Annual Review of Applied Linguistics (2005) **25**, 205–227. Printed in the USA.

11. TRENDS IN ASSESSMENT SCALES AND CRITERION-REFERENCED LANGUAGE ASSESSMENT

Thom Hudson

Two current developments reflecting a common concern in second/foreign language assessment are the development of: (1) scales for describing language proficiency/ability/performance; and (2) criterion-referenced performance assessments. Both developments are motivated by a perceived need to achieve communicatively transparent test results anchored in observable behaviors. Each of these developments in one way or another is an attempt to recognize the complexity of language in use, the complexity of assessing language ability, and the difficulty in interpreting potential interactions of scale task, trait, text, and ability. They reflect a current appetite for language assessment anchored in the world of functions and events, but also must address how the worlds of functions and events contain non skill-specific and discretely hierarchical variability. As examples of current tests that attempt to use performance criteria, the chapter reviews the Canadian Language Benchmark, the Common European Framework, and the Assessment of Language Performance projects.

Two complementary developments in second and foreign language testing relate to issues surrounding characteristics of proficiency or ability scales and how these scales are conceptualized in criterion-referenced performance assessment. These developments have been motivated by a perceived need to produce test results that are more transparent than has traditionally been the case. They are alternative views to those reflected by traditional testing enterprises that provide a single numerical score with an associated indication of the percentile position of the examinee's standing relative to other examinees who took the test. In many ways, each of these developments attempts to address the complexity of language in the assessment of language use. They reflect a current appetite for language assessment anchored in the world of functions and events. These developments interact to promote language assessment that recognizes the need to expand beyond a tradition that has focused on language primarily as a decontextualized cognitive skill or ability. Language takes place in a social context as a social act, and this frequently needs to be recognized in language assessment. This chapter examines three

language testing projects that have attempted to reflect these concerns: the Canadian Language Benchmarks project (Pawlikowska-Smith, 2000, 2002); the Common European Framework (Council of Europe, 2001; North, 2000); and the Assessment of Language Performance project (Norris, Brown, Hudson & Yoshioka, 1998; Brown, Hudson, Norris, & Bonk, 2002).

The focus here is on criterion-referenced testing projects that employ behaviorally oriented scales. As such, the three projects discussed here are not meant to provide a representative sample of all criterion-referenced testing endeavors currently being undertaken. There are many current test projects that are criterion-referenced in their construction process; indeed, *Education Week* (2002) indicates that all but three states in the United States use criterion-referenced tests in their English/language arts assessments. However, their reporting scales often tend to provide more general proficiency or skill descriptors than contextualized performance indicators. Hence, they are not directly addressed here. Their absence should not be taken as an indication that there are not a large number of other criterion-referenced testing projects in the United States or elsewhere.

It must be admitted openly that here are criticisms of, and drawbacks with, both performance assessment and many currently available language ability scales. There are differing views as to the utility and effectiveness of anchoring scale scores directly to performance tasks. However, there are areas of language use where assessment focuses on tasks that cannot be deconstructed into primary traits or skills and still capture the richness of the language performance. For example, tasks such as composing a synthesis from sources or writing a summary of a text inherently involve both reading and writing (Carson, 1993). Rating a synthesis such that the focus is solely on evaluating the characteristics of the composition itself ignores much that is of interest in the task. It may be the case that we need now to explore ways to report language as more complex literacy acts rather than simply reducing performance to one of the traditional four language skills. An analogy for the results of such reduction can be seen in how weather forecasters report the heat index along with measured temperature. If we say that it is 85 degrees because that is what the thermometer reads, when in fact there is 80% humidity with no breeze and the resulting heat index means that it feels like 92 degrees, then by only reporting 85 degrees we have not accounted for the actual effect of the weather on those who are living in it. Similarly, certain performances in Olympic sports such as diving and gymnastics are accorded a "degree of difficulty" rating considered inherent to the particular type of performance. Not all dive routines are considered to be about just whether the athlete moves from the diving board into the water. John Tukey notes that the most important maxim for data analysis to heed is: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (1962, pp. 13–14). There may be times when it will be most productive to look only at what the thermometer reads, while at other times it is most informative to look at the temperature in context and evaluate it as such.

Criterion-referenced tests are designed to assess the learners' knowledge of a well-defined domain of knowledge. As noted in the previous introduction to performance testing concerns, specifying that domain and operationalizing it are often difficult endeavors. How to approach such direct measures of performance is one current focus of criterion-referenced language testing. The general notion of criterion-referenced educational testing is usually traced back to Glaser and Klaus (1962) or to Glaser (1963). Although this is relatively recent given the long history of educational and psychological measurement, criterion-referenced testing has emerged as an important tool in educational testing circles over the past few decades. Although only about 40 years old in educational measurement, the central tenets of criterion-referenced testing have been around through history and pervade the ways humans deal with the world (Brown & Hudson, 2002). Conway Twitty, an American country and western singer, sings, "Don't call him a cowboy 'till you've seen him ride." The basic principles of directly referencing ability to a particular domain of behavior run deep in human interactions, and this anchoring of test results to a domain is the essence of criterion-referenced testing. How to put this criterion referencing in place relates directly to the ability scales that report score results.

Scales

Scales are implicit in measurement, and are central to the validity of test score interpretation (Alderson, 1991). There are *nominal* scales, such as teacher, student, husband, wife; *ordinal* scales, such as first, second, or third; and *interval* scales, such as test scores of 22 points, 75 points, and so on. There have also been scales that are not so clearly interpreted, such as freshman, sophomore, junior, where the scale is sometimes treated as nominal, sometimes ordinal, or sometimes interval in nature. Scales commonly include *size, amount*, *frequency*, *intensity, importance,* or *rank* related to the depth or breadth of a demonstrated ability. The two terms, scales and rubrics, are often used in interchangeably. Rubrics are defined in the *Standards for Educational and Psychological Measurement* as: The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items (AERA/APA/ NCME, 1999, p. 182).

Scales are sometimes used to indicate the numerical characteristics of the measurement. Generally, a rubric may be seen as a set of rules used in scoring performance assessment items, and is generally viewed as most useful for the assessment of tasks requiring responses that are other than selected response tasks. However, scales are frequently used to encompass the range of complex tasks along with the associated numerical scores. The terms scales and rubrics are used interchangeably here because of the way the terms are used in the literature, where sometimes there is a clear distinction and at others the two terms are conflated.

Typically, a rubric or scale

1. Is based on a continuum of performance quality, with a scale of varying potential score points to be assigned

2. Identifies the significant traits or dimensions to be examined and assessed (e.g., *reading*, *vocabulary,* or *listening comprehension*)
3. Provides key criteria of performance for each level of scoring, in "descriptors," which reflect whether and to what extent the key requirements of the performance have been demonstrated

There are several fundamental issues with scales in general that arise as we examine different scales and their contexts. First, there is a basic question of what the underlying nature of the scale is considered to be. Is the scale being used to indicate progression along a trait continuum or the actual achievement of defined and meaningful steps? Second, what kinds of comparisons are to be made with the scale, norm-referenced types of comparisons or criterion-referenced type comparisons? Third, does the scale really represent a set of discrete criteria that are themselves then rated on a numerical scale, as with analytic scoring of compositions or other performances? Fourth, what do the endpoints of the scale represent? Fifth, was the scale developed and evaluated empirically? These are areas that will come up throughout the discussion of scales, and they do not always have clear and satisfactory answers.

Language scales and rubrics can be created for many different functions, from large-scale high stakes tests that function to make decisions regarding university admission or immigration status all the way to self-assessment for purely personal interest. It is well to keep the different functions of scales in mind. Alderson (1991) has indicated that scales have the three different functions of: (1) describing levels of performance; (2) providing guidance for assessors who are rating the performances; and, (3) for guiding test constructors with a set of specifications. Additionally, Mislevy, Steinberg, Breyer, Almond, and Johnson (2002a) point out that the complexity of the variables to be included in the scale depends upon the purposes to which the assessment will be put. They note that, "[a] single variable characterizing overall proficiency might suffice in an assessment meant to support only a summary pass-fail decision" whereas "a coached practice system that helps students develop the same proficiency would require a finer grained student model for monitoring how a student is doing on particular aspects of skill and knowledge for which we can provide feedback" (Mislevy, Steinberg, & Almond, 2002b, p. 367). So, some scales may simply function to provide univariate information while others are designed to communicate more richly contextualized description.

However, not all scales are equally helpful in describing language ability. There is a trade-off in terms of generalizability versus deeper description. This, in part, is because of the complexity of the construct. The attempt to develop a comprehensive scale for language ability that is succinct enough to be easily comprehensible yet is transparent and functional presents us with complications of some depth. Language is perhaps the most complex of human abilities, and consequently we can expect that its assessment will be equally complex. We are, after all, assessing an individual's performance interacting within a very social context.

As Brindley (1998) has pointed out, language scales of achievement or proficiency tend to fall into one of two types. The first type of rating scale is one that is defined independently of content and context. It is derived from a theoretical model of language, and attempts to define a decontextualized ability or proficiency. The following example after Wilds (1975), shows one example of a scale that does not explicitly indicate context, content, or performance conditions.[1] The scale does not anchor either end of the continuum for the different evaluation dimensions.

1. Accent foreign ___: ___: ___: ___ native
2. Grammar inaccurate ___: ___: ___: ___ accurate

The scale proposed by Bachman (1990) and Bachman and Palmer (1983, 1996) in Table 1 is also of this decontextualized scale type, but provides more explicit discussion of the intermediate stages between the two end points of 0 or 4. It matches performance against vocabulary and cohesion in pragmatic competence.

Table 1 Sample of Bachman and Palmer decontextualized scale (1983, 1996)

Pragmatic competence		
Rating	**Vocabulary**	**Cohesion**
0	*Extremely limited vocabulary* (A few words and formulaic phrases. Not possible to discuss any topic, because of limited vocabulary)	*No cohesion* (Utterances completely disjointed, or discourse too short to judge.)
1	*Small vocabulary* (Difficulty in talking with examinee because of vocabulary limitations.)	*Very little cohesion* (Relationships between utterances not adequately marked; frequent confusing relationships among ideas.)
2	*Vocabulary of moderate size* (Frequently misses or searches for words.)	*Moderate cohesion* (Relationships between utterances generally marked; sometimes confusing relationships among ideas.)
3	*Large vocabulary* (Seldom misses or searches for words.)	*Good cohesion* (Relationships between utterances well-marked.)
4	*Extensive vocabulary* (Rarely, if ever, misses or searches for words. Almost always uses appropriate word.)	*Excellent cohesion* (Uses a variety of appropriate devices; hardly ever confusing relationships among ideas.)

Adapted from Bachman (1990)

Scales of this type are terse, efficient, and seemingly straightforward in their application. Their primary advantage is that they specify the language components that are of importance and provide specific reference to them. A potential

disadvantage is that operationalizing terms like "small vocabulary" and "vocabulary of moderate size" clearly becomes normative in nature. Finally, scales of this type consciously exclude mention of context and content. The scale criteria are structured to represent broad learning targets rather than specific tasks.

This perspective follows from Bachman's long held position that in language assessment it is of paramount importance to "clearly distinguish the ability to be measured from the methods or procedures used to elicit evidence of this ability" (Bachman, 1988, p. 150). Such a scale attempts to provide a decontextualized indication of a person's language ability according to different trait competencies that are intended to represent the construct of language ability. Bachman further notes that the "interpretation of test scores is problematic, since traits are frequently difficult to distinguish from methods. This is particularly true with performance tests, such as oral interviews, in which the modality (productive) and channel (oral/aural) of the ability (speaking) match the modality and channel of the elicitation procedure, or test method, thereby making it difficult to clearly distinguish ability from test method" (p. 153). Bachman and Palmer indicate that the "construct definitions from which rating scales can be developed may be based on either a theoretical model of language ability… or on the content of a language learning syllabus. Both of these theoretical construct definitions refer only to areas of language ability, independent of any considerations of the characteristics of the specific testing situation and prompt with which they might be used." (1996, p. 213). More recently, however, Bachman has indicated that performance assessments need to be both construct-based and task-based (2002).

The second type of scale noted by Brindley (1998), the type of primary interest here, is behaviorally based and attempts to describe proficiency according to "real-world" performance in specific contexts. Borman (1986) indicated several different types of behavior-based rating scales. The first type represents the Behaviorally Anchored Rating Scales (BARS). These scales list descriptions of very specific behaviors, and examinees are rated as to whether they reflect these specific behaviors. This type of scale has the potential drawback that some of the description might fit the examinee whereas some components do not. For instance, the descriptor might say "Can carry out an effective fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies." However, this description might be only partially true, in that the examinee might not decide to follow up and probe interesting replies. Thus, another type of scale Borman introduces is the Behavior Summary Scale (BSS). These scales anchor the performance to less specific behaviors that represent a more generalized level of ability by representing a wider range of behavior representative of several specific incidents considered to be at a common level. The third approach that Borman introduces is the Behavior Observations Scale (BOS). This approach takes a different strategy in that observable behavioral statements are presented, and the rater is asked to determine if this is true about the candidate on a scale of frequency, such as *almost never* to *almost always*. Regardless of the particular approach, each is an attempt to link performance to behaviors that represent important behavioral criterial factors of assessment.

The functionality and apparent transparency of behavioral performance scales is the reason that they have generally had the most influence, although this does not absolve them of their problems. The descriptions of the levels often specify the particular tasks associated with each of the levels in the scale. The examinee's performance on the tasks is taken as an indicator that can generalize to a universe of similar tasks. It is this interpretation that is so troubling to many critics of behavioral scales. There are basically two approaches to developing such scales. The first is to assert intuitive orders of language performance, whereas the second approach is to elicit task orders from experts, such as teachers, and then to test them to see which ones order in an empirical manner.

Three Criterion-Referenced Projects

Three current orientations to behavioral scales as represented in the Canadian Language Benchmarks (Pawlikowska-Smith (2000, 2002), the Council of Europe Common European Framework (Council of Europe, 2001; North, 2000), and the Assessment of Language Performance task-based assessment project (Brown et al., 2002; Norris et al. 1998), attempt to address descriptions of language ability in differing ways. All these scales are steeped in notions of communicative competence as it has emerged since Canale and Swain's (1980) seminal article in *Applied Linguistics*. They take into account notions of language competence, strategic competence, sociocultural competence, textual competence, and so on. Additionally, they do not assume that an idealized native speaker is the goal.[2]

The Canadian Language Benchmarks

The Canadian Language Benchmarks Assessment (CLBA; Norton & Stewart, 1999; Pawlikowska-Smith, 2002) represents an example of scales following an intuitive developmental approach. The CLBA is a task-based assessment for adult immigrants to Canada intended to help place adult language learners across Canada in instructional programs appropriate for their level of proficiency in English (Norton & Stewart, 1999). The benchmarks are based on a functional view of language, language use, and language proficiency, explicitly following this behavioral orientation. The developers note that "[s]uch a view relates language to the contexts in which it is used and the communicative functions it performs" (Pawlikowska-Smith, 2002, p. 6). Here, "communicative proficiency is not an abstract concept of absolute language ability. Rather, it depends on situations of language use" (p. 6). The CLBA instruments address several functions, and in this they may be too broadly conceived. These functions are: (1) a descriptive scale of communicative proficiency; (2) a set of descriptive standards; (3) statements of communicative competencies and performance tasks in which the learner demonstrates application of knowledge competence and skill; (4) a framework of reference for learning, teaching, programming and assessing adult English as a second language in Canada; and (5) a national standard for planning second language curricula for a variety of contexts and a common "yardstick" for assessing the outcomes (Pawlikowska-Smith, 2000, p. viii). As such, the benchmarks are intended to be one-size-fits all scales for multiple uses.

The competencies are "directly observable and measurable performance or measurable outcomes of instruction in a curriculum framework" (Pawlikowska-Smith, 2000, p. 25). The competencies and tasks are seen to be only samples indicative of the range of a person's language ability at a particular benchmark level. "Similar competencies require increasing complexity of performance across the three stages of proficiency because of the progressively demanding tasks, contexts and performance expectations" (p. 25). There are three general levels with four benchmark divisions within each of those levels (see Table 2). The structure of Table 2 indicates that there are benchmarks for each of the four skill competencies. One aspect that is somewhat confusing is that in the CLBA descriptors in Table 2, listening and speaking are listed together, but in actuality each does receive a separate benchmark score. The three general levels move from nondemanding contexts and simple texts, through moderately demanding and complex texts, to demanding contexts and complex texts.

Table 2 Organization of Canadian Language Benchmark components

AN OVERVIEW				
Benchmark	Proficiency Level	Speaking and Listening Competencies	Reading Competencies	Writing Competencies
STAGE I: BASIC PROFICIENCY				
1	Initial	Creating/interpreting oral discourse in routine non-demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) • Information	Interpreting simple texts: • Social interaction texts • Instructions • Business/service texts • Information texts	Creating simple texts: • Social interaction • Recording information • Business/service messages • Presenting information
2	Developing			
3	Adequate			
4	Fluent			
STAGE II: INTERMEDIATE PROFICIENCY				
5	Initial	Creating/interpreting oral discourse in moderately demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) • Information	Interpreting moderately complex texts • Social interaction texts • Instructions • Business/service texts • Information texts	Creating moderately complex texts: • Social interaction • Recording information • Business/service messages Presenting information/ ideas
6	Developing			
7	Adequate			
8	Fluent			
STAGE III: ADVANCED PROFICIENCY				
9	Initial	Creating/interpreting oral discourse in very demanding contexts of language use in: • Social interaction • Instructions • Suasion (getting things done) Information	Interpreting complex and very complex texts • Social interaction texts • Instructions • Business/service texts • Information texts	Creating complex and very complex texts: • Social interaction • Recording information • Business/service messages Presenting information/ ideas
10	Developing			
11	Adequate			
12	Fluent			

Note how this approach to scale development explicitly incorporates performance conditions. For example, Table 3 shows the performance conditions associated with Benchmark 6. These explicitly situate the speech sample to be evaluated by noting such conditions as "Interactions are face to face or on the phone."

Table 3 Performance conditions from Canadian Language Benchmarks (adapted, Pawlikowska-Smith, 2002)

Speaking: Stage II Benchmark 6
Performance Conditions: • Interaction is face to face, or on the phone, with familiar and unfamiliar individuals and small informal groups. • Rate of speech is slow to normal. • Context is familiar, or clear and predictable. • Context is moderately demanding (e.g., real-world environment, limited support from speaker). • Circumstances range from informal to more formal. • Setting or content is familiar, clear and predictable. • Topic is concrete and familiar. • Presentation is informal or formal. • Use of pictures or other visuals. • Presentation is five to seven minutes long
Interactions one-on-one • Interactions are face to face or on the phone. • Interaction is formal or semiformal • Learner can partially prepare the exchange.
Interactions in a group • Interaction occurs in a familiar group of three to five people. • Topic or issue is familiar, nonpersonal, concrete. • Interaction is informal or semi-formal.

Also, see how the benchmarks progress as shown through the global performance descriptors in Table 4.

Table 4 An overview of Speaking Benchmarks-Global performance descriptors. First Benchmark for each stage, and final Benchmark example (Source: Pawlikowska-Smith, 2000)

B.1 Learner can speak very little, mostly responding to basic questions about personal information and immediate needs in familiar situations. Speaks in isolated words or strings of 2 to 3 words. Demonstrates almost no control of basic grammar structures and verb tenses. Demonstrates very limited vocabulary. No evidence of connected discourse. Makes long pauses, often repeats the other person's words. Depends on gestures in expressing meaning and may also switch to first language at times. Pronunciation difficulties may significantly impede communication. Needs considerable assistance.

> *Sample Tasks: Hello, how are you? My name is X. Please come in, wait. Please sit down. Excuse me, Bob. Help me please. Answer questions about basic personal information in short interviews with teachers, other learners, and counselors.*

B.5 Learner can participate with some effort in routine social conversations and can talk about needs and familiar topics of personal relevance. Can use a variety of simple structures and some complex ones, with occasional reductions. Grammar and pronunciation errors are frequent and sometimes impede communications. Demonstrates a range of common everyday vocabulary and a limited number of idioms. May avoid topics with unfamiliar vocabulary. Demonstrates discourse that is connected (and, but, first, next, then, because) and reasonably fluent, but hesitations and pauses are frequent. Can use the phone to communicate simple personal information; communication without visual support is still very difficult.

> *Sample Tasks: Respond to small talk comments. Express and respond to compliments and congratulations. Extend an invitation for a coffee, dinner, party. Direct a person to a place with or without maps, diagrams, sketches. Request permission to leave work early or take a day off.*

B.9 Learner can independently, through oral discourse, obtain, provide, and exchange key information for important tasks (work, academic, personal) and complex routine and a few nonroutine situations in some demanding contexts of language use. Can actively and effectively participate in 30-minute formal exchanges about complex, abstract, conceptual, and detailed information and ideas to analyse, to problem-solve, and to make decisions. Can make 15- to 30-minute prepared formal presentations. Can interact to coordinate tasks with others, to advise or persuade (e.g., to sell or recommend a product or service), to reassure others, and to deal with complaints in one-on-one situations. Grammar, vocabulary, or pronunciation errors very rarely impede communication. Prepared discourse is mostly accurate in form, but may often be rigid in its structure/organization and delivery style.

Sample Tasks: Convey appropriately respect, friendliness, distance and indifference in a variety of conversations in a variety of contexts. Give complex instructions on familiar first aid and emergency procedures in the work place. Discuss concerns about your child's progress in school with the child's teacher and school principle.

B.12: Learner can create and co-create oral discourse, formal and informal, general or technical, in own field of study or work, in a broad range of complex situations…Discourse is fluent and "natural" (native-like in phrasing). Language is complex…

The tasks associated with each of these levels are then assessed against a checklist for Effectiveness, Organization, Appropriateness, Grammar, Vocabulary, Legibility /Mechanics, Cohesion, and Relevance. Thus, a profile of scales can be presented as an alternative to the type of scale intended to present only a single numeric score. These CLBA scales are broad and do not use a native speaker as the norm, although they do use such terms as "native-like." Further, they have not yet been empirically validated. However, also note that in this approach, the scales indicate at times what the learner cannot do. For example, B1 says "No evidence of connected discourse" or "Almost no control of basic grammar structures." This differs from the approach taken in the next scale, the Common European Framework.

Common European Framework

Scales such as the *Common European Framework of Reference for Languages* (CEF) developed by the Council of Europe, and the related scales of the Association of Language Testers in Europe, and the Dialang project, provide a framework that allows for more restricted descriptions of language where only partial language knowledge is required (Council of Europe, 2001; North, 2000). This view of scales provides a "can do" approach that recognizes lower levels in the scale as having a place of functional importance. It is the Council of Europe's position that giving formal recognition to these partial and functional abilities will promote plurilingualism through the learning of a wider variety of European languages (p. 1–2). In the CEF, the "ideal native speaker" is not the ultimate model. This is reflected in the form and scope of the scales as shown in Table 5. There are six scales divided into three larger bands. To the degree that they have the three overall levels divided into sublevels, they are similar to the Canadian benchmarks. However, these descriptors do not have negative directionality as with the other scales. That is, there are no statements of the form "has a speaking vocabulary sufficient to respond simply with some circumlocutions; accent, although often quite faulty, is intelligible," or "no evidence of connected discourse."

Table 5 Common European Framework—global scale

Proficient	C2	Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
User	C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic, and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors, and cohesive devices.
Independent	B2	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.
User	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes, and ambitions and briefly give reasons and explanations for opinions and plans.
Basic	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g., very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
User	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Source: Council of Europe (http://culture2.coe.int/portfolio//documents/0521803136txt.pdf)

The CEF aims to provide a comprehensive, transparent, and coherent framework of reference for language learning, teaching, and assessment. It is designed to be useable across the many different languages of Europe. The purpose of the CEF is to: (1) promote and facilitate cooperation among educational institutions in different countries; (2) provide a basis for mutual recognition of language specifications; and, (3) assist learners, teachers, course designers, examining bodies, and educational administrators to situate and coordinate their efforts.

In constructing these scales, a comprehensive survey of over 30 existing language scales was carried out. The contents of each of the reviewed scales were broken up into sentences. Each sentence in these scales was analyzed to determine what category it seemed to be describing. In looking at the various scales, six levels emerged and were adopted. The over 2,000 potential descriptors that emerged were converted into statements that could be answered *yes* or *no*. Duplicated descriptors from across the scales were eliminated. In a series of workshops, teachers evaluated the descriptors and indicated which were desirable and which were not.

Finally, the descriptors were evaluated against examinee videotaped performances. Teachers observed the tapes and scored a form of the observation questionnaire (North, 2000). This process yielded scores for each of the descriptors in terms of the examinee's language ability level. Items that did not scale or that did not fit the model were eliminated. The descriptors were then calibrated using Multifaceted Rasch (MR) model analysis (Linacre, 1992). From this analysis, the scale level descriptors were created. That is, the sentences in each level of the scale were ordered hierarchically. Thus, here is an empirically evaluated scale of language ability. This is a primary difference from the CLBA, which is not empirically based. Some of the descriptors can be seen in Table 6. Note again, that the statements are all "can do" statements designed to indicate the positive aspects of the learner's language. This places a focus on the functional abilities that the examinee has as opposed to focusing on the examinee's linguistic shortcomings.

Table 6 Example descriptors for the CEF (Adapted from North, 2000)

C1:
Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
Can relate own contribution skillfully to those of other speakers.
Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.
Can carry out an effective fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies.
Can follow the essentials of lectures, talks, and reports and other forms of academic/professional presentation that are propositionally and linguistically complex.
Can develop an argument systematically with appropriate highlighting of significant points, and relevant supporting detail.
A2:
Can write simple notes to friends.
Can ask and answer questions about personal details, such as here they live, people they know, and things they have.
Can reply in an interview to simple direct questions spoken very slowly and clearly in direct nonidiomatic speech about personal details.
Can indicate time by such phrases as next week, last Friday, in November, three o'clock.
Can understand instructions addressed carefully and slowly to him/her and follow short, simple directions.

However, as a result of the MR analysis, the Council of Europe test developers were unable to include certain aspects of language use, areas such as literary appreciation and several pragmatic and strategic aspects of language. These areas appeared to represent different factors or aspects of language use than language proficiency. Consequently, there are questions about the comprehensiveness of the scale as a full description of language user abilities. A second potential problem with interpreting the initial CEF is that whereas the descriptors were empirically scaled based on performance ratings, the particular descriptors were not subsequently cast as actual test prompts and then calibrated again to determine if they still scale hierarchically.

Assessment of Language Performance

Central to performance assessment throughout the CLBA and CEF scalar models is the concept of the language task. *Real-world* tasks play a central role in the design of various types of performance assessments. The Assessment of Language Performance (ALP) project at the University of Hawai'i focused on how real-world tasks can function to reveal an examinee's language ability in use for pedagogical goals (Brown, Hudson, Norris, & Bonk, 2002). The project recognized that ultimately task accomplishment is a focus for evaluating much of human performance. It follows that L2 general performance assessment and task-based approaches to language assessment will likely share a great deal of theoretical and practical common ground. After all, task-based language teaching has received increasing recognition in the second language acquisition and second language pedagogy literature over the past two decades. By employing the communicative task as the basic unit of analysis for motivating syllabus design and L2 classroom activities, advocates claim that contemporary theories of language learning and acquisition that are supported by empirical findings can be effectively implemented. Here, task-based tests are held to be assessments that require students to engage in some sort of behavior which simulates, with as much fidelity as possible, goal-oriented target language use outside the language test situation. Performances on these tasks are then evaluated according to pre-determined real-world criterion elements (i.e., task processes and outcomes) and criterion levels (i.e., authentic standards related to task success) (Brown & Hudson, 2002; Brown et al., 2002). In task-based performance assessment, the goals are generally to: (1) discuss a means whereby examinee performance on real-world language tasks can be validly assessed in terms of real-world criteria; (2) illustrate the potential for using task-based performance assessment to generalize about examinees' L2 abilities; and 3) facilitate a direct link between L2 classroom learning and real-world language use.

The ALP incorporated Skehan's condensation of prior definitions of task, in which he presents the following parameters as fundamental for a task activity:

a. Meaning is primary
b. Learners are not given other people's meanings to regurgitate
c. There is some sort of relationship to comparable real-world activities
d. Task completion has some priority

e. The assessment of the task is done in terms of outcome (Skehan, 1998, p. 147).

Additionally, the ALP addressed estimated task difficulty through adaptations to Skehan's *language code complexity*, *cognitive complexity*, and *communicative stress* factors. Tasks were selected and examined in relation to these variables to determine whether a task should be predicted to be more or less demanding for the examinees. For example, language code complexity in the following task could be adjusted as indicated.

> *Task: You would like to try out the fancy new Italian bistro "Il Gondoliero" tonight. Look up the phone number of the restaurant in the phone book and call to reserve a table for one at an appropriate time this evening.*

Low language code difficulty conditions could involve presenting the examinee with a simple telephone book layout with restaurants identified in the yellow pages format. The telephone message could involve standard formulaic expressions, simple single word responses to the telephone message. High difficulty code could involve a linguistically difficult message, low frequency vocabulary, or heavily accented speech with the telephone message delivered at a high rate of speech.

A number of test and item specifications, modeled after Popham (1981), specifying real-world task simulations and scales to assess examinee performance on each one, were developed to represent exemplars of the approach[3]. The examinees worked with a number of tasks which varied in complexity from fairly easy to very demanding. An example task in which an examinee must assist a friend who has injured his hand is shown in Table 7.

Table 7 Sample simulation task F05 (Source: Norris et al., 1998)

Situation: Your friend John has broken a bone in his hand. He cannot write (see photo of John). You told him that you would help him with writing. Now, he wants you to fill out a *change of address* form for him. Study the form provided. Be prepared to listen for the information requested on the form. John said he would leave the information on your answering machine.

Task: Play the message from John. Listen for the information from the *change of address* form. Fill in the form for John. You may listen to the message as many times as you need to get the correct information.

Time: You have **10 minutes** to complete this task.

Product: Completed *change of address* form.

Such a task requires multiple modalities, both listening and writing. Table 8 presents a rating scale designed specifically for that task. Each task on the ALP had a separate task-dependent rating scale. Categories on these scales refer specifically to success of task requirements.

Table 8 Example task-dependent rating scale for task F05

1	2	3	4	5
Inadequate		**Able**		**Adept**
Examinee incorrectly fills out change of address form such that any essential elements (listed in the *able* descriptor) are not processable by the post office (this might include illegibility, incorrect placement of information, absence of information, etc.	Examinee performance contains some elements from the *inadequate* descriptor and some elements from the *able* descriptor.	Examinee fills out change of address form according to information given by John, minimally including with correct spelling and correct locations (see form for details) —name —new address —old address —starting date —signature and printed name (either John Harris or examinee's own name).	Examinee performance contains some elements from the *able* descriptor and some elements from the *adept* descriptor.	Examinee correctly fills out change of address form with ALL applicable information given by John on the answering machine message (see form for details).

Additionally, a task-independent scale that reflected raters' evaluation to how the examinee performed across all of the tasks during the course of the test was developed. That is shown in Table 9. Both scales ranged from a designation of *inadequate* to *adept*. The task dependent scale related to categories of task success, an outcome that can only be realized if the examinee controls the language of both the input and the output.

Table 9 Example of task-independent rating scale across all task performances

1.	*Inadequate*: A rating of insufficient indicates that the student seems generally incapable of coming to terms with the particular processing component (code, cognitive, communicative) on tasks like those found on the (test).
2.	Student performance contains some elements from the *inadequate* descriptor and some elements from the *able* descriptor.
3.	*Able*: A rating of able indicates that the student seems generally capable of coming to terms with the particular processing component on tasks like those found on the (test).
4.	Student performance contains some elements from the *able* description and some elements from the *adept* descriptor.
5.	*Adept*: A rating of adept indicates that the student seems quite capable of coming to terms with the particular processing component on tasks like those found on the (test); additionally, the student seems to have little or no difficulty in accomplishing such tasks in terms of the processing component.

In developing the task dependent scales, a criteria identification team was formed of three people familiar with the types of tasks on the test. These were a highly experienced ESL/EFL teacher, an advanced L2 user of English with much experience in accomplishing the types of tasks on the test, and a member of the university community with experience working with international students. Over a period of time, the members of the criteria team met and: (1) became familiar with the specific tasks; (2) produced drafts of what minimally sufficient, insufficient, and efficient task performances would look like; (3) worked with the drafts trying to rate actual performances; and, (4) jointly revised on agreed upon scoring rubrics for each item. Note that the task-independent descriptor does not actually mention the particular abilities that are to be sampled on the test. Rather, they refer the rater to the performance of the examinee across a range of tasks, and thus to a more global concept of language ability in context. Interestingly, correlations between the task-dependent scale and the task-independent scale turned out to be about .90. So, these two scales represent different approaches to language task performance scale development, although they do overlap a great deal in the actual rating variance they account for. Also, test results indicated high correlations between predicted difficulty and examinee performance.

Discussion and Conclusions

The topics addressed regarding performance and task based criterion-referenced scale development raise several issues. Three specific instances of scale development and performance assessment from different perspectives have been discussed. It is clear that there are very strong criticisms and questionings of such

scales and task-based testing from a theoretical perspective that asks just what the goal of measurement is. In many ways, these reservations reflect the earlier mention of Bachman's criticism of behavioral scales. The association of language ability with authenticity of language use and setting raises real issues regarding the relationship of competence and performance, given that competence can only be inferred via some sample of performance (Shohamy, 1995). Further, a concern has grown about the extent to which successful task performance inherently involves nonlanguage abilities (Bachman, 1990; McNamara, 1996).

As noted, the difficulties in proceeding with behavioral scale application to task-based assessment need to be acknowledged, and it is certainly not appropriate for all language assessment to be task-based. However, also as noted earlier, not all language use is meaningfully interpretable as representing one of the traditional language skills. Rather, complex performances in social contexts require that they be interpreted with fidelity to what those performances mean in that context. Language traits interact with context, and a great deal of research has shown that identified research variables often show unpredicted and seemingly inexplicable interactions. These interactions can cause mischief as we try to identify the precise construct that we are measuring. However, there are ways we can try to remedy this indeterminacy. Those interactions may be partially dealt with using the statistical machinery available to us now. Multifaceted Rasch analysis and G-theory show a great deal of promise in finding and accounting for the relative effects of contextual features that we identify. In fact, if contextual features do affect how the particular language ability construct is engaged, then we should be purposefully seeking those features. Additionally, newer and emerging approaches such as the "evidence-centered design" model proposed by Mislevy et al. (2002a) that is aimed to design complex tasks, evaluate students' performances, and draw valid conclusions from them may be of assistance in the future.

Much more research needs to be done in several areas:

1. What are the relationships between task-dependent scales and task independent scales?
2. How do task specifications, task content, and scoring criteria interact?
3. How are examinee performances affected by task difficulty, task complexity, task conditions, task characteristics, examinee's perceptions of task and raters?
4. To what extent can multifaceted Rasch model, G-theory, and the "evidence-centered assessment design" approach assist in disentangling factors that affect performance?

Certainly, the literature points out numerous disadvantages and problems with performance assessment in general. It has been observed that performance tests: (a) are difficult to create, (b) typically require more time and resources to administer and score than do other test types, (c) are accompanied by a variety of logistical problems (e.g., transporting and storing materials and realia), (d) may cause

formidable reliability problems (because of both test administration and scoring inconsistencies), (e) may only lead to very restricted kinds of test-based interpretations, and (f) often face increased test security risks (Khattri, Reeve, & Kane, 1998).

Such disadvantages or problems notwithstanding, performance assessment of some sort seems essential to meet the kinds of assessment demands that are increasingly associated with L2 education contexts. Despite the difficulties, performance assessment is directly concerned with construct validity in its approach to finding tasks that can be generalized to real world language tasks. In contrast to the disadvantages just presented, several specific advantages often associated with language performance tests are that they: (a) can be designed to simulate authentic language use with high fidelity, (b) may compensate for negative effects often associated with traditional standardized testing, and (c) may initiate positive washback effects on language pedagogy and curriculum design. Those tasks in an educational setting can be directly linked to the curricular objectives that are increasingly communicatively oriented (Khattri, et al., 1998).

This discussion has attempted to show some of the different concerns that are addressed with the development of language proficiency scales and criterion-referenced task-based assessment. It has indicated how behavioral scales are an attempt to be explicit about what language learners are capable of doing with the language that they have. Clearly there are potential problems with this sort of endeavor. Finding descriptors that do actually relate to particular levels along the scale is difficult. However, the CLBA, CEF, and the ALP approaches are examples of attempts to do just that.

The concerns faced by language testers are very well illustrated by Mislevy, Steinberg, and Almond in relation to contemporary measurement in general:

> Standard procedures for designing and carrying out assessments have worked satisfactorily for the assessments we have all become familiar with over the past half century. Their limits are sorely tested today. The field faces demands for more complex inferences about students, concerning finer grained and interrelated aspects of knowledge and the more complicated conditions under which this knowledge is brought to bear. (2002a, p. 126)

The literature presents some evidence that these demands may be addressed through more attention to the nature of our measurement scales and criterion-referenced behavior and task selection, as well as with more attention to the growing technology that may provide measurement tools.

Notes

1. It should be noted, however, that this is simply a short-hand reporting and record keeping scale. The original Wilds scale does have a separate set of descriptive and somewhat contextualized scale descriptions.

2. This discussion does not treat the FSI/ACTFL/ILR language scales. Those scales have been discussed in depth elsewhere (ACTFL, 1989; Bachman, 1988; FSI, n.d.; Lantolf & Frawley, 1985; Lee & Musumeci, 1988; Park, 1999). Additionally, although these scales have been claimed to be criterion-referenced, they are more properly seen as proficiency-referenced scales with the task criteria selected to reflect proficiency levels in a norm-referenced manner.

3. Note that the tasks developed here were intended to display a test development approach and methodology. They were not developed from a specific target language use needs analysis. We do argue that in any programmatic application of the approach, a comprehensive needs analysis is essential to fit the assessment to the appropriate context.

ANNOTATED BIBLIOGRAPHY

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.). *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

A chapter that clearly presents many of the issues in the development and interpretation of scales that describe language ability. It presents the underlying differences between decontextualized scales and behavioral scales that describe performance.

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawai'i Press.

This book presents the follow-up results of the project initially described in Norris, et al. (1998). It describes the extent to which the ALP tasks demonstrated hierarchical orders of difficulty that were predicted. It also presents reliability and validity results for the study.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

This text presents a systematic approach to the construction and analysis of criterion-referenced language testing. It comprehensively relates language criterion-referenced testing to criterion-referenced testing in other educational areas, discussing where the two are complementary and where they have differences.

Council of Europe. (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

This text is the definitive description of the Common European Framework scales, the uses for the scales, and the political context in which they were developed. It discusses the full scope of the scale use, including the language portfolio framework that is an extension of the language testing component of the program. It discusses how the framework provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, and so on. across Europe.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

This study presents the development of the Common European Framework. It presents the rationale of the project, the scale development process, the piloting of the scales, through to the final process of which language features were included and which were not. Anyone interested in the detailed methodology of the Framework will profit from reading this text.

Pawlikowska-Smith, G. (2000). Canadian language benchmarks. English as a second language—adults. Retrieved March, 2003 from the Centre for Canadian Language Benchmarks www.language.ca/bench.html.

Pawlikowska-Smith, G. (2002). Canadian language benchmarks 2000: Theoretical framework. Retrieved March, 2003 from the Centre for Canadian Language Benchmarks www.language.ca/pdfs/final_theoreticalframeworks.pdf.

These two works present the benchmarks themselves as well as a comprehensive theoretical framework and rationale for the Canadian Language Benchmark Assessment. The works show how the benchmarks describe a learner's communicative proficiency as the skills of speaking, listening, reading, and writing over three stages of progression in the specific competency areas of social interaction, giving and receiving instructions, suasion, and information. They notes that the benchmarks are considered to be directly observable and measurable performance outcomes.

OTHER REFERENCES

American Council on the Teaching of Foreign Languages (ACTFL). (1989). The ACTFL provisional proficiency guidelines. In T.V. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 219–226). Lincolnwood, IL: National Textbook Company.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North. *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*, 149–164.

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

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *2,* 5–18.

Bachman, L. F., & Palmer, A. S. (1983). *Oral interview test of communicative proficiency in English*. Urbana, IL: Photo-offset.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 100–120). Baltimore: The Johns Hopkins University Press.

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

Carson, J. (1993). Reading for writing: Cognitive perspectives. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 85–104). Boston: Heinle and Heinle.

Education Week. (2002, January 10). Editorial projects in education, 17. Retrieved Septemeber 3, 2004, from http://nces.ed.gov/programs/digest/d02/tables/dt153.asp.

Foreign Service Institute (FSI). (n.d.) Absolute language proficiency ratings. In M. L. Adams & J. R. Frith (Eds.), *Testing kit: French and Spanish* (pp. 13–17). Washington, D.C.: Department of State.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521.

Glaser, R. & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in systems development* (pp. 419–474). New York: Holt, Rinehart, & Winston.

Khattri, N., Reeve, A., & Kane, M. (1998*). Principles and practices of performance assessment*. Mahwah, NJ: Erlbaum.

Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. The *Modern Language Journal*, *69*, 337–345.

Lee, J. F. L., & Musumeci, D. (1988). On hierarchies of reading skills and text types. *The Modern Language Journal*, *72*, 173–187.

Linacre, J. M. (1992). *Many-faceted Rasch measurement*. Chicago: Mesa Press.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002b). Making sense of data from complex assessments. *Applied Measurement in Education. 15*, 363–389.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002b). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development,* (pp. 97–128). Mahwah, N.J.: Erlbaum.

Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998*). Designing second language performance assessments*. Honolulu: University of Hawai'i Press.

Norton, B., & Stewart, G. (1999). Accountability in language assessment of adult immigrants in Canada. *Canadian Modern Language Review*, *56* (2), 223–244.

Park, S. (1999). *Testing the EFL skills and text hierarchy of the ACTFL reading guidelines*. Unpublished masters thesis, Department of English as a Second Language, University of Hawai'i.

Popham, W. J. (1981). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188–211.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1–67.

Wilds, C. P. (1975). The oral interview test. In, B. Spolsky & R. Jones (Eds.) *Testing language proficiency* (pp. 29–44). Washington, DC Center for Applied Linguistics.