example student response

The inter-rate reliability associated with part one of this test is high, because the 70 categorical questions have one correct answer. Part two of the test, the written questions, would appear to have low reliability. I can find very little to distinguish between the answers given above, as all of them would appear to satisfy the minimum criteria required by the instructions.

The real question is whether the expected response actually discriminates among proficiency levels, corresponding with construct (after Rovinelli and Hambleton, 1976). My informant maintains that the test is "doing its job", since the administration re-assigns the student level only once in roughly every 20 responses on the basis of the written task (personal communication).

My suspicion is that most students are intermediate, with the novice level test-takers guessing at the correct answers. In their case, their writing in terms of sentence structure would give them away. In such a case, a true novice learner probably scores higher than true ability on the first part of the test. Advanced learners score lower than ability, since they may be operating on a "get it wrong for the right reason." These two groups are lower in frequency, hence the perception that the test is "working", when in fact it does not discriminate higher and lower levels accurately.

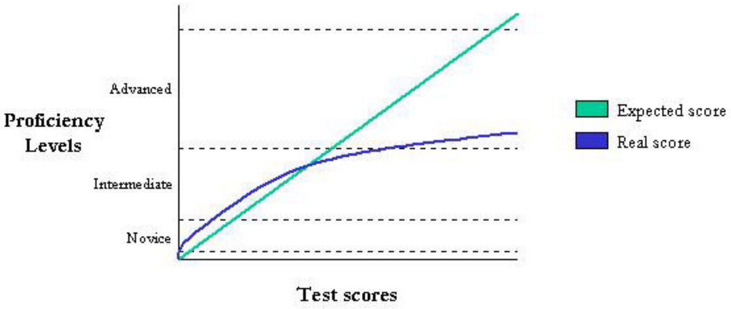


Fig 4. Relationship between proficiency level and test performance

A question that comes to mind is why a proficiency test would even be used whose contents were not related to the program. Some reasons might be to inflate certificate's credentials in the eyes of industry. The informant suggested that such tests were used to give a sense of face validity to the screening process to the students. Pairing the results with an exit-test to gauge learner progress would be legitimate practice, but not one used here. In any case, such reasons say nothing with regard to student's ability specific to course content. Using such tests as baseline performances usually have little use either, as research indicates that little progress is made with such global assessments (Brindley, 1989).

If the test developers had used tasks and items culled from the syllabus contents, the screening would be tied to specific topics, providing focused assessment on competences and skills, specific target language domains. In turn, this can show student gains, because baseline performance is specific.

Authenticity. As mentioned above, the discourse length of the entrance test is sentence length for reading. The written answers require 2 to 3 sentences as a response, and no feedback: is given in either case. The topics are familiar, daily-life routine (ie schedule, activities) in both the gap-fill and written sections. Students are working within a task context of code rules only for the gap-fill section, since the depth of analysis is descriptive, and free elaboration and reasoning is minimal. The written section is more balanced in favor of context rules.

The task context of the gap-fill questions are decontextualized, and comprehension is based on within-sentence cohesion only. There is no treatment of field, tenor, mode in these questions. The way this task treat examines proficiency is purely an academic construct. I cannot imagine a real-life situation where this kind of activity would be necessary for understanding, except in other such tests (such as the TOEFL). The written section does implicitly involve aspects of field, tenor and mode: the exercise is communicative, with the teacher as audience. Whether the scoring of the responses uses these elements for defining communciativeness

The communicative purpose of the gap-fill task is purely teacher-centered. The learners have no choice in answering these questions. There is no indication regarding the relative satisfaction students feel with the test as being reflective of their real communicative abilities. The students are given latitude with the writing section. They can choose which 4 of the ten questions they want to answer. However, scoring proceeds at the discretion of the teacher, and the topics are designed, administered, and evaluated at the sole discretion of program administration. The students receive no feedback on their performance, which could be used to dispute scoring and resultant placement.

Impact. I was not able to gauge the experience of test takers, or their perceptions of the validity of entry test tasks. Feedback is not described in the form of competencies, such that students "get an overall grade – [and] discrete skills etc [are] not specified. In the English for Work course, which is more "modular", the certificate specifies their result in each module (i.e. Speaking, Grammar, Listening etc)" (personal communication). Bearing in mind the discussion above concerning the construct validity and authenticity of the test items, the overall experience could lead to a perception that language is a discrete knowledge domain (Schmitt and Slonaker, 1996).

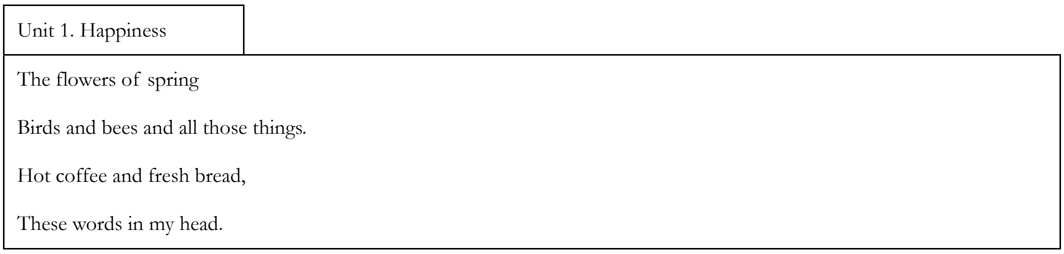
There is no evidence to suggest that negative perceptions of English-speaking culture results. Presumably, those who take the course and pay money to learn a language voluntarily have some positive associations with it. Given the pre-packaged nature of the program, users opinions and perceptions of the test tasks are not solicited, and thus there are no options available for them to affect.

From the information gathered, teachers are not concerned about the degree of correspondence between test goals and learning goals. Testing is done for "cosmetic reasons" - to give a semblance of legitimacy to the learning context. Scoring and grading is meant to be more motivational than concerned with validity or reliability.

Technical problems associated with the noticeboard system have prevented teachers from posting and sharing concerns and debate regarding scoring protocol. On the teacher's noticeboard, teachers can post examples of the scoring and other teachers are invited to scrutinize and comment. Until recently, this web-based feature had not been operating properly. The participant hopes that, once running, the noticeboard feature will facilitate greater norming of the scoring and diagnostic feedback protocol for all 4 types of testing categories evident in the program.

4.3 Findings: within-course exercises

Construct definition. The construct definition above supports the idea of whole text discourse. Within-course exercises show a greater attention to text level construction, but inconsistently. Unit 1 presents a song text (listening mode) in its entirety - a full-length text. Inspection of some of the textual features show characteristics of song genre diverge from the more typical academic genres in written or spoken discourse, such as lack of a main predicate and non-standard use of punctuation for sentence fragments. Instead of drawing attention to these features, the program has the users do a gap-fill exercise, where they are to click and drag blanked-out words into the correct location in the text.



Unit 1. Happiness

The flowers of spring
Birds and bees and all those things.
Hot coffee and fresh bread,
These words in my head.

Unit 1: Task text

The written exercises have the students write phrases of favorite things ("my father's cooking", "the taste of beer on a summer's day"), and then construct single sentences using a modal auxiliary (ie can). Students are also encouraged to use a noticeboard (similar to a discussion post on WebCT) to post their ideas and comment on other students' work. Inspection of the noticeboard shows only 5 postings in the last 4 months for the mid-intermediate level, and none of these items bore relation to course work. Students that did post are not provided feedback by tutors on this effort. Feedback obtained at random from one of the students gives credence to this assertion: this particular student said he did not finish all the exercises because there was "no one to chat with". This tendency for students to not fully utilize IMM available in websites is not uncommon to web-based learning. Smith and Salam (2000), in their review of web-based sites, and Jung (2000), in her review of web-based learning in university environments, found that little or no interaction took place between students either.

Unit 2 presents an interview dialogue, in one sentence question/answer format. The task designed to check comprehension of the text requires students to determine factually correct statements about the interview, by clicking the appropriate answers. Automatic, dichotomous feedback is generated.

Unit 2. Job interview

Interviewer: So you have worked in a Muslim culture before?

Interviewee: Yes, that's right. In Indonesia.

Interviewer: And you speak French and Indonesian, as well as English, is that right?

Interviewee: Yes. I also studied Turkish when I was at University.

Interviewer: But no Arabic?

Interviewee: No, but I'm a good language learner, and I'm sure I will be able to pick it up fast.

Unit 2 task text

Unit 3 presents a conversation (in listening mode) between two interlocuters about the use of mobile phones in different contexts. The text is clearly aiming at full discourse in spoken mode, and the activities designed to check comprehension deal with topical summary (ie click on the topics mentioned) and discerning supporting facts (ie click on the facts mentioned).

Unit 3: Phone home

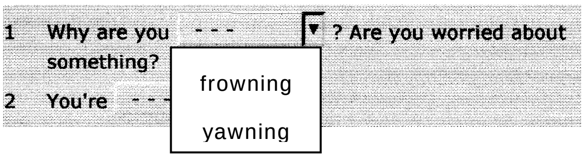
Karl: … that's a good point but what about mobile phones? I've just bought a mobile phone and it's great. Before, if I wanted to speak to a friend I had to be at home or in a call box and they had to be at home. Um, now, I can make a telephone call wherever I like, whenever I like and we can get in touch. I think that's great.

Tash: … Well, I mean, I don't know, I know that some people think mobile phones are very convenient, but I find them very antisocial. I think the noise is very intrusive and I hate sitting in a café, or in a park, or on a bus or whatever, reading my book and having somebody talking away to a piece of black plastic next to me.

Unit 3: Task text

Placed side-by-side, the various activities from units 1 to 3 appear to sequence textual treatment gradually (from phrases to sentence to paragraph), rather than assume such comprehension of whole texts exists as implied by the original proficiency construct.

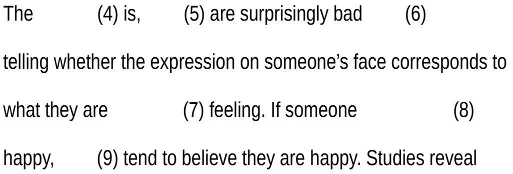
The mid-level tests serve as review exercises for the students, and do not have any overall achievement weighting. The units focus on sentence level comprehension and tests of recall. In fig.5a below, the student clicks on the arrow, and two words appear. She clicks on the one she thinks fits the sentence. After completing 20 of these, the student proceeds to the exercise in fig 5b. In this exercise, a text that the student has already viewed in the reading section of the unit is presented again. The student is expected to remember the exact wording that appeared in the text. This is a variant of the cloze exercise, where the test developer chooses particular content words he feels are essential for text comprehension (Alderson, 2000). An inspection of the text revealed that most of the blanked words in this particular text were not dependent on the comprehension of full discourse, but on adjacent and within-sentence cohesive devices.



1 Why are you --- ▼ ? Are you worried about something?

2 You're ---

Fig 5a: Exercise 1 sentence gapfill



The (4) is, (5) are surprisingly bad (6)

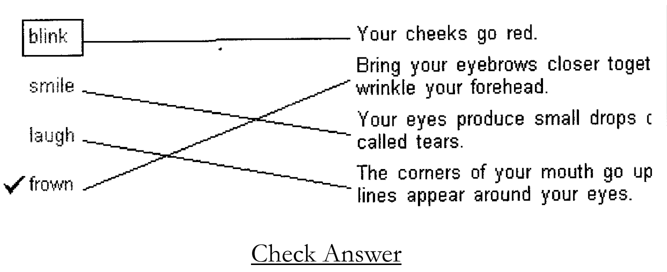
telling whether the expression on someone's face corresponds to

what they are (7) feeling. If someone (8)

happy, (9) tend to believe they are happy. Studies reveal

Fig 5b. Exercise 2: cloze gapfill

Reliability. The course- work utilizes Applet scripts at the server end. Designers use this technology to create forms which provide immediate, dichotomous feedback. An example of this type is found in the vocabulary section of unit 3, which is reproduced here in abbreviated form. Students link the word to its correct meaning. When the task is complete, the students click on the check answers button, which provides feedback shown. Reasons why certain answers are wrong are not provided. The mid-level test tasks follow the same procedure. Exercises include vocabulary matching, sentence completion, and cloze exercises, all of which presume a single correct answer. Contextually-correct answers do not receive credit. As an aside, this technology also provides some linguistic support to students through hypertext links, which opens embedded windows containing helpful information, such as word definitions, and simple grammar explanations. Oliver et al (1996) note that the kind of technology used by this program has the capacity to record diagnostic information on student performance. However, as noted table 1 above, within-course exercises and the mid-level tasks are not used in this manner.



blink	Your cheeks go red.
smile	Bring your eyebrows closer toget wrinkle your forehead.
laugh	Your eyes produce small drops c called tears.
✓ frown	The corners of your mouth go up lines appear around your eyes.

Fig 6. Instant feedback

Authenticity. Many of the exercises above require learner output consisting solely of mouse clicks, which contain no opportunities for the production of comprehensible linguistic output (Chapelle, 1997). With the focus of much of the exercises being on sentence level comprehension, it is perhaps predictable that students do not take advantage of communicative features such as chat and noticeboards. Without any feedback from tutors on this work, and not receiving any credit for participating in them, students are not motivated to use them. As such, these exercises rarely if ever rise to the level of engaging communicative competency.

In order to motivate usage of such technology, students need to see that their participation is valued in some way by the administration. Perhaps frequency of postings could be one assessment criteria, with some minimum performance threshold amount - perhaps 2 postings per unit. As well, having feedback from tutors would also support an improved perception. As noted above, teachers are not interested in discussing aspects of teaching already. I would submit that, unless tangible rewards are given to teachers for this added requirement in time and effort, they will probably be loath to participate in this way either.

Impact. I was informed that, at any given time, there are around 6 active users participating in online courses, spread over all of the proficiency levels. Around 20 so-called "dormant" users still have access to the site, but have not participated or submitted work for some time. The lure of online teaching has been advertised as learning "when you want it" (Zielinksy, 2000), and yet without a critical mass of participating students logged in at the same time, I would expect that gradually most students eventually fall prey to the "dormant" syndrome – requiring the use of assessments that alleviate the "loneliness of online learning" and keeping students interested: "The bottom line is 'Are they still there?'" (personal communication, July 24, 2001).

Programs like these demonstrate an on-going tendency for producers of language education resources to place greater stock in the intrinsic value of the materials themselves rather than in the processes with which they are used (Gatton, 1999). Thus, while the exercises in this program of themselves will not influence the degree to which this trend continues, it probably does reinforce in the minds of these particular students that distance learning, web-based or otherwise, is not a pleasant experience (for examples, see Corcoran, 1999; Zielinsky, 2000; Daugherty and Funke, 1998). And sadly, these same users may conclude that it is not the context that is to blame, but the nature of learning languages, or even worse, their own inability to cope with learning a foreign language.

Klopsis (as cited in Ellis, 1997, 222ff) evaluated the effectiveness of the tasks she used in her class. The point of interest is that conducting evaluations, both of the tasks and the interactive process of student and task, was essential for her in understanding her students. In the web-based program reviewed here, teachers are effectively removed from this diagnostic tool. Students proceed on their own, with teachers having no access to the process by students are engaging the materials. Teachers generally see the students at the end of the process. Thus, questions about why students write the way they do, or answer certain questions incorrectly, remain a mystery. The effect of having students working independently impacts the degree to which teachers can meaningfully assess the correspondence of materials with objectives, and make informative judgements about improving the program. In a sense, they are working from the disadvantage of construct under-representation: they have so few items of student performance available that they cannot make any meaningful assessments of the construct of learner achievement. Increasing the points of contact between students and teachers increases the likelihood that informed judgements will occur (Alderson, 2000).

Perhaps a hidden agenda is that teachers are not needed in this process to provide diagnostic information. Recall the fact that many programs seem to be based on the assumption that students will tend to work through the computer-based curriculum independent of any classroom interaction, a key feature of web-based learning (Warschauer and Healey, 1998). Jones (1998) notes that successful users in self-access learning environments are usually characterized by an ability to diagnose their own communicative deficits. Finally, adult students, thought to be independent and self-sufficient learners, (ie Knowles et al, 1998), and thus may be thought capable of making these judgements themselves. Whether or not students actually can do these things is irrelevant here. The critical point is that the role of the online teacher is no longer as a materials designer, a facilitator of communication, or even a diagnostician. The teacher is a technician who tabulates scores. Unfortunately, I was not able to question particular teachers or students on their individual perceptions of their roles.

4.4 Findings: Tutorials

In these exercises, students are asked to compose a text on a topic specified by the program, trying to incorporate or recycle the grammar and vocabulary they had learned in the unit. Teachers use a scoring protocol for these kind of tutorials which typically follow five dimensions:

Task achievement - does the content match the specifications? (4 marks)
Range, complexity and accuracy of grammar (4 marks)
Vocabulary range and accuracy (4 marks)
Style and layout (3 marks)
Overall effect on the reader - is it coherent and communicative? (5 marks)

An example of such a task, the student's response, and feedback is presented below. The numbers scattered through this text represent the instructor's comments:

Task

Describe an experience you had while traveling. Write at least 6 sentences, and try and use some of these expressions:

I have to be able to   I don't have to have   I have to have   I should be able to
I don't have to be able to   I don't need to be able to   I need to have
I need to be able to

Rosa's tutorial

Last summer I travelled around Peru with my friends, Alicia and Eva. We were in Nazca and we wanted to go to Arequipa (south of Peru). This journey took about eight hours so, we decided to do it at night. The people in our hotel made the reservation (1) of our bus tickets. We always do it by ourselves, but they were so kind that we entrust (2) it to them. During the day we heard some things that made us to think (3) there were (4) something wrong but we really couldn't imagine what was in store. We paid a ticket for a comfortable coach, "Royal Class", as they said.

We were supposed to leave (5) at 10.00 p.m. but the bus arrived at 11.15 p.m. The first impression wasn't very good. It was a very dark and old bus. (6) We left our backpackes in the boot and got into the bus. There was a door which separated the bus driver from the rest of the seats. I was the first person who openned the door. I was literally knocked out by a terrible smell. It could be said that all livestock of Peru (7) had been there.

Fortunately we had the firsts seats and the bus was so dark we couldn't see where we had to sat (8) down. Of course the seats weren't folding as the people told us. They were very hard and I can remember I couldn't move my head because they were impregnate (9) with an abominable smell. It was very cold but we kept an openeneded window (10). We couldn't sleep during all the night.(11)

More or less at 1.00 a.m. the bus stopped. There were (12) a police control. A shooting had been happened (13) some days ago and our bus driver had received a shot. The police asked him if he was well and if any suspect could stay in the bus.(14) As the bus driver answered no, we continued our journey. Alicia and I looked (15) each other but we didn't say any (16) word. It was five minutes later when the bus stopped again and two men left the bus. Perhaps they weren't guilty but that was what they looked like. The bus made several stops. We could hear a lot of people talking with the bus driver. My friends and I didn't see the moment to get off the bus. It was a very large night. We felt swindled, really idiots. We asked ourselves how was possible we had allowed to cheat us on such way.(17) I have to say we remembered the people of the hotel and all their relatives. In the end, (18) we arrived in Arequipa at 9.00 a.m. For the future we decided to trust only what our guide book says and to be a litle bit more distrustful.

I could write you more, but I think you can make an idea how it was, can't you? (Please, tell me if it is too long, ok?)

Tutor's feedback

(1) Correct!

(2) entrusted

(3) made us think

(4) was

(5) Correct!

(6) very dark old bus ("and" not necessary in English)

(7) I would re-word this "It was as if all the livestock of Peru..."

(8) sit

(9) impregnated (although this is a bit formal - maybe you could say "because they stank abominably")

(10) we kept a window open

(11) all night (no "during")

(12) was

(13) had happened (or had occurred) - "had been happened" is a kind of passive instruction, impossible with an intransitive verb like "happen"

(14) not sure what you mean - perhaps "if a suspect had been able to stay on the bus" or "might have stayed on the bus".

(15) at

(16) a word, (not "any")

(17) ...how was it possible we had let ourselves be cheated in such a way...

(18) Eventually... would be better than "in the end"

> This was a very dramatic account of a disatrous journey - I can't help thinking that it was based on the truth? There are a number of minor errors, and some tense and agreement mistakes that you shouldn't be making at your level! (e.g. 4, 8, 12), but the more intersting errors are syntactical and idiomatic (7, 13, 14, 17). On thw whole, the story comes across clearly despite these slips - I will give it 14 out of 20.

Construct validity. Of all the exercises in the program, these written tutorials correspond most strongly to the construct definition The students are writing a composition for an audience. The writer makes her communicative intent explicit in the text itself: "I could write you more, but I think you can make an idea how it was?" The scoring protocol also corresponds with the definition: ie fluency, creativity, etc. The extent that the student has gone beyond the specified length of the task (it calls for 8 sentences) is perhaps an indication of the student's approval of the task and her eagerness to comply with it.

Reliability. Characteristics of the task setting which may lead to variable performance is the possibility of support from other, more knowledgeable confidantes in crafting their texts. The result may therefore be a misrepresentation of their true ability. However, this is a condition common to most school settings with take-home tasks. I doubt whether the anonymity provided by the setting would increase the likelihood. In any case, it does not appear to be the case here. Routine errors in spelling and grammar are evident, as noted in the tutors feedback. If someone had edited the text, these points would probably have been picked up. One way of getting around this situation is requiring students to complete the task online – in a chat room, for example. The Applet script used by this technology also allows the teacher to record and inspect these online tasks as archived discussion logs. As discussed above, going to greater lengths to ensure the identity of the user is probably not worth doing, and do not represent a practical solution to ensuring reliability of task result.

Problems in the test rubric, however, pose a real threat to reliability. The scoring criteria for this task are not made explicit to the students at any point in the program. The nature of the feedback given here is also typical of tutorial feedback correspondence. Note the lack of any reference to these constructs in the tutor's feedback above. While a score of 14 out of 20 is given, no breakdown of the score is offered. The constructs themselves are ambiguous, the specifications for the task not being made clear in the program in the first place. Without such benchmarks, how does one know when the target has been reached? The constructs are thus open to interpretation. What constitutes appropriate style and layout? What is coherent? What is communicative? The terminological clarification and recategorization recommended by Alderson (2000) is needed here.

With such ambiguity, teachers are left to resort to their intuitions about what constitutes standard mid-intermediate performance. Indeed, the tutor providing the feedback above stated that he had relied on such intuition without regard to the scoring constructs. Instead, he appealed to a

> holistic impression, taking into account factors such as accuracy and complexity, as well as a sense of how well she was doing (i.e. a comparison with previous work) and what kind of encouragement she was likely to be needing at this point in the course, given the loneliness of online learning… and by the amount of time available. (Participant, personal communication)

In all likelihood, there is probably little agreement among the student assessors on how these constructs for scoring such exercises should be interpreted, much less applied.

Authenticity. The response characteristics include full-length discourse, topical size and depth that include surface descriptions as well as personal feelings. Although the tutor mentions that she has made "some tense and agreement mistakes that [she] shouldn't be making at her level!", this is more likely an indication of the reliance on context rules rather than code rules (Widdowson, 1979, 194), an indication that the task is tapping her communicative competencies.

The diagnostic feedback provided an analysis of semantics and syntactical errors without reference to context. Recalling the mid-intermediate band descriptor from the construct validity section, it is difficult to imagine how such feedback could encourage the development of whole text understanding. Instead, her organization of the text could have been commended, taking note the smooth progression from event to event as facilitated by the appropriate use of sequence markers (ie "last summer", "during the day", etc). I find it ironic that the tutor alleged his feedback was based on motivating the student, and yet did not find anything specific to commend in the writing itself. Instead, he notes that the story was "dramatic", and "comes across clearly", leaving the question: aren't these communicative criteria more important than the grammatical mistakes? If the mistakes do not render account any less coherent, then why de-value the text to 14 out of 20?

Thus, the feedback seems to be of relatively little value. The grammatical and semantic (ie "idiomatic usage") focus in the feedback had no content links with the unit, and tenuous theoretical links with the scoring rubric: idiomatic expression has typically been associated with higher levels of proficiency (ie Breiner-Sanders et al, 2000). The fact that the student is attempting to use it should commend her performance, rather than being categorized as an "error." The one instance where the tutor does refer to context (#9 "impregnated"), he suggests that this word use is too "formal". This feedback seems to miss the point: the student appears to be saying that the smell was as much a part of the seat as the baby is inside a woman. The visceral, emotive appeal of the wording renders the text very communicative. It would appear in this case that style and communicativeness as constructs are counter-productive.

Impact. While the task is authentic in how it was processed by the test-taker, in its implied communicative intent, the feedback and scoring techniques seem to be in juxtaposition. By focusing Rosa's attention on code rules, the tutor seems to be thwarting the communicative purpose of the task. The perception taken away by the student, then, may be that code rules are more important to follow that context rules. As Byram (1997) notes, expecting students to perform without error places an unachievable target for them to reach. If code rule errors are a natural phenomena associated with learning a foreign language, as suggested by Widdowson (1979) and many others, and not an unnatural one (ie "you shouldn't be making [them] at your level!"), then the expectation of correct use of code rules is unrealistic, and may lead to the student having a negative self-concept of language ability (Byram, 1997).

The student's point of view with regard to the scoring is not taken into account. Whether students agree with the protocol used is not solicited, and thus the interpretation of the student's true meaning in cases of communication breakdown (such as the intent of "impregnated", described above) become indiscernible to the teacher.

As noted above in the reliability section, since the constructs have not been clarified, and at times are even routinely ignored in favor of more intuitively derived assessments, it is very doubtful that students' performances are treated similarly. The case may be argued that the approach being used is learner-centered, and therefore students should not be treated the same, such that tailoring of feedback to individual needs supercedes the imperative of standardized feedback. Variable treatment of student performance is thus justifiable. However, I think that it is inconsistent to endorse learner-centered assessment in one aspect of testing, but then deny its viability for the design of the rest of the testing protocol (ie entry tests and within-course exercises), and for the further development of the program, as implied by its pre-packaged format. It this very inconsistency that may frustrate the learners into becoming "dormant users."

As noted by the tutor, the depth of assessment was also constrained by time. I find this point difficult to understand. There are five tutors, and at present 6 active students, which the informant indicates is typical enrolment scenario.

Why would time, then, be a constraining factor with regard to the depth and applicability of tutorial feedback? The program does not require development (the course is finished and materials proscribed), and the entry and within-course tests discriminate students automatically. It is possible that, on part-time contracts, teachers may be compensated for a few minutes of online feedback a week with any one student. In such a case, the role of the teacher seems to be characterized by an exaggeration of "just in time" approach increasingly necessitated by today's world (for example, see the site of the Oregon Public Education Network at www.open.k12.or.us/jitt/index.html). In the "just in time" paradigm, teachers are expected to incorporate new technology into the classroom as it becomes available, facilitating the development of the "new" literacy (ie Leu, 2000), one that is characterized by skills in using IMM, email, and other cutting-edge communication mediums, while at the same juggling the needs of students who might not be progressing at the speed envisioned by the materials developers.

4.5 Findings: Applicability of TUFF

Evaluating test usefulness proved to be a complicated, yet rewarding task. Complicated, because the process of operationalizing the qualitative construct criteria resembled Zeno's paradox:

> It is impossible to cover any distance, because half the distance must be traversed first, then half the remaining distance, then again half of what remains, and so on. Some portion of the distance to be covered is always left to cover. Therefore, motion is impossible. (Curnutt, 2000)

That is, the more I tried to specify the criteria, the more terms I uncovered that needed specification, making my progress toward a replicable model of evaluation seem out of reach. As well, I recognized how important it is to have access to student and teacher perceptions. So much of my analysis is based on the speculation and practice of one informant (who is, nevertheless, the major stakeholder in the program), that it is hard to escape from the notion of how much better and significant the work would be had I the access to teachers and students I desired. As it was, procuring participants was difficult enough. Even after accepting the proposal, the participant was constrained by time and perhaps by proprietary concerns to divulge more information.

Rewarding, because it reinforced my pre-conceptions of the value of teachers and their expertise in this process. If teacher expertise had been utilized in creating the materials, if they had been given more time and resources, the web-site would probably look completely different.

I think that authenticity has been shown to be most interesting, if not the most important, dimension in the evaluation. Construct validity and reliability are concerns which seem relatively undifferentiated between the web-based and classroom-based contexts. With an elimination of these F2F interactions, perhaps the greatest differentiating feature between web-based and classroom is in the dimension of authenticity. Further research could be done which restricted itself to this domain of inquiry, building upon the criteria I have described at length here.

5.0 Conclusion

The TUFF was very helpful for drawing broad guidelines for analysis, but still seemed to require extensive background knowledge and expertise in the field. My desire, then, to provide a straightforward, simplified technique for assessing test usefulness, remains unsatisfied. Ironically, the "just in time" approach I criticize in the previous section could characterize such a desire. Perhaps it is fitting, then, that such quick fixes are not so easy to come by, thereby securing the value of teacher practitioners for at least a while longer.

The participating web-site, echoing the findings of Gatton (1999), seems more concerned with the financial bottom-line, developing products with high face validity but little construct validity. Testing practices are concerned with assessing discrete language knowledge domains, without regard to issues of stakeholder accountability, but possessing high reliability in many aspects due to the frequency of items with dichotomized test responses. This, in turn, has severely curtailed the ability of assessment outcomes to demonstrate that acquisition in any communicative sense has taken place. Having isolated the students and teachers from each other in the learning process, web-based learning may have placed the objective of learning a language out of reach as well.

Appendix 1: Brief annotated international bibliography of IT integration into education

Location: United States

Glennan and Melmed, 1996: RAND Corporation

Developed out of RAND Critical Technologies Institute for research agenda and develop a national educational technology plan; issues discussed include use and effectiveness of IT, strategies, and government roles

US White House, 1997

Provides a "snapshot of the use and effectiveness of technology in American schools." Topics include: school access, how technology is used, the effectiveness of technology, teacher issues and cost

Education Testing Service, 1999

Provides a snapshot of the use and effectiveness of technology in American schools. Topics include: type, allocation, & patterns of use of teachnology; recommendations for improvement of use; educational impact; quality and educational standards; associated costs of implementation

Kolatch, 2000: Internet Policy Institute (US Dept of Ed.)

Extensive annotated bibliography of the state of technology in US schools

Chapman, 2000: Brookings Institute

Examines Elementary & Secondary Education Act (ESEA) and Technology in Education Program

Location: South Korea

Murchison, 1996

Describes a Canada-Korea federal-level IT joint venture program

Presidential Commission on Education Reform (PCER), 1997

Comprehensive proposal for policy of enhancing education through use of computers and network technologies in publicly funded learning institutions

Korea Education Research Institute (KERI), 1998

White Paper with describes comprehensive plan for enacting PCER recommendations

Jung, 2000: World Bank

Comprehensive examination of web-based learning in Korea, including case studies of university implementation

Location: Australia

Alexander, 1999

Evaluates current state of Australia's technology-based projects

Harper et al, 2000

Provides case studies of IT in education; describe online teaching strategies and issues of online use; discuss resource and policy implications; evaluate outcomes and need for further research

Location: United Kingdom

Claxton, 1999

Presents various anecdotal case studies of web-based learning, focusing especially on the UK's Education Departments' Superhighways Initiative (EDSI) and other cases in Europe

Appendix 2: Research venture proposal

Dear Sir:

Nowadays, there are more and more programs being offered on the internet to help people improve their foreign language ability. Many programs have at least 2 problems. First, they tend to look the same. Instead of defining their own approach, many programs offer the same kinds of materials and activities. As a result, customers do not gain from choosing one program over another. And second, many programs suffer from constant customer turnover. Most programs do not have repeat customers, and this is where most revenue is generated. Repeat customers are more likely to be satisfied with the program, and tend to the enhance company reputation by word-of-mouth.

To assist you in getting an edge over your competition and cut down on your customer turnover, I would like to interest you in a joint venture proposal:

In exchange for allowing me access to your site for about two months, I will provide you, FREE OF CHARGE, detailed feedback on your testing techniques using my innovative Test Usefulness Feedback Framework (TUFF).

TUFF will give your program a frame of reference for improving customer satisfaction in at least several ways:

(1) You will be able to design feedback tied to student performance, increasing the applicability of your program;
(2) You will be able to tie your testing procedures to specific student needs more effectively, increasing the product value;
(3) Your program will appear more transparent, increasing your product credibility;
(4) You will be able to make more specific predictions about the benefits of your program to your customers;
(5) You will gain more insight about future directions your R&D needs to take in order to gain greater market share.

I have developed foreign language programs for world-leading companies as LG Chemical and steel-maker POSCO, and I am now conducting research into the nature of web-based language learning programs. I have been studying web-based language learning for several years now, and have designed TUFF in order to investigate the industry's current testing approaches. My goal is to publish a research paper that provides a generalized picture of the web-based industry's current testing standards. In order to validate TUFF, I would like to apply it to your program.

TUFF has 4 dimensions: construct validity, reliability, authenticity, and impact. Based on current theory in testing research, and specifically informed by the innovative and respected work of Bachman and Palmer's (1996) Language Testing in Practice, my research instrument will provide specific criterion-based feedback on your testing procedures.

For example, with respect to the instructional rubric used for test items, TUFF will be able to provide explicit item-by-item detail on range of clarity and relevance of the accompanying instructions. Such information will help you design test items that prevent scoring and testing bias, increasing the accuracy and predictive value of your test. And this is but one example of how TUFF can assist and inform your program development.

Over a period of approximately 6 to 8 weeks, I will examine your program's teaching and assessment content, and will hold email interviews with instructors about their scoring and feedback methods. After this period of investigation, I will provide you with a detailed report of my findings.

My research will the hold the identity and copyrighted specifics of your program in the strictest confidence. It will greatly help to guide the research and development agenda for your program's future!

I look forward to working with you!

Your sincerely,

Mr. Gerry Lassche
University of Wollongong
02 4283 3689

Appendix 3: Interview Questions

Entry-level test section

1. In general, how would you describe (holistically) mid-intermediate communicative proficiency?

2. What aspects of language ability are being tested in this test (ie the 70 gap-fill questions)? That is, having administered the test, what are you now able to say about specific strengths and weaknesses in the test takers?

3. How were these questions developed? That is, was there any pre-testing of the items, to see how different items discriminate among different levels of ability?

4. How are the questions structured, in terms of relative difficulty (ie grammar, vocabulary, topic, etc), and matched for level (ie what question corresponds to what intended proficiency level)?

5. What criteria are used for evaluating the 4 written responses? Could you provide me an example of such feedback?

6.Would all the stakeholders involved with the program agree with your evaluations made in the above questions?

Unit 1 Happiness

7. What are the language goals for this unit?

8. In what terms are these goals evaluated?

9. Are students aware of these goals?

10. Are students aware of the scoring procedures, ie do they know why certain answers are correct or not?

11. Given your answer in #1, how well do the exercises in parts 2, 3, and 4 correspond to

and/or scaffold language learning? This is probably a difficult question to ask and answer without having the benefit of your response to #1, so I will come back to this question later.

12. Would all the stakeholders involved with the program agree with your evaluations made in the above questions?

Tutorial feedback

The feedback in the tutorials appears to be much more informal, and tailored to specific student needs.

14. What criteria did you use to develop your tutorial scores?

15. To what extent does this exercise result impact a student's final overall assessment in the course (ie 5% of final score)?

16. What language tasks are being targeted in this exercise, ie after having the student write this task, what can you tell me about their language ability?

17. To what extent do the other tutors share your scoring rationale and explanation of the task purpose?

Impact

18. With regard to the certificate issued at the end of the course, to what extent is it recognized in the industry (ie does it have the endorsement of International House)?

19. To what degree of specficity does it describe what the learner can do as a result of the training they have received from the program?

20. Approximately how many clients have been served since 1998? How many are usually being served at any one period of time? How many teachers/markers are currently employed by netlanguages?

21. What are the nationalities of your clients? Best guesses on demographic profiles are ok.

22. I noticed that at the end of each section, there is a feedback section. Have there been any summative accounts of that information that I might have access to?

23. To what extent do the teachers know the criteria for scoring and objectives of learning? Two areas come to mind here:

1. For example, the description of mid-intermediate proficiency that you provided earlier: do the other teachers in your program share this view?

2. Related is the criteria for written work: for example, #4 style and layout; #5 coherent and communicative; these constructs could be interpreted differently by different teachers. To what extent has the interpretation been standardised?

24. Your concern about student motivation as the primary influence on assessment protocol: do the other stakeholders (ie do students, other teachers, administration) hold a similar view?

References

Alderson, J. (2000). Assessing Reading. Melbourne: CUP.

Alexander, S. (1999). An evaluation of innovative projects involving communication and IT in higher education. Higher Education Research and Development, 18 (2), pp 173 - 183.

Alexander, S. and Hedberg, J. (1994). Evaluating technology-based learning: Which model? In Beattie, K., McNaught, C. and Willis, S. (eds.) Multimedia in higher education: Designing for change in teaching and learning. (pp 233 – 244). Amsterdam: Elsevier.

Bachman, L. and Palmer, A. (1996). Language Testing and Practice. Hong Kong: OUP.

Bain, J. (1999). Introduction. Higher Education Research and Development, 18 (2), pp 165 – 172.

Brett, P (1998) An intuitive, theoretical and empirical perspective on the effectiveness question for multimedia. In Cameron, K. (ed.) Multimedia CALL: theory and practice. (pp 81 – 93). Exeter: ElmBank Publications. Available URL: http://pers-www.wlv.ac.uk/~le1969/

Brindley, G. (1989). Assessing achievement in the learner-centered curriculum**.** Sydney: NCELTR.

Breiner-Sanders, K., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines – speaking: revised 1999. Foreign Language Annals, 33 (1), 13 – 18. Available URL: http://www.actfl.org/public/articles/Guidelinesspeak.pdf

Brown, B. (1998). Is vocational education making a difference for high-risk populations? ERIC Myths and Realities. Washington, DC: Eric Clearinghouse on Adult, Career, and Vocational Education.

Busbee, E. (2001). The computer and the internet: Are they really destined to play a major role in English teaching? English Teaching, 56 (1), 201-225.

Byram, M. (1997). Teaching and assessing intercultural communicative competence. Clevedon, UK: Multilingual Matters:.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1, 1-47.

Canning-Wilson, C. and Wallace, J. (2001). Interactivity: the relationship between EFL and the human-computer interface. ELT Newsletter, 55 (April). Available URL: http://www.eltnewsletter.com/back/April2001/art552001.htm

Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms? Language Learning and Technology, 1 (1), pp. 19-43.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. Ch 2 (pp. 32-70) in L. Bachman and A.. Cohen (Eds.), Second language acquisition and – language testing interfaces. Cambridge: CUP.

Chapman, Gary (2000). Federal support for Technology in K-12 education. In Brookings papers on education policy: 2000. Brown Center for Education Policy. http://www.brookings.org/press/bpep/bpep00_intro.htm

Clarke, D. (1989). Communicative theory and its influence on materials production. Language Teaching, 22, 73 – 86.

Clarkson, R. and Jensen, M. (1995). Assessing achievement in English for professional employment programs. Ch. 7 (pp 165 – 194) in Brindley, G. (ed.) Language assessment in action. Sydney: NCELTR.

Claxton, C. (1999). Is the Internet the next big step forwards for MFL education? Unpublished Master's dissertation, King's College, London, UK. Available URL: http://www.dove-tail.com/claxton/Disserta.htm

Corcoran, C. (1999, May 9). Studying up on online courses: An audit of 4 classes. Mercury News, p. 1E. Available URL: http://www.mercurycenter.com

Cucchiarini, C., Strik, H., Boves, L. (2000) Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal. of the Acoustical Society of America, 107 (2), 989 – 999. Available URL: ftp://lands.let.kun.nl/pub/strik/publications/a67.pdf

Cummins, J. (1999). Biliteracy, empowerment, and transformative pedagogy. In J. V. Tinajero & R. A. DeVillar (Eds.), The power of two languages: 2000. (pp. 9-19). New York: McGraw-Hill. Available URL: http://www.iteachilearn.com/cummins/biliteratempowerment.html

Curnutt, L. (2000). Zeno's paradoxes. Available URL: http://www.scidiv.bcc.ctc.edu/Math/Zeno.html

Curtis, S., Duchastel, J., and Radic, N. (1999). Proposal for an online language course. ReCALL, 11 (2), pp. 38 – 45.

Daugherty, M., & Funke, B. (1998). University faculty and student perceptions of web-based instruction. Journal of Distance Education, 13 (1).Available URL: http://cade.athabascau.ca/vol13.1/daugherty.html

Dunn, R.; Griggs, S. (1995). Multiculturalism and learning style. teaching and counseling adolescents. Praeger: Westport, Connecticut.

Educational Testing Service. (1999). Computers and Classrooms: The Status of Technology in U.S. Schools. The Educational Testing Service Network, Policy Information Center. http://www.ets.org/research/pic/compclass.html

Eggins, S. (1994). An introduction to systemic functional linguistics.Continuum: New York.

Ellis, R. (1997). SLA research and language teaching. Melbourne. CUP.

Fleta, B., Sabater, C., Salom, L., Guillot, C., Monreal, C., and Turney, E. (1999). Evaluating multimedia programs for language learning: A case study. ReCALL, 11 (3), pp 50 – 57.

Freeman, M., & Capper, J. (1999). Exploiting the web for education: An anonymous asynchronous role simulation. Australian Journal of Educational Technology, 15 (1), 95-116. Available URL: http://www.asu.murdoch.edu.au/ajet/ajet15/freeman.html

Gatton, W. (1999). Call trends: A post-TESOL view. Retrieved July 1, 2000, from the World Wide Web: http://www.dyned.com/dyned/japan/htm/trend.htm

Glennan, T. & Melmed, A. (1996). Fostering the use of educational technology: Elements of a national strategy. (MR-682-OSTP). Santa Monica, CA: RAND. Available URL: http://www.rand.org/publications/MR/MR682/

Grierson, J. (1995). Classroom-based assessment in intensive English centers. Ch 8 (pp 195 – 237) in Brindley, G. (ed.) Language assessment in action. Sydney: NCELTR.

Gunn, M. (1995). Criterion-based assessment: a classroom teacher's perspective. Ch. 9 (pp 239 – 270) in Brindley, G. (ed.) Language assessment in action. Sydney: NCELTR.

Gunn, C. (1999). They love it, but do they learn from it? Higher Education Research and Development, 18 (2), pp.

Hara, N., & Kling, R. (2000). Students' distress with a web-based distance education course: An ethnographic study of participants' experiences. Information, Communication and Society, 3 (4), 557-579. Available URL: http://www.slis.indiana.edu/CSI/wp00-01.html

Harper, B., Hedberg, J., Bennett, S., and Lockyer, L. (2000). The on-line experience: The state of Australian on-line education and training practices. Kensington Park, Australia: NCVER.

Hedberg, J., Brown, C., Larkin, J., Agostinho, S. (2000). Designing practical websites for interactive training. Ch 27 in B. Khan (ed.) Web-based training. Englewood Cliffs, NJ: Educational technology Publications.

Hedberg, J., Brown, C., and Arrighi, M. (1997). Interactive multimedia and web-based learning: Similarities and differences. In Khan, B. (ed.) Web-based instruction (pp 47 – 58). Englewood Cliffs, NJ: Educational technology Publications.

Hegelhiemer, V. and Chapelle, C. (2000). Methodological issues in research on learner-computer interactions in CALL. Language Learning and Technology, 4 (1), 41 – 59. Available URL: http://llt.msu.edu/vol4num1/hegchap/default.html

Hoekje, B., and Linnell, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. TESOL Quarterly, 28, 103-126.

Johnson, M. (1999). CALL and teacher education: Issues in course design. CALL-EJ Online, 1 (2). Available URL: http://www.lerc.ritsumei.ac.jp/callej/4-2/johnson.html

Jones, F. (1998). Self-instruction and success: A learner profile study. Applied Linguistics, 19 (3), 378-406.

Jones, T. and Paolucci, R. (1999). A research framework for investigating the effectiveness of technology on educational outcomes. Journal of research on computing in education, 32, 17.

Jung, I., & Rha, I. (2000). The impact of information and communication technology in higher education: Experiences in Korea's virtual university. Retrieved November 10, 2000, from the World Wide Web: http://www.com.unisa.edu.au/cccc/papers/non_refereed/jung.htm

Kennedy, D. and McNaught, C. (1997). Design elements for interactive multimedia. Australian Journal of Educational Technology, 13(1), 1-22.

Knowles, M., Holton, E., and Swanson, R. (1998). The adult learner: The definitive classic in adult education and human resource development. 5th ed. Gulf: Houston.

Kolatch, E. (2000). Education and the Internet. Draft document for the Internet Policy Institute. http://www.cs.umd.edu/users/kolatch/education/

Korea Education and Research Information. (1998) White paper: Adapting education to the Information Age. Retrieved 2000/11/12 from http://www.keris.or.kr/web_zine3/index.htm

Korea Herald. (July 27, 2000). CampClick.com offers training in 5 foreign languages. Available URL: http://www.koreaherald.co.kr

Kuehn, P. (2000). Assessment of academic literacy skills: Preparing minority and limited English proficient (LEP) students for postsecondary education. Ch. 5 (pp 31 - 37) in Marcus, D., Cobb, E., and Schoenberg, R. (eds.) Lessons Learned from FIPSE Projects IV. U.S. Department of Education: Jessup, MD. Retrieved 10-04-01 from www.ed.gov/PDFDocs/FIPSE_IV.pdf

Lassche, G. (2000). Web-based language learning in Korea: a pedagogical critique. Korea TESOL Journal, 5, 55 – 76. Seoul: KOTESOL.

Lassche, G. (2001). Wired for Excellence: an evaluation checklist for web-based language learning programs. The English Connection, 5 (1), 1 & 6-7. Jan/Feb. Seoul: KOTESOL.

Laurillard, D., Stratfold, M., Luckin, R., Plowman, L. & Taylor, J. (2000). Affordances for learning in a non-linear narrative medium. Journal of Interactive Media in Education, 2. Available URL: www-jime.open.ac.uk/00/2

Leu, D. (2000). Literacy and technology: Deictic consequences for literacy education in an information age. In M. L. Kamil, P. Mosenthal, P. Pearson, and R. Barr (Eds.) Handbook of Reading Research, Volume III. Mahwah, NJ: Erlbaum. Available URL: http://web.syr.edu/~djleu/Handbook.html

Levy, M. (1997). Computer Assisted Language Learning: Context and conceptualization. Oxford, England: Oxford University Press.

Levy, M. (1999). CALL in context: Moving the research agenda forward. English Teaching [Korea], 54 (4), 239-255.

Lewkowicz, J. (2000). Authenticity in language testing: some outstanding questions. Language Testing, 17 (1), 43-64.

Lieb, S. (1999). Principles of Adult Learning. Available URL: http://www.hcc.hawaii.edu/ intranet/committees/FacDevCom/ guidebk/teachtip/adults-2.htm.

Lightbown, P. and Spada, N. (1993). How languages are learned. Oxford: OUP.

Long, M. and Robinson, P. (1996). Focus on form: theory, research and practice. Ch 2 (pp 15 – 41) in Doughty, C. and Williams, J. (eds.) Focus on form in SLA. Cambridge: CUP.

Lynch, B. (1996). Language program evaluation: Theory and practice. Cambridge: CUP.

Martin, J. (2000). Design and practice: enacting functional linguistics in Australia. Annual Review of Applied Linguistics, 20 (20th Anniversary Volume 'Applied Linguistics as an Emerging Discipline'), 116-126.

McMeniman, M. and Evans, R. (1998). CALL through the eyes of teachers and learners of Asian languages: Panacea or business as usual? OnCALL, 12 (1). Available URL: www.cltr.uq.edu.au/oncall/mcmen121.html

Mena, M. (1993). New pedagogical approaches to improve production of materials in distance education. CADE: Journal of Distance Education, 8 (3). Available URL: http://cade.athabascau.ca/vol8.3/10b_mena-english.html

Merrill, M. (2000). First principles of instruction. Presentation delivered to the AECT international conference, Denver, October 28, 2000. Available URL: http://www.id2.usu.edu/Papers/5FirstPrinciples.PDF

Messick, S. (1988). Validity. In Linn R. (ed.) Educational Measurement. New York: Macmillan.

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

Murchison, K. (1996). Business cooperation opportunities in Korea's information technology arena. Industry Canada International Operations Branch. Available URL: www.achilles.net/impact/korea2.html

Nunan, D. (1988). The leaner-centered curriculum. Melbourne: CUP.

Nunan, D. (1989). Designing tasks for the communicative classroom. Melbourne: CUP.

Nunan, D. (1997). Language teaching and research. 13-21. In Griffee and Nunan (eds.) Classroom teachers and classroom research. JALT Applied Materials. Tokyo: JALT.

Oliver, R., Herrington, J., Omari, A. (1996). Creating effective instructional materials for the World Wide Web. In Debreceny, R. and Ellis, A. (eds.), Proceedings of AusWeb'96. Lismore, NSW: Norsearch.

Oppenhiemer, T. (1997). The computer delusion. Atlantic Monthly, 280, 45-62.

Oregon Public Education Network. (2000). Site evaluation criteria. Available URL: http://www.open.k12.or.us/jitt/jitscore.html

PCER (The Presidential Commission on Education Reform). (1997). Education reform for the 21st century. Seoul: PCER.

Personal communication. (S.T., July 15 – August 10, 2001). Name of informant withheld to protect anonymity.

Phillips, J., Phillips, P., & Zuniga, L. (2000). Evaluating the effectiveness and the return on investment of e-learning. What Works Online, 2Q. Available URL: http://www.astd.org/virtual_community/research/What_Works/

Reeves, T., and Reeves, P. (1998). Effective dimensions of interactive learning on the WWW. In Khan, B. (ed.) Web-based instruction. (pp 59 – 66). Englewood cliffs, NJ: Educational Technology Publications.

Roman-Odio, C. and Hartlaub, B. (1998). Learning about assessment: a case study of a multimedia language program. Ohio 5 Foreign Language Technology Projects. Detroit: Mellon Foundation. Available URL: www.kenyon.edu/ohio5/assess.htm

Rovinelli, R., and Hambleton, R. (1976). On the use of content specialists in the assessment of criterion-referenced test item validity. Paper presented at the annual meeting of AERA, San Francisco. Eric Document # ED121845.

Jones, S. (1993). Cognitive learning styles: Does awareness help? A review of selected writings. Language Awareness, 2 (4), 195 – 207.

Salaberry, R. (1999). CALL in the year 2000: Still developing the research agenda. Language Learning and Technology, 3 (1), July 1999, pp. 104 – 107. Available URL: http://llt.msu.edu/vol3num1/comment/index.html

Seedhouse, P. (1994). Linking pedagogical purposes to linguistic patterns of interaction. International Review of Applied Linguistics, 32 (4), 309 – 326.

Seedhouse, P. (1995). Communicative CALL: focus on the interaction produced by CALL software. ReCALL, 7 (2), 20-28.

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching, 10, 209-231.

Schmitt, C., & Slonaker, L. (1996, January 14). High technology doesn't always equal high achievement. *Mercury New*s. http://www.mercurycenter.com/archives/reprints/edcom011496.htm

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. Language Testing, 13, 298-317.

Smith, M. and Salam, U. (2000). Web-based courses: a search for industry standards. CALL-EJ Online, 2 (1). http://www.lerc.ritsumei.ac.jp/callej/5-1/msmith&salam.html

Stilborne, L. and Williams, L. (1996). Meeting the needs of adult learners in developing courses for the internet. Available URL: http://www.isoc.org/isoc/whatis/conferences/inet/96/proceedings/c4/c4_2.htm

Taylor, J.C. (1994). Novex Analysis: A cognitive science approach to instructional design. Educational Technology, 34 (5), 5-13.

Trokeloshvili, D. and Jost, N. (1997). The internet and foreign language instruction: Practice and discussion. I-TESL Journal, 3 (8). Available URL: http://www.aitech.ac.jp/~iteslj/Articles/Trokeloshvili-Internet.html

US White House. (1997). Report to the President on the use of technology to strengthen K-12 education in the United States (White House Papers). Available URL: http://www.whitehouse.gov/WH/EOP/OSTP/NSTC/PCAST/k-12ed.html

Van Lier, L. (1988). The classroom and the language learner: ethnography and SL classroom research. London: Longman.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An over-view. Language Teaching, 31, 57-71. Retrieved November 14, 2000, from the World Wide Web: http://www.lll.hawaii.edu/web/faculty/markw/overview.html

Warschauer, M., & Meskill, C. (2000). Technology and second language learning. In J. Rosenthal (Ed.), Handbook of undergraduate second language education (pp. 303-318). Mahwah, New Jersey: Lawrence Erlbaum. Available URL: http://www.gse.uci.edu/markw/tslt.html

Weir, C. (1993). Understanding and developing language tests. New York : Prentice Hall.

Widdowson, H. (1979). Explorations in applied linguistics. Oxford: OUP.

Wideman, H., Owston, R., Handscombe, J. and Solomon, D. (1999). Web-based ESL learning: An assessment of the networked English language learning project. Technical Report 99-2. http://www.edu.yorku.ca/csce/nell/tech99-2.html

Wilson, G. and Cole, P. (1996). "Cognitive teaching models", in Handbook of Research in Instructional Technology. New York: Scholastic Press.

Yoon, A.S., & Kwon, H.C. (1998). Web-based language learning program: What to consider and how to develop? Korea Research Foundation Project #98-72. Retrieved November 10, 2000, from the World Wide Web: http://www.sccs.chukyo-u.ac.jp/ICCS/olp/o4-11/o4-11.htm

Zhao, Y.; Englert, C.; Chen, J.; Jones, S; and Ferdig, R. (2000). The development of a web-based literacy learning environment: A dialogue between innovation and established practices. Journal of Research on Computing in Education, 32 (4).

Zielinsky, D. (2000). The lie of online learning. Training, (February), 38-40.