2 *Construct definition and validity inquiry in SLA research*

Carol A. Chapelle

In second language acquisition (SLA) research, some form of measurement is frequently used to produce empirical evidence for hypotheses about the nature and development of communicative competence. For example, SLA researchers test learners to investigate such aspects of interlanguage vocabulary[1] as the acquisition of semantic (Kellerman 1978) and syntactic (Ard & Gass 1987) features of words, the structure of the L2 lexicon (Meara 1984; Singleton & Little 1991), lexicon size (Nation 1993), strategies associated with vocabulary use (Blum-Kulka & Levinson 1983), and automaticity of lexical access (Chitiri, Sun, Willows, and Taylor 1992). Tests are used for investigating vocabulary, as well as for SLA research in general, to elicit learners' performance in a defined context. In other words, taking the complement to Douglas's (Chapter 6, this volume) view of language tests as SLA elicitation devices, I consider SLA elicitation devices from the perspective of two principles that underlie language testing: construct definition and validation.

These principles are important in SLA research because learners' performance on elicitation devices – like performance on language tests – is used to make inferences extending beyond the observed performance. For example, inferences are often made concerning the learner's underlying competence, which is a construct. Justification of the inferences made on the basis of performance is validation. On SLA elicitation devices, researchers would seldom rely on a single observation of performance to make inferences. Instead, performances are typically summarized across observations to produce scores or other descriptions of performance consistency. In SLA research – like language testing (LT) research – demonstration of performance consistency is significant because when learners' consistent performances across observations are summarized, the resulting score or profile is more dependable than any one idiosyncratic observation can be. When researchers summarize performance consisten-

1 Interlanguage vocabulary is used here to indicate learners' vocabulary that is less developed than that of a native speaker. As explained in this chapter, this term can be viewed from all three perspectives of construct definition. Interlanguage vocabulary refers to one component within a broad definition of communicative language ability.

cies and use those summaries to make inferences beyond the actual performance, they are working within the domain of psychological measurement – a domain that supports theory and practices relevant to the use of measurement in SLA research. In particular, measurement theory offers perspectives on (1) defining the construct(s) believed to be reflected by consistent performance and (2) justifying test performance as a valid indicator of the construct(s).

The purpose of this chapter is to explore how principles of construct definition and validity inquiry apply to SLA research. Using research on interlanguage vocabulary as an example, the first part examines the nature of construct definition. It explains three theoretical perspectives toward construct definition – trait, behaviorist, and interactionalist – by demonstrating how interlanguage vocabulary can be defined within each and how each is reflected in vocabulary testing. In my view, current theory in applied linguistics favors an interactionalist approach to construct definition.[2] Therefore, the second part of the paper explains the implications of such a definition for validity inquiry. It defines validation as the ongoing process of justifying particular interpretations and uses of test results, and it explores implications of an interactionalist approach to construct definition for investigating construct validity and the consequences of testing.

Construct definition

Because SLA researchers study interlanguage constructs, it is crucial to define the term. A *construct* is a meaningful interpretation of observed behavior. When a researcher interprets a learner's score on a vocabulary test, for example, as an indicator of vocabulary knowledge, then "vocabulary knowledge" is the construct that gives meaning to the score. The fundamental requirement for interpreting observed behavior as a construct is that the behavior reflect performance consistency. The consistency requirement has caused some researchers to question the usefulness of tests in SLA research and practice (e.g., Lantolf & Frawley 1988; Swain 1990) because of the variable and changing nature of interlanguage. However, as Bachman (Appendix, this volume) points out, the intention of SLA research is to document and explain the learner's changing interlanguage, and to do so, researchers need reliable descriptions of language at its various stages of development. Reliable pictures of interlanguage

2 I interpret the work on communicative competence and communicative language ability as pointing to the need to hypothesize an interactionalist construction for language measurement. Note, in contrast, however, that Eckman (1994) assumes that the trait-oriented construct definition (i.e., the linguist's notion of "competence") is the obvious way of approaching construct definition in SLA research.

are obtained when consistent performance is observed because consistencies allow researchers to

> move from the level of discrete behaviors or isolated observations to the level of measurement. This is not to say that scores for individual items of discrete behaviors are not often of interest but, rather, that their meaning and dependability are fragile compared with response consistencies across items or replications. (Messick 1989: 14)

Observation of performance consistency is therefore fundamental to the use of empirical performance data to infer constructs such as "interlanguage vocabulary." The problem of construct definition is to hypothesize the source of performance consistency.

Theorists' various perspectives of construct definition, therefore, can be understood by identifying how they explain response consistency (Messick 1981). *Trait theorists* attribute consistencies to characteristics of test takers, and therefore define constructs in terms of the knowledge and fundamental processes of the test taker. A trait perspective on interlanguage vocabulary must include the dimensions of vocabulary knowledge (e.g., size) and fundamental processes (e.g., lexical access) that have been investigated in SLA research. *Behaviorists* attribute consistencies to contextual factors (e.g., the relationship of participants in a conversation), and therefore define constructs with reference to the environmental conditions under which performance is observed. A behaviorist definition of interlanguage vocabulary must specify features of the context that SLA researchers have identified as affecting vocabulary use. *Interactionalists* see performance as the result of traits, contextual features, and their interaction. An interactionalist definition of vocabulary will include dimensions of both trait and context, although each will be constrained by the other. In addition, this definition must include the strategies required to mediate between the person and the context (e.g., goal setting based on an assessment of the context).

In the following sections, the three approaches to construct definition are introduced by citing the SLA research supporting each and outlining the components of each type of definition. Since SLA researchers' definitions of interlanguage have implications for the types of tests they choose (e.g., Skehan 1987), some of the principles and methods of measurement implied by each type of definition will be explained. Discussion of the three types of definitions and their uses will show that much of the research on interlanguage vocabulary appears to assume a trait-oriented definition, but it is the interactionalist perspective that is consistent with current theory in applied linguistics (Bachman & Cohen, Chapter 1, this volume). Taking the interactionalist perspective as most relevant to future SLA research, I then discuss how validity inquiry for tests used in SLA research can be informed by an interactionalist construct definition.

A trait perspective of interlanguage vocabulary

A trait definition of interlanguage would attribute test performance to the characteristics of the learner, because "for trait theorists (and cognitive theorists as well), scores are largely signs of underlying processes or structures" (Messick 1989: 15). Strictly speaking, a trait is defined as "a relatively stable characteristic of a person – an attribute, enduring process, or disposition – which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (Messick 1989: 15). The notion of stability is not essential to the trait definition, however, because a trait can be expected to change as the result of deliberate study (Carroll 1993: 7) or other special conditions, as in the case of second language (L2) development. The primary identifying feature of a trait is that it can be defined as a person characteristic that is displayed across a variety of settings and therefore can be defined without careful specification of the settings in which it might be observed.[3]

Interlanguage researchers have hypothesized trait-type theories of interlanguage by attributing performance across relevant measures to underlying characteristics. Trait-type construct definitions of language have been articulated and used by SLA researchers such as Bialystok and Sharwood Smith (1985), who describe interlanguage as a system that "depends upon the dual influence of the learner's level of analysis of knowledge and control of cognitive procedures." Learners' performance or "product is a description of essentially the values which occur on these dimensions" (p. 116). In their view, the focus of construct definition for SLA researchers is describing the learner's underlying "system," which is responsible for "products." Figure 2.1 illustrates the assumed relationship between a construct conceived as an underlying system and observed products which appear as performance consistency. In other words, performance consistency is attributed to underlying characteristics of the learner.

TOWARD A TRAIT DEFINITION OF VOCABULARY

A trait theorist's view of interlanguage vocabulary attempts to define "implicit knowledge" (Ard & Gass 1987) and fundamental processes that would be relevant to vocabulary performance across a variety of contexts. A trait-oriented definition would therefore consist of the four knowledge and process dimensions that SLA researchers have investigated,

3 Of course, all performance is the result of contact between person and context; therefore, even a trait theorist would agree that context does influence performance. The point is that the focus of construct definition (to interpret performance consistency) is on person characteristics, and particularly on those characteristics believed to behave similarly in different contexts.

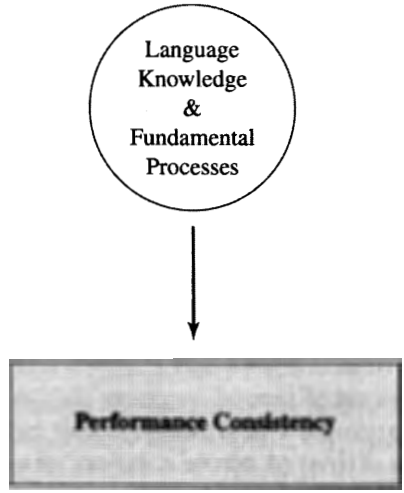

Figure 2.1

each of which is defined without reference to a context of language use. The first dimension, *vocabulary size,* denotes the absolute number of content words a person knows. Estimates of absolute native speaker vocabulary size differ (Aitchison 1987), but as learners' second language develops, the size of their vocabulary grows (Blum-Kulka & Levinson 1983).

The second dimension is *knowledge of word characteristics,* including phonemic, graphemic, morphemic, syntactic, semantic, and collocational features. During the process of acquisition, knowledge of specific words can be incorrect, incomplete, or unanalyzed (Bialystok & Sharwood Smith 1985). Incorrect knowledge refers to lexical representations that do not correspond to the target language, such as erroneous orthographic or semantic representations. Incomplete knowledge refers to the gaps in the learner's knowledge of a word, which may result in confusion between two words with similar forms (Laufer 1990). Unanalyzed knowledge is what the learner knows as a unit but cannot break up or use creatively. For example, in early stages the learner often knows words as they occur in phrases but does not know how to change them morphologically or use them in other phrases.

The third dimension of the trait definition, *lexicon organization,* refers to the way morphemes and words are represented in the mental lexicon, as well as the way they are connected to one another by, for example, semantic and phonological features. The change in vocabulary organization that accompanies acquisition is termed *restructuring* (McLaughlin 1990) or *reanalysis* (Gass 1988). Although debate continues on the representa-

tion of morphemes in the mental lexicon (e.g., Henderson 1985; Sternberger & MacWhinney 1988), most agree that the connections among words in a native speaker's mental lexicon are primarily semantic. In contrast, the lexicon of the low ability L2 learner has been described as more loosely organized (Ard & Gass 1987), with connections made on the basis of phonological features (Meara 1984).

The fourth dimension of a trait definition is a set of fundamental vocabulary processes (Sternberg 1977) associated with lexical access. Vocabulary processes would include the following: attending to relevant vocabulary features in written or spoken input, encoding phonological and orthographic information into short-term memory, accessing structural and semantic properties from the lexicon (e.g., Yang & Givon 1993), integrating the semantic content of the word with the emergent semantic representation of the input text (Marslen-Wilson 1989), parsing words into their morphological components, and composing words morphologically (Olshtain 1987).[4] These fundamental processes are tied closely to the three aspects of language knowledge defined earlier. For example, speed of lexical access is believed to depend, in part, on the organization of the lexicon (Frenck-Mestre & Vaid 1992). Furthermore, Olshtain (1987) describes morphological parsing and composing as both "word formation processes" and "knowledge of word formation rules" (p. 221). In short, the processes appear to be viewed by SLA researchers as specific to the vocabulary trait. What these four dimensions of the trait definition – size, knowledge of word characteristics, organization, and fundamental processes – have in common is that each is defined and studied independently of the context of language use.

MEASUREMENT IMPLICATIONS OF A TRAIT DEFINITION

In interlanguage vocabulary research, trait approaches to construct definition are apparent not only by the way researchers define vocabulary constructs but also by the principles they use for test construction and validity justification. In constructing tests, trait theorists rely on random sampling of content from the relevant domain so that test performance can be considered an accurate sign of vocabulary ability across a wide variety of contexts. For example, taking a trait perspective of vocabulary, Nation (1993) explained ideal procedures for sampling from a dictionary to develop tests of vocabulary size. He notes that if the researcher follows these sampling procedures, "it is possible to make an estimation of vocabulary size that can be generalized beyond the particular dictionary

4 The processes associated with production are hypothesized to be more difficult than those required in comprehension because when a word is produced, its formal properties must be accessed and composed morphologically to fit properly into the linguistic output (Teichroew 1982).

studied" (p. 32). A second principle of test construction for the trait theorist is to minimize the effects of context on performance by placing vocabulary items within a minimal discourse context. For example, the trait "vocabulary organization" has been investigated by researchers who have chosen such tests as the word association test, which presents words in isolation (Meara 1978, 1984), and grammaticality judgments, which present words in isolated sentences (Ard & Gass 1987). Word recognition processes (i.e., accessing semantic features) have been assessed by presenting subjects with isolated pairs of words (one in the native language and the other in the target language) and timing test takers' judgment of whether or not the pair is the same semantically (Yang & Givon, 1993). All three of these tests, in keeping with principles of trait theory, present language in settings other than the type of discourse context that would require learners to perform tasks encountered during normal communicative discourse.

In justifying the use of tests chosen to measure vocabulary traits, researchers use both judgmental and empirical arguments. For example, Ard and Gass (1987) justify their choice of grammaticality judgment tests as follows: "We are attempting to understand the implicit knowledge that learners have about the relationships among words in their mental lexicon. This is only ascertainable through specific probings of intuitions, which provide a much more direct window on implicit knowledge than do other types of data (cf. Bialystok 1981)" (p. 238). Justifying the choice of two tests for investigating learners' "lexical confusions," Laufer (1990) supplies the following rationale: "The fact that test version A tested synforms [i.e., lexical confusions] in sentences, while test version B tested them in isolation, does not mean we wanted to check context effect on synform confusion. The versions were simply two elicitation methods" (p. 285). This use of different test methods solely for the purpose of double-checking that the trait was measured accurately epitomizes the trait theorist's ideal that observed performance be attributable to the underlying capacity of the test taker and not to features of the operational setting (i.e., the test). The researcher states no interest in "context effect" (which here refers to the context of the sentence in the test items) on performance. Instead, performance is to be attributed to a trait dimension – the quality of word representation in the mental lexicon (i.e., knowledge of word characteristics).

A behaviorist perspective of interlanguage vocabulary

A behaviorist's view of interlanguage vocabulary would include the features of context relevant to vocabulary use. In contrast to trait theorists, who interpret performance as a sign of underlying characteristics,

> in psychological measurement, behaviorists and social behaviorists usually interpret scores as samples of response classes. A response class is a set of behaviors all of which change in the same or related ways as a function of stimulus contingencies, that is a class of behaviors that reflect essentially the same changes when the person's relation to the environment is altered (Messick 1989: 15)

For behaviorists, then, the relevant dimensions are not person characteristics, but characteristics of the context in which performance occurs. Accordingly, the scores obtained from tests are interpreted by behaviorists as "derived from responses made to [carefully defined] stimuli for the purpose of predicting responses made to similar naturally occurring stimuli found in vocational, academic, and other settings" (Tryon 1979: 402).[5]

A behaviorist construct definition is implicit in the work of some SLA and LT researchers. SLA researchers focusing on the contextual features influencing performance have discussed the issue in terms of "variability" in interlanguage (e.g., Tarone, Chapter 3, this volume). Although appearing to be the opposite of the performance consistency of interest to LT researchers, some variability – that recognized as "systematic" variability (e.g., Ellis 1989) – refers to performance consistency under particular conditions. Conditions might include the linguistic environments or the sociolinguistic contexts in which particular interlanguage forms are most likely to appear (Ellis 1989). When SLA researchers attempt to explain systematic variability by specifying the contextual factors influencing performance, they are defining performance consistency within particular contexts. Use of contextual factors to explain performance consistency reflects a behaviorist approach to construct definition.

From the same theoretical stance, LT researchers working within the tradition of performance testing attempt to create test methods that replicate the conditions of the settings for which they wish to predict the test taker's future performance. Wesche (1987) describes the rationale as follows:

> Since it appears to be impossible to establish the communicative load of . . . different kinds of knowledge precisely because they interact with each other in a compensatory way in performance, it will probably not be possible to establish context-free, universally fair language tests. Rather one might hope to improve the predictive validity of tests by recreating those contextual features that theory and experience suggest may have an important influence on language performance. (Wesche 1987: 31)

Representing the modern behaviorist position implicit in performance testing, Wesche's statement indicates that the learner's underlying knowledge

5 In SLA research the "other settings" of interest to the researcher are "target language use" settings: the contexts in which learners will use the target language (Bachman & Palmer 1996).

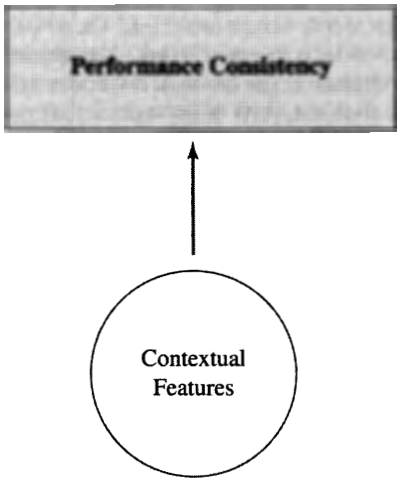

Figure 2.2

is too elusive to define, and therefore, as Figure 2.2 illustrates, the objective of construct definition becomes context definition. In other words, performance consistency is attributed to the context in which it occurs.

TOWARD A BEHAVIORIST DEFINITION OF VOCABULARY

A behaviorist definition of vocabulary would therefore comprise the descriptions of contexts believed to affect vocabulary performance. Relevant contexts are often described imprecisely, by using phrases such as "vocabulary knowledge in reading comprehension" (Luppescu & Day 1993: 263–264) or by referring to local discourse contexts of vocabulary as defined by "phrasal constraints" or "situational utterances" (Nattinger & DeCarrico 1989). Recognizing the importance of context, SLA researchers are beginning to take a more comprehensive view of context definition. For example, in an attempt to better understand the construct of task, Duff (1993) examined the role of task characteristics like "direction of interaction" and "nature of gap between subject and interviewer on task" in predicting interlanguage performance (p. 65). In his work on nonnative speakers' use of inflectional morphology, Young (1989) has suggested that the essential elements to include in such a definition would be those Hymes (1967) identified: *setting, participants, ends* (i.e., purpose), *art characteristics, communicative key, instrumentality* (e.g., spoken vs. written), *norms* (i.e., sociolinguistic rules of conversation), and *genre*. Although his research investigated only the role of participants, Young hypothesized that each of these features of context would affect learners' use of inflectional morphemes (i.e., one aspect of vocabulary),

and therefore that observed performance must be explained with reference to these features of context. Because little work has attempted to add to our understanding of a behaviorist definition of vocabulary, these features of context are adopted here to illustrate the behaviorist approach to construct definition. In other words, from a behaviorist perspective, consistency in vocabulary performance might be hypothesized to be defined by Hymes's eight features of context.

MEASUREMENT IMPLICATIONS OF A BEHAVIORIST DEFINITION

One measurement implication of a behaviorist perspective as explained by Wesche (1987) is that tests must be designed by carefully constructing test environments that mirror the contexts in which the test writer wishes to predict performance (see the "target language use situation" described by Bachman & Palmer 1996). In other words, tests cannot be constructed through random selection of items from a large domain of possible items. The behaviorist does not assume that performance will be generalized over a wide variety of contexts, but only to those contexts which are similar to the test setting. In constructing a test of "reading vocabulary," then, the behaviorist would attempt to mirror the context in which the learner will be reading in the future. If researchers used Young's behaviorist construct definition to design such a test, Hymes's features of context would guide comparison of the elements of the test setting to those of the reading context. For example, a test of reading comprehension vocabulary might be constructed to fit the following contextual specifications. Test takers read at home *(setting)* alone *(participants)* to answer questions at the end of the chapter *(ends);* this would be treated as an important class assignment in a real class, so the text must be read for its meaning *(art characteristics)*. The reading consists of a written text *(instrumentality)* composed by an unknown, respected scientist who uses formal language *(communicative key)* conforming to the norms of academic written language *(norms)* and containing the specific linguistic signals that identify the content and structure of an academic text of its type *(genre)*. Performance on such a test would be interpreted as a sample of performance in target language use settings that reflect similar contextual values.

A second measurement implication follows from the first: Because target language use situations consist in part of language in discourse contexts, the behaviorist's vocabulary test must present and elicit vocabulary within a discourse context. Unlike Laufer (1990), who did not want to "check context effect on synform confusion" (p. 285), the behaviorist attempts to create a test that will check the relevant context effects. The behaviorist considers the well-chosen discourse context in a vocabulary test to be essential to the test, unlike the trait theorist, who views it as a source of error in observed performance.

TABLE 2.1. MEASUREMENT PRINCIPLES ASSOCIATED WITH TRAIT AND BEHAVIORIST DEFINITIONS OF INTERLANGUAGE VOCABULARY

	Trait principles	*Behaviorist principles*
Meaning of performance	A sign of underlying characteristics.	A sample of performance across similar contexts.
Construct definition	Implicit characteristics must be specified independent of context.	Implicit characteristics cannot be specified. Context must be specified.
Test construction	Random sampling of content allows generalizability across all contexts.	Careful selection of content allows generalizability across similar contexts.
	Contextual influences are considered irrelevant.	Contextual influences are considered relevant.
	Minimize these by presenting and eliciting language out of context.	Maximize these by presenting and eliciting language in relevant contexts.
Validity justification	Compare performance on different test methods to distinguish method error from the trait variance.	Compare the test context to the context of interest to identify similarities and differences.

Because the behaviorist views performance as a sample, justification of test use (i.e., test validity) requires the researcher to demonstrate that test performance is a good sample of the behavior that would occur in a real setting. Therefore, some researchers attempt to document the authenticity of the test setting relative to the context in which the test user wishes to make predictions. This approach to testing research has resulted in attempts to define and investigate "authenticity" (see, for example, volume 2, number 1 of *Language Testing,* 1985), which compare characteristics of the test setting with those of the target language use situation. Researchers also attempt to demonstrate that significant variance cannot be accounted for by irrelevant aspects of the operational test setting, but that variance can be attributed to facets of the test methods that reflect the same contextual values present in the target language use setting.

Table 2.1 clarifies the contrasts between the measurement assumptions underlying the trait and behaviorist perspectives. Despite the polarity of the two approaches, SLA and LT researchers increasingly recognize that progress in understanding language development and use rests on understanding how traits and contexts interact during the process of communication. This interactionalist perspective requires that the contrasts be-

tween the trait and behaviorist perspectives be seen as complements in an interactionalist construct definition that specifies how language traits are put into use in contexts.

An interactionalist perspective of interlanguage vocabulary

The third approach to construct definition, the interactionalist approach, requires the researcher to specify the relevant aspects of both trait and context. Interactionalist perspectives of construct definition (e.g., Zuroff 1986) represent "intermediate views, attributing some behavioral consistencies to traits, some to situational factors, and some to interactions between them, in various and arguable proportions" (Messick 1989: 15). Despite the fact that the interactionalist construct definition includes both trait and context, it cannot be derived by simply adding trait and behaviorist definitions together. Instead, when trait and context dimensions are included in one definition, the quality of each changes. Trait components can no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics. From the interactionalist perspective, *performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts.* Moreover, to incorporate a dimension of interaction between trait and context, an interactionalist definition must include metacognitive strategies responsible for mediating between the two. For example, the language user's metacognitive strategies (e.g., "assessing the context") would intervene between the context of language use and the user's knowledge during performance to assess the relevant features of context (e.g., level of formality) and decide which aspects of knowledge (e.g., which words) were needed.

SLA AND LT RESEARCHERS AS INTERACTIONALISTS

There is strong support for an interactionalist perspective in both SLA and LT theory. In SLA, the interactionalist perspective has been fueled primarily by the theory of communicative language use, which suggests that communicative competence refers to both knowledge of language and the ability to put language to use in context (Hymes 1972; Canale & Swain 1980; Widdowson 1983). In addition, many "variability" researchers who examine performance data conclude that performance is the result of "both learner internal and environmental (i.e., input) sources" (Ellis 1989: 25), and therefore a need exists for "studies showing the ways in which these various influences *interact* in normal communication" (Tarone 1988: 136; emphasis in original). Some researchers have attempted to outline the strategies required for communication in context (Faerch & Kasper, 1983) and to identify methods for studying

such strategies (Cohen & Robbins 1976; Cohen 1984; Faerch & Kasper 1987).

In LT the interactionalist perspective has gained empirical support through research which has found that traits and methods (the latter interpreted as realizing contextual variables) each contribute unique variance to test performance (Bachman & Palmer 1982; Bachman, Davidson, & Foulkes 1993). On the basis of this empirical work and communicative competence theory, Bachman (1990) composed a general interactionalist construct definition of communicative language ability, which includes "both knowledge, or competence, and the capacity for implementing, or executing that competence in language use" in context (Bachman 1990: 84). An interactionalist definition includes the language knowledge and fundamental processes of the trait theorist as well as the context of the behaviorist. When these two parts appear in a single construct definition, the need for a component controlling the interaction between the two is apparent. Bachman defined this as "strategic competence," the metacognitive strategies required for assessing contexts, setting goals, constructing plans, and controlling execution of those plans (Faerch & Kasper 1983). In other words, an interactionalist construct definition comprises more than trait plus context; it includes the metacognitive strategies (i.e., strategic competence) responsible for putting person characteristics to use in context.

TOWARD AN INTERACTIONALIST DEFINITION OF VOCABULARY

To hypothesize a framework for an interactionalist definition of interlanguage vocabulary, it will be necessary to revisit the trait and context dimensions outlined earlier. When traits and contexts are considered together, however, both lose their general relevance. Trait dimensions must be specified more precisely with reference to the context of language use, as Table 2.2 indicates. "Vocabulary size," for example, cannot be defined in an absolute sense but instead is a meaningful construct only with reference to a particular context (Dollerup, Glahn, & Hansen 1989). Similarly, the "linguistic characteristics" a learner knows about individual words would be prompted by and would depend on the contextual factors occurring when those words are used. Moreover, from the interactionalist perspective, word knowledge must include the word's pragmatic features – knowledge of the appropriate contexts of word use and the perlocutions the word can produce in those contexts. "Vocabulary organization" would not be fixed at any stage of development, but the connections the language user made between words would depend on contextual factors prompting those connections (Votaw 1992). Fundamental lexical processes would also be defined relative to the context of language use so that a researcher, rather than investigating "lexical recognition," might investigate lexical recognition during reading processes (Chitiri et

TABLE 2.2. HYPOTHESIZED INFLUENCES ON PERFORMANCE ACCORDING TO TRAIT, BEHAVIORIST, AND INTERACTIONALIST DEFINITIONS OF INTERLANGUAGE VOCABULARY

Performance influence	*Trait definition*[a]	*Behaviorist definition*[b]	*Interactionalist definition*[c]
Knowledge	Size Linguistic characteristics Organization		Size in context Linguistic and pragmatic characteristics in context Organization in context
Process	Fundamental lexical processes		Fundamental lexical processes in context
Context		Setting Participants Ends Art characteristics Key Instrumentation Norms Genre	Field Tenor Mode
Metacognitive strategies			Assessment Goal setting Planning Execution

[a]Comprising trait dimensions investigated by interlanguage researchers. One might find additional dimensions that could be added to a trait definition.
[b]Based on Young's (1989) suggestion for the study of interlanguage variation. One might choose other features of context.
[c]My hypothesis of how trait and context might be combined at the theoretical level.

al. 1992) in particular contexts. In short, according to the interactionalist definition, the dimensions of the trait definition of interlanguage vocabulary will differ qualitatively depending on the context in which vocabulary is used, and therefore must be specified with reference to that context.

The context dimension of an interactionalist definition must provide a theory of how the context of a particular situation within a broader context of culture constrains the linguistic choices a language user can make during linguistic performance. Systemic theory provides a general theory to do just that. Encompassing Hymes's (1972) features of context while allowing flexibility to add additional features, Halliday and Hasan

(1989) present a theory of context comprising three theoretical components: field, tenor, and mode. Reflecting applied linguists' understanding of context, these three components are intended to be complex and overlapping. They work together in any language use situation to define a range of potential language that may occur. Field refers to the location(s), topic(s), and action(s) present in a particular language use context. Tenor includes the participants, their relationships, and objectives. Mode includes the channel, texture, and genre of situated language. These three context constructs of systemic theory work together to define a particular "contextual configuration," which consists of specific values for field, tenor, and mode. To use this theory of context in consort with the trait dimension "vocabulary size," for example, it would be necessary to define vocabulary size within a particular field, tenor, and mode. Learners' vocabulary size would be expected to differ depending on whether they were reading a psychology text at home, for example, or listening to a biology lecture in a classroom.[6] These differences, according to the interactionalist definition, are important for defining and assessing vocabulary size.

In addition to trait and contextual features, an interactionalist definition must include the metacognitive strategies required for vocabulary use in context. The strategies Bachman (1990) defined as strategic competence – assessing the situation, setting goals, planning language use, and controlling execution of plans (Faerch & Kasper 1983) – are not specific to vocabulary but are an important part of the definition of interlanguage vocabulary. Metacognitive strategies are different from the fundamental processes defined by the trait theorist because the former are tied directly to the context of language use, and therefore in order to be defined meaningfully must be specified with reference to a particular context of language use. For example, Blum-Kulka and Levinson (1983: 126) have identified some of the strategic plans associated with vocabulary use: circumlocution, paraphrase, language switch, appeal to authority, change of topic, and semantic avoidance. These strategies, unlike fundamental processes such as lexical access or morphological parsing, are "situation bound" (Blum-Kulka & Levinson, 1983: 126). Figure 2.3 illustrates the multiple influences the interactionalist construct definition hypothesizes

6 According to Halliday and Hasan's (1989) theory of context, "reading a psychology text at home" and "listening to a biology lecture in a classroom" could be specified at various levels of delicacy through the field, tenor, and mode constructs. A key problem in an interactionalist definition is finding an optimal level of delicacy for defining contexts that are meaningful to test users and in which consistent performance can be observed. In other words, even when the future language use context (e.g., academic reading [EFL] in international universities) is identified, the question remains of how specifically one needs to define such a context (e.g., "all academic reading," "science reading," "chemistry reading," "research articles in chemistry," "research articles on a particular topic within chemistry" . . .).

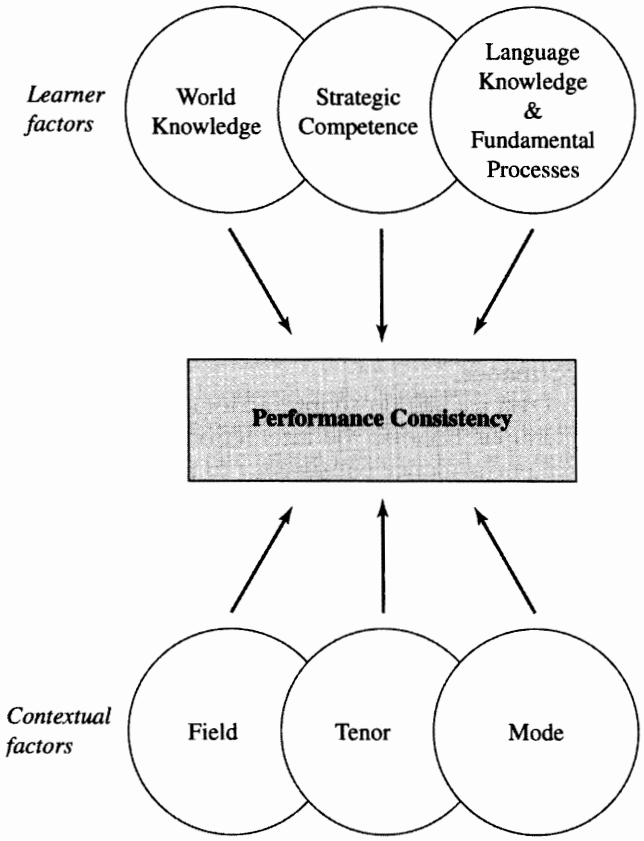

Figure 2.3

for performance consistency. Performance consistency is attributed to learner characteristics (knowledge and processes, metacognitive strategies, and world knowledge) and the values of the contextual variables of field, tenor, and mode.

MEASUREMENT IMPLICATIONS OF AN INTERACTIONALIST DEFINITION

An interactionalist approach to construct definition poses difficult problems for measurement because it combines two philosophies that embody contrasting ideals, as indicated in Table 2.1. With respect to construct definition, interactionalist theory, like trait theory, requires that implicit knowledge be specified. What the researcher must specify are the knowledge and fundamental processes that are required within a particular context as well as the metacognitive strategies controlling performance in

that context. In test construction, interactionalist theory, like behaviorist theory, requires that test content be the result of careful sampling from the context of target language use. However, unlike the behaviorist, who might rely on superficial similarities between the test methods and the future language use context (e.g., "authenticity"; Spolsky 1985), the interactionalist must also consider the underlying abilities judged to be required in the test context. Test construction using language in a discourse context is also the ideal for the interactionalist. But unlike the behaviorist, who simply attempts to mirror the context of future language use to improve prediction, the interactionalist attempts to use discourse to elicit the defined linguistic knowledge, processes, and metacognitive strategies during test performance.

Justification of test use within an interactionalist framework also poses some unique dilemmas because the interactionalist construct definition ascribes observed performance consistency to the combined influence of person characteristics and contexts. This view of performance consistency presents a challenge for traditional testing research. Fortunately, modern theory and methods of test validation meet this challenge with an expanded conception of what validity means and how it can be investigated.

Validity inquiry

SLA and LT researchers alike are concerned that the interpretations they make on the basis of test performance are well justified. With respect to vocabulary research, for example, Sharwood Smith has pointed out that "experimental data cannot of themselves inform us about the nature of the learner's current mental lexicon" (1984: 239). His point is that researchers make interpretations on the basis of their observed data, and as a result, the justification of their interpretations is crucial. For example, the sustainability of Meara's conclusion, mentioned earlier, about the nature of L2 vocabulary organization rests on the validity of his interpretation of performance on word association tests as a sign of vocabulary organization. The value of Young's findings, about participant influence on inflectional morpheme use rests on the quality of the observed performance as a sample obtained from the relevant contextual configuration. In both cases, inferences are made on the basis of performance in a setting that acts as an operational definition of the construct the researcher wants to assess. When the setting where performance consistency is observed is a test, the operational definition can be specified in terms of test method facets (Bachman 1990) or task characteristics (Bachman & Palmer 1996). This is illustrated in Figure 2.4, which shows that performance consistency is observed within a setting that often can be

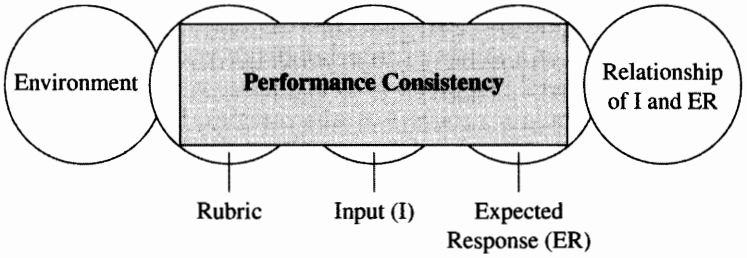

Figure 2.4

described by test method facets or task characteristics. (The test method or task characteristics are good descriptors for operational settings that are language tests or SLA elicitation devices, but it is not clear that they would be the best descriptors for other settings.) Sufficient justification of the interpretations made from test performance in an operational setting is needed so that tests can be used appropriately for decision making in educational contexts or for theory construction in research settings. The process of securing sufficient justification is validation.

Measurement researchers' conceptions of validation have evolved considerably throughout the 1980s and 1990s in ways that are relevant to the use of the interactionalist construct definition. Messick (1989) defines validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). "Empirical evidence and theoretical rationales" refer to justifications for test interpretation and use. Justifications for the interpretation and use of vocabulary tests include evidence about their construct validity and about their relevance and utility in a particular setting. In addition, justifications refer to the consequences resulting from test use. The types of construct validity evidence that can be used to justify test interpretation and use include such arguments as results of content analysis and correlational studies that support the hypothesis that test performance reflects the intended construct. Implications of an interactionalist definition for five types of construct validity evidence are discussed later in this chapter.

Another type of validity justification is the evidence pertaining to the relevance and utility of testing. Such evidence would demonstrate the usefulness of a test for achieving particular objectives in a given context. For example, in a setting where instructors need diagnostic information pointing to the needs of individual students, a vocabulary test that meets those needs from the instructors' and students' perspectives could be shown to have utility in that context. Such evidence, of course, would not stand alone in justifying test use but would need to be supported by construct

evidence indicating that the test measured the desired aspects of vocabulary. For example, in a study of interlanguage vocabulary (Singleton & Little 1991), learners' responses to C-test items were used to make inferences about vocabulary processes and strategies, but because no construct evidence supported such inferences, Chapelle (1994) suggested that the test had no utility for its research purpose – to inform theories of vocabulary organization, a vocabulary strategy, and the strategy's context of use. Because relevance and utility evidence are use-specific, and because there has been little work in this area, I will not discuss this facet of validity inquiry here.

Justification of testing through examination of a test's consequences requires the researcher to clarify the value implications underlying interpretation and use of a particular test as well as its impact on the test use context and beyond. Value implications are attached to the nature of the construct definition a test reflects. As explained earlier, each perspective on construct definition encompasses beliefs about what can and should be defined, how tests should be designed, and what the priorities for validation should be. These beliefs are formulated in socioacademic communities that implicitly support particular perspectives on construct definition through the use of particular types of tests. The objective of clarifying the value implications associated with particular test interpretations and uses is to assess consciously the values implied by the choice of a test for a particular purpose. Other consequences refer to the actions resulting from test use in a particular context and its unintended side effects beyond the immediate test use context.

For example, some tests that begin as research instruments eventually make their way to classrooms, where they provide implicit guidance to teachers about what to teach and to students about what to study. As an illustration, I will examine the value implications and social consequences of the trait-oriented vocabulary tests used in vocabulary research. I will explain why examination of values and consequences associated with a test must be informed by an understanding of the construct definition underlying the test. However, because all validity questions rest on an understanding of construct validity, I begin by explaining methods for investigating construct validity within an interactionalist perspective of construct definition.

Construct validity evidence

Construct validity evidence refers to the judgmental and empirical justifications supporting the inferences made from test scores. Because test performance is used to make inferences about a construct believed to be reflected in test performance, the nature of the researcher's construct definition is fundamental to the manner in which construct validity evidence

is produced and interpreted. The confusion resulting from disparate perspectives toward construct definition is apparent from Shohamy's (Chapter 7, this volume) interpretation of Tarone's (Chapter 3, this volume) finding that subjects performed differently on two versions of a speaking test. Tarone's point, consistent with behaviorist and interactionalist construct definitions, was that the two sets of data varied because they were samples from two different contexts, and therefore could not be interpreted as if they had been affected by the same set of factors. Shohamy, viewing the same results from a trait perspective (as signs of speaking ability), offered the following explanation: "One wonders if Tarone's conclusions can be interpreted as an indication of variation or of method effect" (p. 162). From the trait perspective, of course, "method effects" refer to error introduced into performance data resulting in inconsistent performance (i.e., variation) across test methods. From a behaviorist or interactionalist perspective, these method effects would not be error; they would be evidence of the expected influence of context on performance.

In short, rational interpretation of this and other evidence attempting to explain performance must be tied to an understanding of how the measured construct is defined. This understanding of the theoretical construct definition can then be used to evaluate the adequacy of the operational definition (i.e., the test) in eliciting the relevant performance as signs, samples, or both. The interactionalist perspective, which views the operational definition as eliciting both a sign of learner capacities and a context-constrained sample, is illustrated in Figure 2.5. The figure shows that performance consistency, which is attributed to both learner characteristics and aspects of context, is observed within an operational setting, which can be defined through the use of test method facets or task characteristics.

The process of construct validation requires evidence supporting the use of performance within the operational setting as an indicator of the defined construct. Five types of construct validity evidence are examined here: (1) content analysis, (2) empirical item investigation, (3) task analysis, (4) relationships between test scores and other measures, and (5) experimental research identifying performance differences over time, across groups and settings, and in response to experimental interventions.[7]

CONTENT ANALYSIS

Content analysis consists of experts' judgments of what they believe a test measures at the operational level (e.g., Carroll 1976; Bachman,

7 Another type of construct validity evidence, of course, demonstrates that observed test performance exhibits theoretically appropriate consistency.

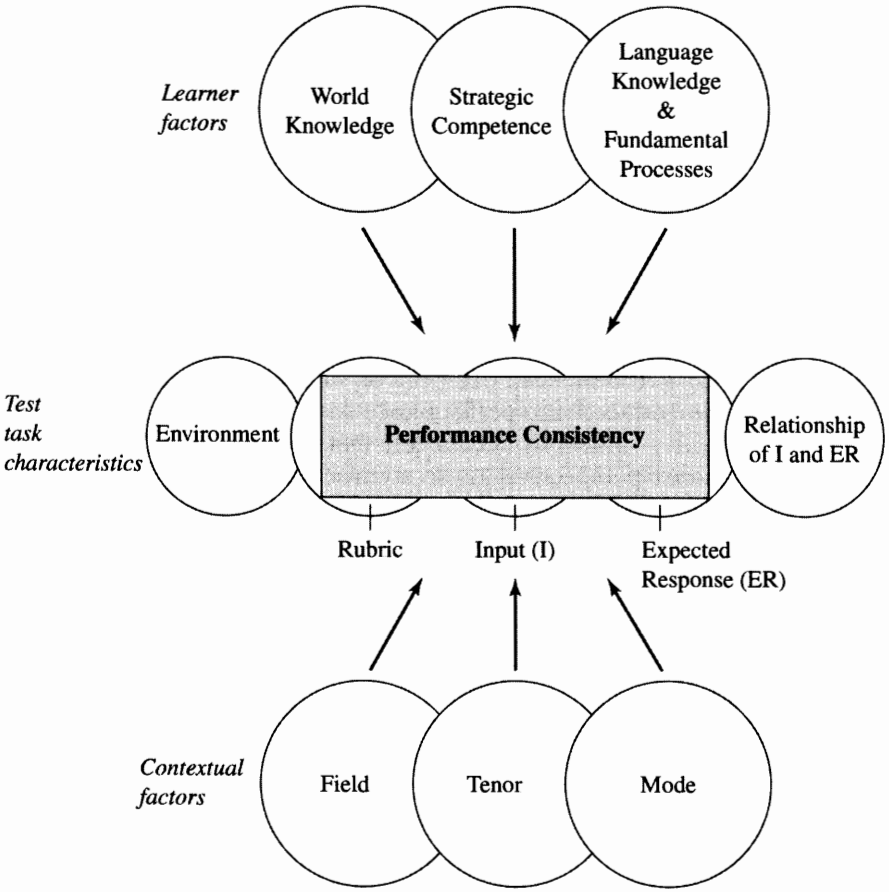

Figure 2.5

Kunnan, Vanniarajan, & Lynch 1988; Bachman, Davidson, Ryan, & Choi 1995). These judgments are then used to argue for or against the use of test performance as an indicator of the theoretically defined construct that the test is intended to measure as well as to inform empirical validity inquiry. With respect to tests of interlanguage vocabulary, for example, performance on word association tests has been judged as indicative of the trait "lexical organization" (Meara 1978, 1984). In Meara's judgment, word association tests "are a useful way of investigating the way a speaker's knowledge of his language is structured and stored" (Meara 1978: 208). It was on the basis of results from word association tests that Meara concluded that learners' mental lexicons were not semantically organized. Another test, the C-test, has also been judged a measure of vocabulary organization by other researchers (Singleton & Little 1991), but the re-

sults obtained from that test led the researchers to a different conclusion – that learners' lexicons were organized semantically. Particularly in view of the contradictory conclusions drawn from the use of the two different tests, a content analysis of each test based on the same interactionalist construct definition would be informative.

An interactionalist content analysis should hypothesize how the test operationalizes – during test taking – all of the theorized influences on performance. Such an analysis would use the test method facets to organize judgments about the operational constructions of the learner factors (language knowledge, fundamental processes, strategies, and world knowledge) and contextual factors (field, tenor, mode). For example, Douglas (Chapter 6, this volume) hypothesizes that the instructions (a test method facet) are associated with the test takers' conception of the discourse domain (the operational level the test taker creates for values of field, tenor, and mode). The discourse domain then influences the test taker's goal setting and planning (the operational level of strategic competence), which in turn – along with other test method facets – influence the language knowledge employed (the operational level of language knowledge).

Because the test taker encounters test method facets sequentially during test taking (i.e., first the environment is perceived, then the instructions are read or heard, then the input is encountered . . .), Chapelle and Douglas (1993) have suggested that content analysis should also be process-oriented, allowing for judgments about the nature of the operational setting as it is constructed during test taking. Indeed, such an analysis becomes very complex. As Palmberg (1987) put it in his discussion of learners' performance on a vocabulary test used in classroom research, "it is hazardous to speculate about the factors that made the pupils think of the very words that they actually produced [on the test]" (Palmberg, 1987: 209). Although it may be hazardous, theory-based speculation about the nature of the operational test performance is essential if a content analysis is to provide results that can be compared to a theoretical interactionalist construct definition. Such a comparison is a crucial source of construct validity evidence.

EMPIRICAL ITEM INVESTIGATION

Empirical item investigation, or identification of factors affecting item difficulty and discrimination (Carroll, 1989), provides statistical evidence relevant to researchers' understanding of the operational level of the construct definition. In vocabulary testing, for example, Perkins and Linnville (1987) investigated the operational construct definition of a multiple choice vocabulary test by assessing the extent to which item difficulty could be attributed to test method facets, including the characteristics of the stimulus words (e.g., frequency, abstractness) and the keyed responses (e.g., number of synonyms for each response). Their findings that the combinations of item characteristics predicting item difficulty

depended on the level of subjects' proficiency led them to conclude that the operational construct meaning of the test varied as a function of proficiency level – evidence useful in the interpretation of scores for learners at different levels. If we were to interpret such item analysis results with reference to the theoretical construct the test is intended to measure, however, the operational variables to be investigated could be chosen on the basis of their hypothesized reflection of aspects of the theoretical definition.

From an interactionalist perspective, the test method facets to investigate would be those reflecting the influences of specific learner factors (language knowledge, fundamental processes, strategies, and world knowledge) and contextual factors (field, tenor, and mode). For example, one predictor of difficulty in a vocabulary test might be frequency of the lexical item (as Perkins and Linnville hypothesized); however, that variable defined from an interactionalist perspective would have to refer to frequency within a particular context, and the results would be interpreted relative to that context. Another test method facet that has been investigated in this manner in cloze tests is the distance from the blank of the clue needed for supplying the correct lexical item. The distance variable might be chosen to represent the metacognitive strategies required for figuring out what goes in each blank; the prediction is that the farther the clue is from the blank, the more difficult the item. Results are interpreted to indicate that the predicted metacognitive strategies do indeed come into play in the operational definition. The construct validity question is whether or not these similar strategies are also included in the theoretical construct definition of what the test is supposed to measure.

This type of validity inquiry places difficult demands on construct definition because it requires predictions linking theoretical sources and levels of difficulty with empirical item difficulty. Such a theory of item difficulty (Campbell 1961; Carroll 1989; Abraham & Chapelle 1992) can best be informed by a construct definition with components specified in terms of their levels of development. For example, the trait or interactionalist definition informed by SLA research would specify that idiomatic senses of words would be acquired later than literal senses (Kellerman 1978). This developmental assertion predicts that a test response requiring knowledge of an idiomatic sense will be more difficult than responses requiring knowledge of literal senses. Accordingly, item analysis research might choose the item characteristic "literal/idiomatic" as a predictor of difficulty. Similarly, native speakers' acquisition of word meaning has been found to be a gradual process (Marshalek 1981) so items requiring vague knowledge of meaning (e.g., recognition items) are hypothesized to be easier than those requiring precise knowledge of meaning (e.g., items requiring test takers to produce definitions). Work is needed to hypothesize the developmental dimensions of other aspects

of the interactionalist construct definition to make it more applicable to our understanding of operational definitions through the study of item difficulty.

EMPIRICAL TASK ANALYSIS

As Cohen (Chapter 4, this volume) explains, empirical task analysis attempts to document the metacognitive strategies that learners use as they complete test tasks. The primary methodology is qualitative, probing the metacognitive strategies employed in the operational setting. The operational performance is then compared to an interactionalist construct definition at the theoretical level. For example, Feldman and Stemmer (1987) documented the test-taking strategies used in an operational setting to complete the lexical items deleted from a C-test (see also Stemmer 1991). They found that a wide variety of metacognitive plans were executed to complete blanks when lexical items were not retrieved automatically. The results of such research, suggesting that the plans (and therefore knowledge) used to arrive at responses vary across items and learners, present a challenge for construct validity inquiry. Messick describes the problem as follows:

> The notion that a test score [or profile] reflects a single uniform construct interpretation or that validation should seek to defend (or challenge) a single test-construct match becomes illusory. Indeed, that a test's construct interpretation might need to vary from one type of person to another (or from one setting or occasion to another) is a major current conundrum in educational and psychological measurement. It suggests that the purview of construct validation should include delineation of the alternative construct meanings of test scores. (Messick 1989: 55)

These multiple construct meanings implied by the idiosyncratic metacognitive strategies used in an operational setting are the precise target of empirical task analysis.

The kind of information provided by empirical task analysis makes it uniquely suited to investigating the operational settings associated with interactionalist construct definitions. For example, Dollerup et al. (1989) present a definition of "vocabularies in the reading process" that includes "(a) a 'word knowledge store,' (b) strategies for decoding words, and (c) the special linguistic context. It implies that individual vocabularies in reading exist instantaneously, and that they are, in effect, fluid entities which change every time they are generated by the reading of specific texts. *Vocabularies differ not only in time but also from text to text with the same reader*" (pp. 30–31; emphasis in original). Investigation of a test associated with this interactionalist conception of vocabulary would have to probe the strategies individuals used in the operational setting of the test. Findings would address the conundrum of individual differences in score meaning.

CORRELATIONS WITH OTHER TESTS AND BEHAVIORS

Similarities among different operational settings can be identified by calculating correlations among sets of scores obtained in operational test (and other) settings. Insofar as the researcher understands the connections between operational tests and their underlying theoretical construct definitions, such correlations can provide a source of validity evidence. The need to use other methods of validity inquiry to hypothesize the bridge between operational and theoretical levels was voiced by early educational measurement researchers (e.g., Gulliksen 1950) and has been echoed more recently by LT researchers (e.g., Vollmer 1983; Grotjahn 1986; Bachman 1990; Alderson 1993). Grotjahn expresses the problem of using correlational evidence at the operational level as follows:

> Construct validation of a language test with the help of other language tests presupposes the construct validity of these tests, which is normally at best only partially established, and if at all, then only with the help of correlational analysis. The potential circularity of this approach should be obvious. (Grotjahn 1986: 161)

The operational circle can be broken when the theoretical constructs underlying performance on tests inform the design and interpretation of correlational research. The best-known paradigm for systematizing theoretical predictions of correlations obtained at the operational level is the multitrait-multimethod (MTMM) research design (Campbell & Fiske 1959). To conduct research within this paradigm, as some language testing researchers have done (e.g., Bachman & Palmer 1982; Arnaud 1989; Swain 1990), the researcher administers several tests – each intended to measure a language trait. Tests must be chosen intentionally so that each trait is measured using several different methods. Observed variance attributed to the context (which is realized by the test methods) is viewed as evidence against the test's validity. In other words, context (or method) variance is viewed as measurement error. Observed variance attributed to the trait that the test is intended to measure is considered evidence for the test's validity. Trait variance is good. The researcher has found evidence for the construct validity of the tests if the correlations among tests of the same trait are higher than are correlations among tests of different traits or among tests using the same method of measurement. In other words, this research design, consistent with the trait theorist's measurement principles, treats method effects as error, predicting that ideally strong covariance should occur only as the result of similar traits underlying test performance.

The MTMM research design offers valuable perspectives for the use of correlational methods in validation research; however, to work within the interactionalist construct perspective, researchers need to reconceptualize the "trait as good variance" and "method as systematic error vari-

ance" as more general notions – "construct-relevant variance" and "construct-irrelevant variance," respectively. According to the interactionalist definition (as illustrated in Figure 2.3), variance should be contributed by both test-taker and context factors. To the extent that the test method facets play a role in operationalizing both of these aspects of the theoretical definition, they cannot be treated as error. Instead of assuming that only trait variance is construct relevant and that all method variance is irrelevant, the researcher working within an interactionalist construct definition is obligated to theorize exactly what should be considered relevant sources of variance and what should be considered irrelevant.

Moreover, because the interactionalist perspective hypothesizes multiple contributors to observed performance consistency, the researcher must consider the expected proportions of relevant variance that should be shared with various types of tests. For example, Chapelle and Green (1992) have suggested the need to hypothesize the expected shared variance between language and nonlanguage tests (e.g., Marshalek 1981; Chapelle & Abraham 1990) in order to interpret observed correlations. Such hypotheses must rest on an understanding of both theoretical and operational construct meaning of all the tests involved because interpreting such correlations requires a construct theory that allows a researcher to hypothesize a "network of relationships of a test to other measures" (Embretson 1983: 180). The general interactionalist construct definition moves in this direction by including all of the factors believed to affect performance; however, it fails to specify the predicted relative strengths of influence from any of the factors. This additional specification would be needed to make predictions about the strengths of observed correlations. Correlational research that systematically selects operational definitions of various aspects of the interactional construct definition can help to develop an understanding of the general construct definition.

EXPERIMENTAL STUDIES OF TEST PERFORMANCE

Experimental manipulations of subjects or test methods enable the researcher to examine hypotheses about test performance by systematically modifying test conditions to verify that observed performance behaves in consort with theory-based predictions. This type of research can investigate differences in test performance between native speakers and nonnative speakers, between learners before and after instruction, or between the same learner's performance on different forms of a test. The latter type of research can be most useful when it is informed by the interactionalist construct definition. The interactionalist definition provides a framework for hypothesizing aspects of the theoretical definition as they emerge within the operational setting. In a study comparing a general and a field-specific speaking test, for example, Douglas and Selinker (1992) demonstrated the use of the interactionalist perspective in this type of research:

> It is our view that testing language for specific purposes involves more than just changing content; specific purpose testing requires a change in discourse domain (Douglas & Selinker 1985), which involves the language user's assessment of the communicative situation and her subsequent planning of a linguistic response to the situation [i.e., the operational level of strategic competence] . . . This change in discourse domain is brought about in turn by contextualization cues, culturally conventional, highly redundant language signals such as voice tone, pitch, tempo, rhythm, code, topic, style, posture, gaze and facial expression, that interactants attend to in assessing the communicative situation (Gumperz 1976). (Douglas & Selinker, 1992: 318).

In an operational setting, these contextual features (which fall within the field, tenor, and mode constructs) are identified through test method facets and their interpretation by the test taker. In other words, in a given test the features of the input (a test method facet) would specify particular values for cues such as voice tone (an aspect of tenor), and a particular test taker would interpret those values to create a discourse domain during test taking.

An important question to be investigated in validity research is whether changes in test method facets can produce the theorized changes in all aspects of the operational setting, including the test takers' construction of context and use of strategic competence. As Douglas (Chapter 6, this volume) points out, an understanding of the role of the test method facets in creating an operational setting is essential to progress within the interactionalist perspective:

> I have argued that we need to *capitalize* on [method effects] by designing tests for specific populations – tests that contain instructions, content, genre, and language [i.e., test method facets] directed toward that population. The goal is to produce tests, useful to both LT and SLA professionals, that would provide information interpretable as evidence of communicative competence in context. (p. 153)

Research investigating the role of subtle differences in test method facets as operational instruments of theoretical learner and context factors will provide an essential foundation for work seeking to realize this objective.

Each of these research methods used for investigating construct validity offers a unique type of evidence pertaining to the meaning of consistent test performance. In each case, however, a rational approach to designing and interpreting the research rests on a clearly articulated construct definition. The collective possibilities for construct validation may appear overwhelming to researchers wishing to use tests rather than endlessly investigating their construct validity. As Shepard (1993) expresses the problem, "If construct validity is seen as an exhaustive process that can be accomplished over a 50-year period, test developers may be inclined to think that any validity information is good enough in the short run" (p. 444). The alternative Shepard suggests is to focus on those questions

essential to the consequences of testing. For example, Perkins and Linnville's (1987) object of investigation was a test used for placement at different levels; therefore, their most immediate concern was the consistency of the operational construct definition across different levels, causing them to choose to investigate validity through a study of empirical item difficulty. If testing consequences are to direct validity inquiry, however, it will be necessary to identify the consequences of tests used in SLA research.

Consequences of test use

According to Messick's (1989) definition of validity, the consequences of testing serve as additional justifications that can build upon our understanding of score meaning by identifying the value and sociological implications of testing. In particular, Messick identifies two interrelated questions that should underlie investigation of testing consequences: What are the value implications of the interpretations made from testing? What are the social consequences of test use? These difficult questions are just beginning to be explored by LT researchers. Addressing educational uses of tests, Alderson and Wall (1993) point out that specific consequences of language tests have not been clearly documented, and therefore an essential aspect of consequential validity inquiry is to hypothesize the nature and scope of test consequences. Addressing test consequences from the perspective of educational and psychological measurement in general, Shepard suggests that "the validity framework appropriate to [an educational] test use will have a different focus than one for the same measure used to operationalize a construct in a research setting" (1993: 445). In applied linguistics research, however, the distance between laboratory and classroom is often negligible; the constructs developed by researchers often end up affecting the classroom. I will therefore consider research on testing to affect both L2 research and classrooms.

I will illustrate the testing impacts relevant to validity by discussing the apparent consequences of two test methods associated with the trait-oriented perspective of interlanguage vocabulary: the word association test and the Y/N vocabulary recognition test. The former method has been used in a number of different forms, but the basic principle is that the test taker is required to produce one or more words that come to mind each time the experimenter presents a word. Meara (1978, 1984) believes that the word association test assesses the trait dimension "vocabulary organization." The Y/N test requires learners to indicate whether or not they think they recognize words that are presented to them in isolation. Test items are composed of both actual target language words and sequences of letters that are not real words in the language. Meara believes that this test measures the trait "absolute vocabulary size" (Meara & Buxton 1987;

Meara 1989). Both tests, consistent with their underlying construct definitions, present vocabulary items out of context and choose items without regard to the context in which they might occur in language use.

The validity of interpretations and uses of these tests should be examined by considering evidence for their construct validity with reference to the construct each is intended to measure, their relevance and utility in a particular setting, and their consequences. Here, however, I will consider only the consequences in terms of value implications and their societal consequences.

SOCIETAL IMPLICATIONS OF INTERLANGUAGE VOCABULARY TESTING

In SLA research, the tests used to elicit language samples are typically viewed as instruments acting as one piece of a larger research design. As Loevinger (1957) pointed out many years ago, however, the tests used in research help to define researchers' views of the constructs they investigate. The use of a given instrument is tied to a particular perspective toward construct definition, which is built upon assumptions and values about constructs and their measurement. Tests and their associated construct definitions, in turn, affect the nature of the theories that L2 researchers develop. One purpose of validity inquiry, therefore, is to clarify the values underlying the interpretations made from a given test in order to compare these implicit test-embedded values with the researcher's explicit values and with those of others in the field. I explain here how identifying the nature of the construct definition underlying a particular test is essential to an investigation of the values underlying test interpretation. In particular, I will speculate on two value-related implications of the trait-oriented word association and Y/N test: their impacts on the perceptions of researchers and on the nature of SLA theory.

The word association test and its underlying trait definition appear to have influenced the perceptions of some interlanguage vocabulary researchers. Since Meara's 1978 study, other researchers have continued to use varieties of word association tests in L2 lexicon research (e.g., Palmberg 1987; Soderman 1989) and to discuss "vocabulary organization" from a trait perspective (e.g., Ard & Gass 1987; Gass 1988; Singleton & Little 1991). In fact, after finding that a word association test was inadequate as a proficiency measure, Kruse, Pankhurst, and Sharwood Smith (1987) concluded that word association tests "may conceivably be a useful tool in future explorations of lexical interlanguage. A study of the associative responses given by learners might still contribute to the discovery of developmental processes that take place in language learners' mental lexicons" (p. 153). On the basis of this recommendation and the lack of additional insight into the validity of word association tests for making inferences about lexicon organization, future researchers may

indeed continue to investigate the trait conceptions of vocabulary using word association tests.

Meara, on the other hand, later advocated the Y/N test, which is intended to measure the trait dimension "absolute vocabulary size," but he also suggests it for a proficiency measure.

> The prototype version of the test [estimates vocabulary size] in 10 minutes, scores itself automatically and produces very high correlations with an extended test of overall ability in EFL which normally requires an hour and a half to administer. We think this is a significant development, particularly for research purposes, since it will allow researchers to assess the proficiency level of their subjects in a very simple way and help us to get away from the messy and unreliable labels which characterize much current research (Meara & Buxton 1987: 150–151)

It is not yet clear whether or not other researchers will adopt the Y/N test for measuring vocabulary size (or for proficiency). What is apparent is that the trait-oriented construct definitions underlying both of these tests must be recognized and their impacts evaluated.

A validity study investigating the values underlying the use of a word association test or the Y/N test would first identify the trait-oriented construct definition underlying the tests and then examine the impact of this trait definition on the test use context. Because the word association test and the Y/N test are believed to affect perceptions of interlanguage vocabulary researchers, consistency between the field's conception of interlanguage and the construct definition implied by a test should be examined. These tests apparently play a role in maintaining a trait perspective of interlanguage vocabulary despite the fact that the interest in applied linguistics has shifted to an interactionalist perspective. One might argue against the use of such tests in research on the basis of this discontinuity.

A second value-related consequence of tests used in vocabulary research is their impact on the nature of theory about the acquisition of L2 vocabulary. In the case of the Y/N test, the "neat and reliable" underlying trait definition extends into a trait-oriented *theory of L2 vocabulary development.* Meara's theory of vocabulary development hypothesizes that "the underlying structure [of vocabulary acquisition] can be seen as a transitional probability matrix" (Meara 1989: 73). The research associated with this theory would focus on prediction and observation of increases in vocabulary size without reference to other aspects of vocabulary knowledge, strategies, or contexts of acquisition. Moreover, Meara recommends the Y/N test and its associated research agenda to other vocabulary researchers:

> It is unlikely that such study will get very far, unless we adopt some common standards which might help to make diverse [research] programs relatively compatible. These are: 1) adoption of a week as the basic unit of time;

2) adoption of the T2-T3 transition matrix as the standard datum; 3) adoption of a neutral, minimal vocabulary assessment instrument, based in Meara and Buxton (1987). (Meara, 1989: 73).

Other researchers working within the trait perspective have suggested that other dimensions of the trait also be investigated. For example, Wesche and Paribakht (1993) described the state of research on L2 lexical development:

A problem facing empirical researchers is the lack of sensitive assessment procedures for tracing the development over time of specific vocabulary knowledge, either in terms of stages characterizing how well given words are known, or in terms of the different kinds of knowledge it is possible to have about given words. (Wesche and Paribakht, 1993: 2)

This observation addresses the need to investigate L2 lexical development using methods other than those advocated by Meara in order to construct a theory of L2 vocabulary development that extends beyond a quantitative theory of expanding vocabulary size. However, in doing so, Wesche and Paribakht propose to expand research to other dimensions of the trait rather than to reconceptualize the construct to include strategies and contexts of vocabulary use. A validity study of the word association or Y/N test should investigate their impact on a theory of interlanguage vocabulary development. Having identified the one-dimensional, trait-oriented definition underlying each test, one might make a consequential validity argument against the use of word association and Y/N tests because they have helped to focus theory on one trait dimension while other researchers document the need to understand multiple dimensions of trait. Moreover, as Tarone (Chapter 3, this volume) points out, tests that reflect a trait definition confine the focus of theory to dimensions of knowledge and fundamental processes, causing researchers to overlook the roles of contexts and strategies in performance and acquisition. The perceptions of interlanguage researchers and the nature of the acquisition theory they construct are two of the value implications of the inferences made from the tests chosen for vocabulary research. As I have suggested, however, the impacts of these tests are felt beyond the research community, extending into educational practices, where their social consequences must also be evaluated.

SOCIAL CONSEQUENCES OF INTERLANGUAGE VOCABULARY TESTS

Because much work in applied linguistics is aimed at addressing real language problems and needs, the impact of tests used in applied linguistics research should be expected to extend into classroom settings. Direct impacts might be seen when an SLA study is conducted in a classroom by administering tests to students in their language class. Impacts of research

in the classroom also occur when a research test such as the word association test is investigated as a potential proficiency measure for educational use. Kruse et al. (1987) were "concerned with the practical research issue of whether the word association test can indeed be used as a reliable measure of language proficiency" (p. 145). Despite their conclusion that the word association test would not make a good proficiency measure (and might therefore be suitable only for L2 vocabulary research), their study illustrates the blurred distinction between research and classroom in applied linguistics and therefore the potential for today's research tests to appear in tomorrow's classrooms. Even if research tests are not adopted in L2 classrooms, they are likely to influence L2 teaching practices indirectly. For example, from his first reported study of the use of word association tests for vocabulary research, Meara attempted an extension to the classroom:

> The production of word associations is not so clearly related to ordinary language activities. My own feeling, however, is that all the various types of language activity are reflections of the same underlying, basic skills, and that if we could develop learning methods that, as a side-effect produced learners with native-like association patterns, we would also be producing learners who were better able to communicate in their foreign language. (Meara 1978: 210–211)

The word association test's successor, the Y/N test, is now claimed to be ideal for educational uses. Meara and Buxton (1987) and Meara (1989) claim the usefulness of the Y/N test despite any apparent limitations:

> There are a number of problems with [Y/N vocabulary tests] – notably that they measure passive vocabulary, rather than active vocabulary skills. However, they do have some important practical advantages – ease of construction, simplicity of assessment, time necessary for completion, possibility of large sample rates, and so on – *which seem to outweigh most of the theoretical disadvantages.* (Meara 1989: 72; emphasis added)

According to our current conception of validity, practical advantages do not outweigh "theoretical" disadvantages, such as a test's impact on learners' perceptions of what they should study. One might argue that a Y/N recognition test encourages students to study dictionaries and especially spelling, to the neglect of the interactional abilities that are no less essential for communicative development. Investigation of the social consequences of tests would hypothesize and document these types of testing impacts on students' and instructors' perceptions of language and on their classroom actions (Canale 1987).

Despite the importance of consequences as a facet of validity inquiry, their investigation as a credible mode of validity inquiry has not yet been well developed by educational measurement researchers. The first step in investigating test consequences is to identify the potential intended and unintended consequences of test use. The word association and Y/N tests

provide good examples of research tests in applied linguistics with consequences extending beyond a particular research setting. As I suggested earlier, they may affect the perceptions of interlanguage researchers, interlanguage theory, and classroom practices. In examining the values and consequences of these tests, an understanding of their underlying construct definitions was essential. Educational measurement researchers will continue to refine their definition of – and research methods for studying – testing consequences. As this work continues, applied linguists, whose research concerns are situated in classrooms and other contexts of language use, are ideally suited to develop the precedents for consequential validity inquiry.

Implications for SLA and LT research

Current conceptions of construct definition and validity inquiry offer directions and challenges for SLA and LT research. Tests used for both educational and research purposes need to be subjected to the processes of validity inquiry to reveal the quality of any given operational setting for producing the relevant signs and samples of learners' performance. The validation process is ongoing; varieties of justifications pertaining to appropriate test use are continually revisited. This means that there is no "validated test" – one that has been proved valid for all time, and therefore can be picked up and used without further examination. Instead, researchers are at all times responsible for examining construct validity, relevance and utility, value implications, and social consequences. For example, when Kruse et al. (1987) investigated the validity of using a word association test for assessing language proficiency, in addition to the two traditional construct validity methods they used (correlations of the word association test with other language tests and comparison of performance by native and nonnative speakers), other aspects of validity could have also been discussed (e.g., how the use of word associations as a proficiency measure would affect learners' perceptions of proficiency). The eventual impacts of tests in applied linguistics research may even help to prioritize investigation of construct validity inquiry.

Construct definition plays an essential role in rational design and interpretation of validation research. Despite the support in applied linguistics for an interactionalist approach to construct definition, little systematic inquiry has been attempted to better define specific person and context dimensions. Validity researchers, therefore, are challenged to better understand the nature and implications of an interactionalist construct definition because it offers unique dilemmas for each method of construct validity inquiry. Content analysis requires that the person and context sources of learners' performance in the operational setting be hypothe-

sized, but the analytic procedures for making such process-oriented hypotheses have not been developed. Empirical item investigation demands selection of operational variables believed to represent context and person as well as their interaction as sources of difficulty, but the developmental definitions required to make such selections have not been specified. Task analysis investigating the strategies used in operational settings forces all researchers to recognize what Messick (1989) terms the "conundrum of educational measurement" – that strategies can vary across people and tasks even when the same results are achieved. Correlational studies require hypotheses about the strengths of correlations expected among language tests and other tests, but such hypotheses rest on unknown degrees of contribution from the multiple relevant sources of performance consistency. Experimental studies require specification of test method facets that will aid in better understanding the role of theoretical factors in the operational setting, but little theoretical work has addressed questions pertaining to the connection between the two levels of measurement.

An essential step toward substantive progress in these areas is a clear distinction between the measurement data sought from SLA elicitation devices and those from other methods of observation in SLA research. Only when performance consistency is observed do researchers move from fragile, idiosyncratic observations to measurement. As a consequence, estimates of reliability are essential to understanding when data can be summarized to act as dependable indicators of constructs and when construct validity evidence is needed to support construct-related inferences. Although inconsistent performance is interesting in its own right, it cannot be treated as a dependable indicator of a construct. To make progress in understanding the interactionalist construct definition, it is necessary to better understand the nature of operational settings across which consistent performance can be observed.

Teichroew's review of vocabulary research concluded that "it does seem clear that the form of the test has considerable influence on the results gleaned" (1982: 22). Over fifteen years later we can add that the nature of the test also has considerable influence on our implicit construct definitions, which in turn affect the perceptions of researchers, the nature of SLA theory, and classroom practices. As a consequence, validity inquiry informed by current interactionalist construct definition is essential.

References

Abraham, R. G., & C. A. Chapelle. 1992. The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal, 76,* 468–479.

Aitchison, J. 1987. *Words in the mind: An introduction to the mental lexicon.* New York: Blackwell.

Alderson, J. C. 1993. The relationship between grammar and reading in an EAP test battery. In D. Douglas & C. Chapelle (eds.), *A new decade of language testing research* (pp. 203–219). Arlington, VA: TESOL.

Alderson, J. C., & D. Wall. 1993. Does washback exist? *Applied Linguistics, 14,* 115–129.

Ard, J., & S. Gass. 1987. Lexical constraints on syntactic acquisition. *Studies in Second Language Acquisition, 9,* 233–351.

Arnaud, P. J. L. 1989. Vocabulary and grammar: A multitrait-multimethod investigation. *AILA Review, 6,* 56–65.

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

Bachman, L. F., F. Davidson, & J. Foulkes. 1993. A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL test batteries. In D. Douglas & C. Chapelle (eds.), *A new decade of language testing research* (pp. 25–45). Arlington, VA: TESOL.

Bachman, L. F., F. Davidson, K. Ryan, & I. Choi. 1995. *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study.* Cambridge: University of Cambridge Local Examinations Syndicate / Cambridge University Press.

Bachman, L. F., A. Kunnan, S.Vanniarajan, & B. Lynch. 1988. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency tests. *Language Testing, 5*(2), 128–159.

Bachman, L. F., & A. S. Palmer. 1982. The construct validation of some components of communicative competence. *TESOL Quarterly, 16*(4), 449–465.

Bachman, L. F., & A. S. Palmer. 1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bialystok, E. 1981. The role of linguistic knowledge in second language use. *Studies in Second Language Acquisition, 4,* 31–45.

Bialystok, E., & M. Sharwood Smith. 1985. Interlanguage is not a state of mind: An evaluation of the construct for second language acquisition. *Applied Linguistics, 6,* 101–117.

Blum-Kulka, S., & E. Levinson. 1983. Universals of lexical simplification. In C. Faerch & G. Kasper (eds.), *Strategies in interlanguage communication* (pp. 119–139). London: Longman.

Campbell, A. 1961. Some determinants of the difficulty of non-verbal classification items. *Educational and Psychological Measurement, 21,* 899–913.

Campbell, D. T., & D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Canale, M. 1987. Language assessment: The method is the message. In D. Tannen & J. E. Alatis (eds.), *The interdependence of theory, data, and application* (pp. 249–262). Washington, DC: Georgetown University Press.

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

Carroll, J. B. 1976. Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (ed.), *The nature of intelligence* (pp. 27–56). Hillsdale, NJ: Lawrence Erlbaum.

Carroll, J. B. 1989. Intellectual abilities and aptitudes. In A. Lesgold & R. Glaser

(eds.), *Foundations for a psychology of education* (pp. 137–197). Hillsdale, NJ: Lawrence Erlbaum.

Carroll, J. B. 1993. *Human cognitive abilities: A survey of factor analytic studies.* Cambridge: Cambridge University Press.

Chapelle, C. A. 1994. Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*(2), 157–187.

Chapelle, C. A., & R. G. Abraham. 1990. Cloze method: What difference does it make? *Language Testing, 7*(1), 121–146.

Chapelle, C. A., & D. Douglas. 1993. Interpreting L2 performance data. Paper presented at the second Language Research Colloquium, Pittsburgh, PA, March.

Chapelle, C. A., & P. Green. 1992. Field independence/dependence in second language acquisition research. *Language Learning, 42,* 47–83.

Chitiri, H-F., Y. Sun, D. M. Willows, & I. Taylor. 1992. Word recognition in second language reading. In R. J. Harris (ed.), *Cognitive processing in bilinguals* (pp. 283–297). New York: Elsevier.

Cohen, A. 1984. On taking language tests: What the students report. *Language Testing, 1*(1), 70–81.

Cohen, A., & M. Robbins. 1976. Toward assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and the learners' explanations. *Language Learning, 26,* 45–66.

Dollerup, C., E. Glahn, & C. R. Hansen. 1989. Vocabularies in the reading process. *AILA Review, 6,* 21–33.

Douglas, D., & L. Selinker. 1985. Principles for language tests within the "discourse domain" theory of interlanguage: Research, test construction and interpretation. *Language Testing, 2,* 205–226.

Douglas, D., & L. Selinker. 1992. Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20*(3), 317–328.

Duff, P. 1993. Tasks and interlanguage performance: An SLA perspective. In G. Crookes & S. M. Gass (eds.), *Tasks and language learning: Integrating theory and practice* (pp. 57–95). Philadelphia: Multilingual Matters.

Eckman, F. R. 1994. The competence-performance issue in second-language acquisition theory: A debate. In E. E. Tarone, S. M. Gass, & A. D. Cohen (eds.), *Research methodology in second-language acquisition* (pp. 3–15). Hillsdale, NJ: Lawrence Erlbaum.

Ellis, R. 1989. Sources of intra-learner variability in language use and their relationship to second language acquisition. In S. Gass, C. Madden, D. Preston, & L. Selinker (eds.), *Variation in second language acquisition. Psycholinguistic issues* (Vol. 2, pp. 22–45). Philadelphia: Multilingual Matters.

Embretson, S. 1983. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197.

Faerch, C., & G. Kasper 1983. Plans and strategies in foreign language communication. In C. Faerch & G. Kasper (eds.), *Strategies in interlanguage communication* (pp. 20–60). London: Longman.

Faerch, C., & G. Kasper (eds.). 1987. *Introspection in second language learning.* Clevedon: Multilingual Matters.

Feldmann, U., & B. Stemmer. 1987. Thin_____ aloud a_____ retrospective da_____ in C-te_____ taking: Diffe_____ languages-diff_____ learners-sa_____ approaches? In C. Faerch and G. Kasper (eds.), *Introspection in*

second language research (pp. 251–267). Philadelphia: Multilingual Matters.

Frenck-Mestre, C., & J. Vaid. 1992. Language as a factor in the identification of ordinary words and number words. In R. J. Harris (ed.), *Cognitive processing in bilinguals* (pp. 265–282). New York: Elsevier.

Gass, S. 1988. Integrating research areas: A framework for second language studies. *Applied Linguistics, 9*(1), 198–217.

Grotjahn, R. 1986. Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 3*(2), 159–185.

Gulliksen, H. 1950. Intrinsic validity. *American Psychologist, 5,* 511–517.

Gumperz, J. J. 1976. Language, communication and public negotiation. In P. R. Sanday (ed.), *Anthropology and the public interest* (pp. 273–292). New York: Academic Press.

Halliday, M. A. K., & R. Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective.* Oxford: Oxford University Press.

Henderson, L. 1985. Toward a psychology of morphemes. In A. W. Ellis (ed.), *Progress in the psychology of language* (Vol. 1, pp. 15–72). London: Lawrence Erlbaum.

Hymes, D. 1967. Models of the interaction of language and the social setting. *Journal of Social Issues, 23*(2), 8–28.

Hymes, D. 1972. *Towards communicative competence.* Philadelphia: Pennsylvania University Press.

Kellerman, E. 1978. Giving learners a break: Native speaker intuitions as a source of predictions about transferability. *Working Papers on Bilingualism, 15,* 59–92.

Kruse, H., J. Pankhurst, & M. Sharwood Smith. 1987. A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9,* 141–154.

Lantolf, J. P., & W. Frawley. 1988. Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10,* 181–195.

Laufer, B. 1990. "Sequence" and "order" in the development of L2 lexis: Some evidence from lexical confusions. *Applied Linguistics, 11*(3), 281–296.

Loevinger, J. 1957. Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Luppescu, S., & R. R. Day. 1993. Reading, dictionaries, and vocabulary learning. *Language Learning, 43*(2), 263–287.

Marshalek, B. 1981. Trait and process aspects of vocabulary knowledge and verbal ability. Technical Report No. 15, Aptitude Research Project. AD A102757. Stanford, CA: School of Education, Stanford University.

Marslen-Wilson, W. 1989. Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (ed.), *Lexical representation and process* (pp. 3–24). Cambridge, MA: MIT Press.

McLaughlin, B. 1990. Restructuring. *Applied Linguistics, 11*(2), 113–128.

Meara, P. 1978. Learners' word associations in French. *Interlanguage Studies Bulletin, 3,* 192–211.

Meara, P. 1984. The study of lexis in interlanguage. In A. Davies, C. Criper, & A. P. R. Howatt (eds.), *Interlanguage* (pp. 225–239). Edinburgh: Edinburgh University Press.

Meara, P. 1989. Matrix models of vocabulary acquisition. *AILA Review, 6,* 66–74.

Meara, P., & B. Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing, 3,* 142–154.
Messick, S. 1981. Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89,* 575–588.
Messick, S. 1989. Validity. In R. L. Linn (ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Nation, P. 1993. Using dictionaries to estimate vocabulary size: Essential but rarely followed procedures. *Language Testing, 9,* 27–40.
Nattinger, J., & J. DeCarrico. 1989. Lexical phrases, speech acts, and teaching conversation. *AILA Review, 6,* 118–139.
Olshtain, E. 1987. The acquisition of new word formation processes in second language acquisition. *Studies in Second Language Acquisition, 9,* 221–232.
Palmberg, R. 1987. Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition, 9,* 201–220.
Perkins, K., & S. E. Linnville. 1987. A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*(2), 126–141.
Sharwood Smith, M. 1984. Discussant: The study of lexis in interlanguage by P. Meara. In A. Davies, C. Criper, & A. P. R. Howatt (eds.), *Interlanguage* (pp. 236–239). Edinburgh: Edinburgh University Press.
Shepard, L. 1993. Evaluating test validity. *Review of Research in Education, 19,* 405–450.
Singleton, D., & D. Little 1991. The second language lexicon: Some evidence from university-level learners of French and German. *Second Language Research, 7,* 62–81.
Skehan, P. 1987. Variability and language testing. In R. Ellis (ed.), *Second language acquisition in context* (pp. 195–206). Englewood Cliffs, NJ: Prentice-Hall.
Soderman, T. 1989. Word associations of foreign language learners and native speakers – a shift in response type and its relevance for a theory of lexical development. *Scandinavian Working Papers on Bilingualism, 8,* 114–121.
Spolsky, B. 1985. The limits of authenticity in language testing. *Language Testing, 2,* 29–40.
Stemmer, B. 1991. What's on a C-test taker's mind? Mental processes in C-test taking. Bochum, Germany: Universitätsverlag Dr. N. Brockmeyer.
Sternberg, R. J. 1977. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Lawrence Erlbaum.
Sternberger, J. P., & B. MacWhinney. 1988. Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics* (pp. 101–116). San Diego, CA: Academic Press.
Swain, M. 1990. Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? In J. Alatis (ed.), *Georgetown University Round Table, 1990* (pp. 401–412). Washington, DC: Georgetown University Press.
Tarone, E. 1988. *Variation in interlanguage.* London: Edward Arnold.
Teichroew, J. M. 1982. A survey of receptive versus productive vocabulary. *Interlanguage Studies Bulletin, 6,* 3–33.
Tryon, W. W. 1979. The test-trait fallacy. *American Psychologist, 34,* 402–406.
Vollmer, H. 1983. The structure of foreign language competence. In A. Hughes

& D. Porter (eds.), *Current developments in language testing* (pp. 3–29). London: Academic Press.

Votaw, M. C. 1992. A functional view of bilingual lexicosemantic organization. In R. J. Harris (ed.), *Cognitive processing in bilinguals* (pp. 299–321). New York: Elsevier.

Wesche, M. B. 1987. Second language performance testing: The Ontario test of ESL as an example. *Language Testing, 4,* 28–47.

Wesche, M., & T. S. Paribakht. 1993. Assessing vocabulary knowledge: Depth versus breadth. Paper presented at the American Association of Applied Linguistics conference, Atlanta, GA, April.

Widdowson, H. 1983. *Learning purpose and language use.* Oxford: Oxford University Press.

Yang, L., & T. Givon. 1993. Tracking the acquisition of L2 vocabulary: The Keki language experiment. Technical Report No. 93–11. Eugene: Institute of Cognitive & Decision Sciences, University of Oregon.

Young, R. 1989. Ends and means: Methods for the study of interlanguage variation. In S. Gass, C. Madden, D. Preston, & L. Selinker (eds.), *Variation in second language acquisition. Psycholinguistic issues* (Vol. 2, pp. 65–90). Philadelphia: Multilingual Matters.

Zuroff, D. C. 1986. Was Gordon Allport a trait theorist? *Journal of Personality and Social Psychology, 51,* 933–1000.