VI Research Methods

21 Defining and Measuring SLA

JOHN NORRIS AND
LOURDES ORTEGA

1 Introduction: A Framework for Understanding Measurement in SLA Research

Research within the social and cognitive sciences frequently calls upon measurement to provide a systematic means for gathering evidence about human behaviors, such that they may be interpreted in theoretically meaningful ways. The scientific value of resulting interpretations, which explain what is observed in light of what is known, depends in large part on the extent to which measurement practice within a given research domain adheres to standards for the development, use, and evaluation of measurement instruments and procedures (e.g., AERA, APA, NCME, 1999). Where such standards are not in place, or where they are not rigorously followed, measurement practice will produce research "findings" which lack interpretability and generalizability, which do not contribute to the accumulation of knowledge, and which therefore, as Wright (1999) has observed, provide little more than "a transient description of never-to-be-reencountered situations, easy to doubt with almost any replication" (p. 71).

Measurement is used within second language acquisition (SLA) research to elicit, observe, and record the language (and language-related) behaviors of L2 learners, and to enable the interpretation of resulting evidence in light of explanatory theories of the language acquisition process. Although by no means in a state of theoretical accord, the field of SLA is, on the whole, interested in describing and understanding the dynamic processes of language learning (learning used here in its broadest sense) under conditions other than natural, first language acquisition (Beretta, 1991; Bley-Vroman, 1989; Crookes, 1992; Ferguson and Huebner, 1991; Gregg, 1993; Lambert, 1991; Long, 1990, 1993; McLaughlin, 1987). Accordingly, measurement in SLA research generally provides evidence for interpretations about: (i) a learner's linguistic system (i.e., the underlying mental representations of the L2); (ii) development or change

(or the lack thereof) in a learner's linguistic system; and (iii) factors which may contribute to or hinder a learner's developmental approximations of the target L2.

Despite similarities, theoretical accounts of SLA differ widely according to the ways in which acquisition is defined and the types of evidence that are brought to bear in associated research; so, too, do measurement practices differ systematically according to the varying theoretical premises. Although a number of these measurement practices have enjoyed rather lengthy traditions of use within particular SLA research communities, doubts continue to be voiced regarding the extent to which: (i) theoretical constructs are being defined in measurable ways (e.g., Bachman, 1989; Bachman and Cohen, 1998); (ii) measurement instruments and procedures are being systematically developed and implemented (e.g., Polio, 1997); (iii) measurement practices are being subjected to adequate validity evaluation (e.g., Chapelle, 1998); and (iv) the reporting of measurement-based research is adequate for enabling scientific replication and knowledge accumulation (e.g., Norris and Ortega, 2000; Polio and Gass, 1997; Whittington, 1998). Furthermore, it is likely that advances in measurement theory are not afforded consistent attention within measurement-based SLA research (see, e.g., discussions in Bachman and Cohen, 1998; Grotjahn, 1986; Hudson, 1993; Paolillo, 2000; Saito, 1999; Shohamy, 2000), as has been noted with respect to other social science research domains (see, e.g., Embretson, 1999; Thompson, 1998).

The purpose of the current chapter is to address these concerns and to discuss how SLA researchers might organize their thinking about measurement in order better to serve the research endeavor. In the remainder of this first section, we present a framework which defines the scope and process of measurement and which we use throughout the chapter to analyze measurement practices in SLA. We then present an overview of the primary epistemological approaches to be found in the field. This overview establishes the link between the nature of SLA theories, the ways in which acquisition has been defined, and the types of evidence brought to bear in interpretations about "acquisition." We then examine measurement practices and problems associated with SLA research, and we offer recommendations for resolving problems and generally improving measurement practice. Where applicable throughout the chapter, we also indicate recent advances in measurement theory which seem pertinent to the measurement of L2 acquisition. Finally, we end with a discussion of several implications for the future of measurement-based SLA research.

1.1 Constructs, data, and the measurement process

Measurement is at once a data- and theory-driven undertaking (Messick, 1989). This implies, on the one hand, that the kinds of theoretical interpretations to be made have been defined, and on the other, that the kinds of data to be accepted as relevant evidence for such interpretations have been specified. The

first of these assumptions is treated traditionally under the notion of *construct definition*, and the second concerns the nature of *measurement data*.

Historically, constructs were considered unobservable explanatory entities residing within theory and only inferred via the interactions between sets of observable variables. More recently, however, the notion of construct has evolved to acknowledge the interplay between a theoretical explanation of a phenomenon and the data that may be gathered about the phenomenon (see Angoff, 1988; Cronbach, 1988; Cronbach and Meehl, 1955; Loevinger, 1957; Messick, 1975, 1989). This current view is reflected by Chapelle (1998), who maintains: "[a] construct is a meaningful interpretation of observed behavior" (p. 33). Construct definitions, then, provide an explicit delineation of the interpretations that are intended to be made on the basis of a measure. As such, they dictate the theoretical meanings which may be attached to measurement data; without construct definitions, measurement data are meaningless.

Measurement *data* are composed of repeated observations of particular patterns in behaviors (Chapelle, 1998; Cronbach, 1980), and these observations are condensed into scores of some kind, which can be defined as "any coding or summarization of observed consistencies on a test, questionnaire, observation procedure, or other assessment device" (Messick, 1989, p. 14). The types of data which constitute acceptable evidence about a construct are typically drawn from an empirical knowledge base. For example, accumulated findings from a series of longitudinal descriptive studies of the given phenomenon may lead to an association between particular observable behaviors and a theoretical explanation for those behaviors. Given such an empirical association and an explicit definition of the construct, the kinds of data which may serve as evidence for interpretations can be specified.

Measurement, then, involves the collection of data, the transformation of those data into evidence, and the use of that evidence for making a theory-based interpretation. In practice, measurement proceeds according to several interrelated but distinguishable stages (see discussions in Bennett, 1999; Messick, 1989, 1994; Mislevy, 1994, 1995; Mislevy et al., forthcoming), which are outlined in figure 21.1. Note that the measurement process there begins and ends with interpretation; thus, intended interpretations are the starting point for developing appropriate measures, and actual interpretations are the culmination of using measures. Note also that the arrows in figure 21.1 proceed only in one direction, with each stage feeding into the next. This unidirectionality shows the chronological progression of stages in measurement development and use. At the same time, the graduated shading in the model and its cyclical composition indicate that the process is not static; while individual stages are primarily conceptual or primarily procedural, decisions and discoveries at each stage of the process may influence developments at all other stages. Finally, the ultimate outcomes of the measurement process obviously feed back into revised theoretical interpretations.

Each of the stages in figure 21.1 implies particular actions on the part of researchers. The first three stages require that researchers *conceptualize* the

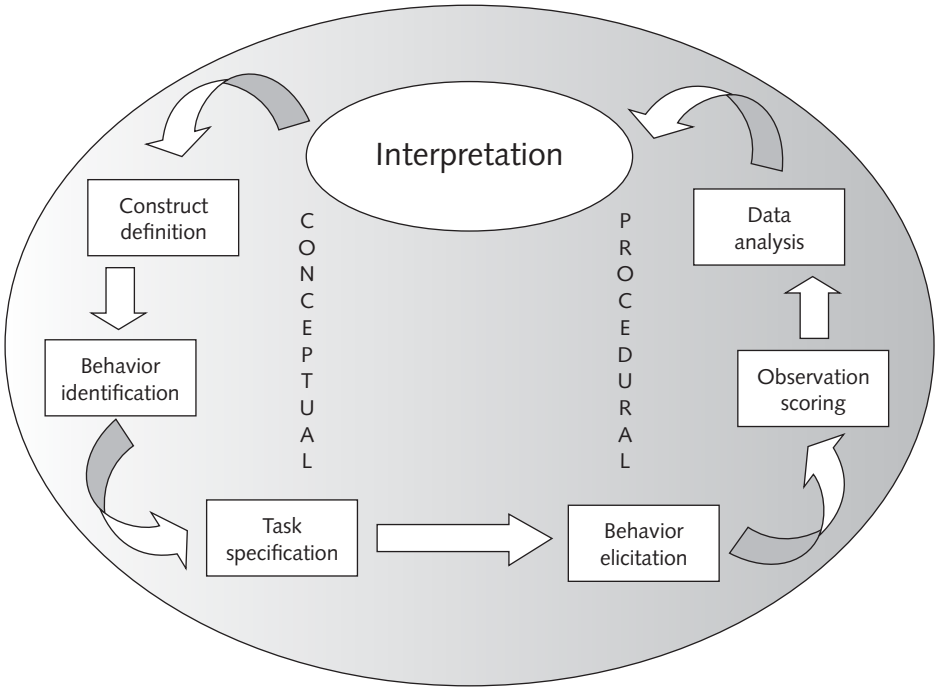

Figure 21.1 The measurement process

evidence to be provided with a given measure, by defining intended construct interpretations and linking them with observable behaviors:

1 *Construct definition*: For a given measure, researchers explicate exactly what it is that they want to know based on what kinds of interpretations are going to be made. Constructs should be defined in specific terms, such that observable behaviors may be obviously linked with them, and they should provide a clear indication of the theoretical assumptions that they represent.
2 *Behavior identification*: Researchers decide what particular behavior or constellation of behaviors needs to be observed, as well as what qualities or variations in those behaviors are important, in order to provide sufficient evidence for a given construct interpretation. The link between target behaviors and constructs emerges from an empirical knowledge base; that is, researchers draw on accumulated knowledge about the construct in order to identify evidentiary requirements in the form of behaviors.
3 *Task specification*: The researcher specifies a particular set of tasks or situations for the elicitation/observation of targeted behaviors. Tasks/situations should also be linked to behaviors via an empirical knowledge base.

In practice, this implies the careful analysis of tasks/situations in order to determine whether they can provide the behavioral evidence required of them (see Bachman and Palmer, 1996; Mislevy, Steinberg, and Almond, 1999; Norris, Brown, Hudson, and Yoshioka, 1998; Skehan, 1998). Tasks/situations should be defined in terms explicit enough to enable exact replication.

In the next three stages, researchers *proceduralize* the outcomes of the conceptual stages, implementing mechanisms for the elicitation, scoring, and analysis of behavioral data in order to provide evidence for interpretations:

4 *Behavior elicitation*: Data on targeted behaviors are elicited, observed, and recorded via the administration of tasks or the observation of situations, while the potential influence of other variables is carefully controlled or accounted for (this incorporates the whole of instrument operationalization and administration; see practical guides in AERA, APA, NCME, 1999; Bachman and Palmer, 1996; J. D. Brown, 1996, forthcoming; Linn, 1989; Popham, 1981; Seliger and Shohamy, 1989).
5 *Observation scoring*: Data are attributed initial construct-relevant meaning by researchers classifying variations in observed behaviors according to the range of previously identified criterial values; the score should summarize observations in a way that may be clearly linked to intended interpretations. In practice, scoring is based on the use of numeric scales which reflect meaningful values, including categorical, ordinal, interval, and ratio types (see Angoff, 1984; Bachman and Palmer, 1996; Brindley, 1998; J. D. Brown, 1996; Wright, 1999). The reliability of scoring is also evaluated, in order to establish the extent to which score summaries represent systematic versus unknown or unintended sources of variability, by estimating classical and other sorts of reliability (see Feldt and Brennan, 1989; Hambleton, Swaminathan, and Rogers, 1991; Orwin, 1994; Shavelson and Webb, 1991; Traub, 1994).
6 *Data analysis*: Individual scores and patterns of scores are compared and summarized in light of various categorical and probabilistic properties. Behavioral predictions from the construct definition stage (e.g., in the form of hypotheses) are evaluated using various techniques (statistical description and inference, implicational scalar analysis, etc.; see J. D. Brown, 1988, 1996; Hatch and Lazaraton, 1991; Tabachnick and Fidell, 1996; Woods, Fletcher, and Hughes, 1986).

In a final stage, which forms the culmination of the cyclical measurement process outlined in figure 21.1, measurement outcomes are incorporated as evidence for construct interpretations. At this point, researchers (and the research community) discuss the outcomes from their measures in light of theoretical predictions, and they integrate the new evidence into an existing research knowledge base.

1.2 Construct validation

The objective of proceeding through each of the measurement stages above, carefully building on the foundations of the previous stage, is to produce a *warranted* interpretation about the construct of interest. An interpretation is warranted when researchers can demonstrate that a measure has provided trustworthy evidence about the construct it was intended to measure. Of course, the intended construct interpretation, as originally defined from the point of view of theory, is susceptible to becoming *un*warranted at any and all of the stages in measurement on *each* occasion of measurement use. As such, it is incumbent on individual researchers as well as the research community to investigate the construct validity of measurement, asking to what extent their practices in developing and using a measure result in an interpretation or set of interpretations that may be warranted (see AERA, APA, NCME, 1999; Messick, 1989). Comprehensive validation in educational measurement generally involves an evaluation of the entire process of test use, including the social consequences and values implications of applied test use and the relevance/utility of particular test scores for decisions and other actions (see, e.g., Kane, 1992; Linn, 1997; Messick, 1989; Moss, 1992; Shepard, 1993, 1997). However, when measures are employed as research tools, validation may be usefully constrained to a focus on the measurement stages outlined above and on the resulting construct interpretations (indeed, it is these interpretations which generally define the extent to which research measures are intended to be used).

The major threats to construct validity in measurement are of two types. *Construct underrepresentation* indicates the "degree to which a test fails to capture important aspects of the construct," whereas *construct-irrelevant variance* is the "degree to which test scores are affected by processes that are extraneous to its [*sic*] intended construct" (AERA, APA, NCME, 1999, p. 10). Problems of construct underrepresentation typically occur during the conceptualization of a measure (stages 1–3 above), when researchers fail adequately to consider (and demonstrate) a relationship between intended interpretations and the observable behaviors which will provide evidence about them. Construct-irrelevant variance is usually introduced during the proceduralization of a measure (stages 4–6 above), when researchers fail to control or account for the potential influence of the act of measuring itself (including scoring and analysis, as well as elicitation) on construct interpretations.

In order to engage in sound measurement practice in SLA research, and to better understand the extent to which their interpretations may be threatened by construct underrepresentation or construct-irrelevant variance, researchers will need to understand the relationship between SLA theories and the ways in which each of the stages in the measurement process is pursued. Therefore, we now turn to an examination of the link between SLA theories and their definitions for acquisition, the types of evidence brought to bear upon acquisition constructs, and the measurement practices employed within acquisition research.

2 What Counts as L2 Acquisition? Conceptual Bases for Measurement in SLA

Since the inception of SLA as a field (see discussion in Huebner, 1991; Larsen-Freeman, 2000), theories of acquisition have multiplied, reflecting both a broadening scope of inquiry and interdisciplinary excursions by researchers. Diverging epistemologies have also led, undoubtedly, to "conflicting views about the 'best' way to gather data and/or the 'correct' questions to be asked" (Gass, 1988, p. 199). As a consequence, what counts as L2 acquisition – including what constructs are of interest, how they are defined, and what kinds of observable data are accepted as evidence – has become increasingly complex, varied, and at times disputed.

A persistent concern of many SLA researchers has been the relevance of linguistic theory for explaining L2 acquisition, and vice versa (for example, see articles in Huebner and Ferguson, 1991). As Huebner (1991) pointed out, "to the extent that linguistic theories are concerned with diachronic change, language development, language universals, or the nature and acquisition of grammatical and communicative competence, the phenomena involved in SLA must be of central concern to linguistic theory" (p. 4). Since the 1970s, the predominant linguistic theory, at least in the US, has been of a Chomskian generativist bent. However, as Lightbown and White (1987) observed, it was not until the mid-1980s that some SLA researchers paid more than lip service to generative linguistics, in vague references to a universal grammar, and started developing a research agenda for a formal linguistic theory of learnability in SLA (see, e.g., Eubank, 1991; Gass and Schachter, 1989; Rutherford, 1984). Thus, generative SLA researchers have begun to investigate the extent to which purportedly innate Universal Grammar (UG) principles and parameters are accessible in L2 acquisition (see White, 1996, 2000, this volume, for an overview of the various positions). Another line of research has concentrated on investigating the fundamental similarity or difference (Bley-Vroman, 1989) not only between L1 and L2 acquisition, but also between child L2 and adult L2 acquisition (e.g., Schwartz, 1992). Finally, an area of research receiving increased attention in generative SLA concerns the hypothesis of a critical period and associated maturational constraints on the attainment of nativelike, UG-constrained competence by non-native speakers (e.g., Birdsong, 1999; Hyltenstam and Abrahamsson, this volume; Sorace, 1993; White and Genesee, 1996).

For other researchers, linguistic theory alone has not been epistemologically sufficient. The need for SLA to explain differential success and, often, failure among second (particularly adult) language learners fostered a two-fold focus on linguistic and non-linguistic (social, affective, and cognitive) variables that influence the L2 acquisition process. From such research concerns stemmed a second theoretical strand that has gained prominence since the early 1980s: that of interactionist SLA (or interactionalist SLA; see Chapelle, 1998). Interactionist

approaches to SLA focus on the relationship between learner-internal and external processes in L2 acquisition. Input, interaction, and output were the essential external variables identified within initial social interactionist research agendas (see Krashen's, 1981, input hypothesis; Long's, 1980, interaction hypothesis; and Swain's, 1985, 1995, output hypothesis). More sociolinguistically oriented research has investigated the influence of social context on acquisition, as in IL variation theories (R. Ellis, 1985; Tarone, 1988), and the interaction of learner variables with social context, as in Gardner's (1979) social psychological model and Schumann's (1978) acculturation model. Interest in the role of learner-internal variables, influenced by theories of learning within an information-processing approach to cognitive psychology, has spurred the development of cognitive interactionist theories of SLA, such as a skill theory of L2 acquisition (Bialystok, 1991; McLaughlin, 1987), a psycholinguistic theory of universal operating principles for L2 acquisition (Andersen, 1984), and a processing constraint theory of L2 acquisition (Pienemann, 1984, 1998).

Until recently, these two distinct theoretical perspectives, generativist and interactionist, comprised the SLA research mainstream. The 1990s brought two new types of theories into the field, along with unique epistemologies: emergentism and sociocultural theory. Sociocultural theories maintain that learning of any kind (including language learning) is an essentially social process rather than one generated within the individual. Second language, like first language and thought itself, develops in the social, inter-mental plane, and only subsequently is it appropriated by the individual into the intramental plane (Lantolf, 1994; Vygotsky, 1986). Because research driven by sociocultural theories of L2 acquisition does not, in general, employ measurement of the sort discussed in this chapter, we make no further reference to such work (although sociocultural approaches are by no means exempt from the concerns raised in this chapter, wherever measurement is employed). Emergentist theories view L2 learning, like all human learning, as the outcome of a neurobiological tendency of the brain to attune itself to primary sensory experience through the strengthening and weakening of connections among the billions of neurons that it typically develops. Linguistic knowledge (or the phenomenological experience thereof) emerges as a by-product of the establishment of networked connections upon exposure to probabilistic patterns underlying the (L1 or L2) linguistic input (e.g., N. Ellis, 1998, 1999). In fact, emergentism is radically different from both generativist and interactionist epistemologies. On the one hand, it is incompatible with generative SLA because it denies symbolism, modularity, and innatism, and it removes linguistics from the center of the research domain, replacing it with cognitive architecture. On the other hand, in spite of the shared interest in functionalist explanations and cognitive constructs, emergentist theory resonates little with interactionist SLA. The highly specialized neurobiological treatment of cognitive processes, the lack of a traditional dichotomy between representation and access, and the absence of interest in non-cognitive variables (social, affective, educational, etc.) all differentiate emergentist from interactionist perspectives.

Although fundamental differences in the theories outlined above often lead to sharp divisions among SLA researchers according to what may or may not count as acquisition, it is not our intention here to address theory construction or evaluation (see Beretta, 1991; Crookes, 1992; Gregg, 1993; Long, 1990, 1993). Instead, we maintain that whatever theoretical questions are posed and however data are gathered, where measurement is used, careful construct definition and adherence to measurement standards will provide a rational guide for enabling and improving the research process. Therefore, we turn now to an examination of the first three conceptual stages of the measurement process outlined in figure 21.1, asking of SLA research:

i How are constructs defined via the interpretations made about acquisition from different theoretical perspectives?
ii Have criterial behaviors and behavioral qualities been identified which can provide sufficient evidence for making such construct interpretations?
iii Are measurement tasks/situations designed to elicit adequate and accurate behavioral data?

2.1 Construct definition: interpretations about L2 acquisition

In order to define acquisition as a construct for measurement purposes, the particular interpretations to be made about L2 acquisition must first be sought within existing SLA theories. Table 21.1 summarizes some of the essential features (interpretive as well as evidentiary) for three main theoretical approaches to SLA.

Generative SLA views language as a symbolic system, autonomous from cognition, and too complex to be acquired by training or through inductive or deductive learning from the input. Since it adheres to the tenets of first language nativism, generative SLA research aims at elucidating empirically whether learners can have indirect, partial, full, or no access to the principles of Universal Grammar in the process of acquiring an L2, and it prioritizes interpretations about linguistic competence, not language performance (Gregg, 1990; Schwartz, 1993; White, 1991). Further, this epistemological approach to L2 acquisition focuses on constructs which describe and explain the origins of linguistic mental representations (the "competence problem" central in a property theory) and does not concern itself so much with interpreting how such representations unfold or become available to the learner in a predictable route (the "developmental problem" central in a transition theory) (see Gregg, 1996). Therefore, generative SLA research confines itself to formal descriptions of interim learner grammars (i.e., syntax) as reflected in a learner's tacit ability to judge ungrammaticality in the L2, because it assumes that the goal of SLA as a theory is to explain how learners can acquire a full mental representation of many of the complexities of the L2, and why they cannot acquire all aspects of an L2 syntax (and precisely which aspects learners may fail to acquire).

Table 21.1 What counts as L2 acquisition for three types of SLA theories

Stage	*Generative SLA*	*Interactionist SLA*	*Emergentist SLA*
Epistemology and construct interpretations	Language as symbolic representation which is autonomous from cognition Learning mediated by UG and L1 Grammatical competence Property theory: initial state and end state in L2 acquisition	Language as symbolic representation which is constrained by cognition Learning mediated by social, affective, and cognitive variables Communicative competence Transition theory: developmental course of L2 acquisition (For information-processing theories) automatization of declarative knowledge	Language as complex rule-like behavior, epiphenomenal result of functional needs Learning as interaction of the organism with the environment Neural networks Transition theory: specification of input frequency and regularity plus learning mechanisms
Target behaviors	Tacit intuiting of what is ungrammatical in the L2	Appropriate and fluent performance when using the L2 communicatively (and in controlled tasks)	Accurate and fluent performance in laboratory tasks Output that matches attested learning curves and eventually matches characteristics of fed input
Elicitation tasks/situations	Grammaticality judgment tasks of various kinds	Spoken and written discourse production Tests of implicit and explicit knowledge: verbalization of understanding of rules; controlled performance on comprehension and production tasks; grammaticality judgment tasks	Implicit memory tasks and forced-choice reaction-times tasks with human learners in laboratory Computer simulations of neural networks

Generative linguistic studies of SLA are likely to rely almost exclusively on the outcomes of grammaticality judgment tasks of various kinds, where *acquired* means nativelike levels of rejection of illegal exemplars of the target grammar.

Interactionist SLA, on the other hand, is based on functionalist views of language as a symbolic system that develops from communicative needs (Tomlin, 1990; Tomasello, 1998a). Language is believed to be a complex faculty that is acquired by the learner through engagement with the environment, through inductive and/or deductive learning from input, and in a constructive process (in the Piagetian sense) constrained by general cognition (see Long, 1996; Richards and Gallaway, 1994). Hence, language acquisition is thought of as a gradual process of active form/function mapping, and the traditional dichotomy between competence and performance is not maintained; instead, language learning is inextricably related to language use in that performance is viewed as driving competence (Hymes, 1972; see papers in G. Brown, Malmkjaer, and Williams, 1996; and discussion in McNamara, 1996, ch. 3). Interactionist epistemologies, drawing on functionalist linguistic theories, such as variationist sociolinguistics (Preston, 1989), functional grammar (Givón, 1979), and discourse analysis (Sinclair and Coulthard, 1975), focus not so much on the origin and description of linguistic representation as on the "developmental problem" (e.g., Pienemann, 1998). Not only, therefore, do interactionist SLA theories need to describe and explain learner transitional grammars, but their interpretations must also invoke non-linguistic (i.e., cognitive and environmental) constructs thought to be crucial in accounting for how learning of an L2 takes place on a predictable route and with differential ultimate success. Interactionist SLA researchers maintain that acquisition of L2 forms cannot be demonstrated until such forms are productively used in a variety of contexts in spontaneous performance; a multiplicity of performance data is therefore required to produce a complete picture of language development. In addition, this type of theory argues that incremental, non-linear *changes* (not necessarily target-oriented improvements) in patterns of language use can be taken as indications that gradual learning is taking place (e.g., Mellow, Reeder, and Forster, 1996). Consequently, interactionist studies (at least logically ought to) draw on measures of implicit and explicit memory for L2 forms (i.e., recognition tasks where *acquired* means detected or noticed), measures of explicit knowledge of rules (i.e., metalinguistic verbalization tasks, where *acquired* means understood with awareness), and measures of the use of L2 forms in spontaneous, meaning-driven discourse (i.e., comprehension and production tasks involving sentence-level and, preferably, text-level performance, where ability for use is demonstrated). In sum, under interactionist approaches to SLA, *acquired* may mean a number of gradual and non-linear changes in the linguistic (and, in some theories, metalinguistic) behavior that characterize the developmental course of L2 acquisition, based on construct interpretations such as: (i) a form has "emerged," has been "detected," "noticed," "attempted," or "restructured"; (ii) a learner is "aware" of a form or a form-related pattern; and/or (iii) a learner is "able to use a form appropriately and fluently."[1]

Finally, emergentism provides a combined functional and neurobiological approach to language acquisition that views grammar as a complex, rule-like, but not rule-governed system arising from the interaction of very simple learning mechanisms in the organism (the architecture of the human brain) with the environment (massive exposure to input). Emergentist theories of L2 acquisition seek to explain the frequency and regularity of linguistic input to which the learner must be exposed in order for the processing system (i.e., the brain) to develop a functional set of weights (i.e., degree of interconnectivity among nodes) that will match patterns underlying that input (Sokolik, 1990). Speeded, accurate production of output that matches the input provides evidence that such functional sets of weights in the neural networks have been established on the basis of simple learning algorithms and exposure to positive input alone (N. Ellis, 1998). Consequently, emergentist-connectionist studies typically employ computer modeling experiments and trials with human subjects under laboratory conditions, with interpretations based on reaction-time decision tasks involving carefully controlled input (e.g., N. Ellis and Schmidt, 1997). *Acquired*, for emergentists, means fast, accurate, and effortless performance attained along attested learning curves that reflect non-linear, exemplar-driven learning.

Obviously, each of the preceding theoretical approaches to SLA defines acquisition in unique ways and calls for particular construct interpretations to be made on the basis of measurement data. Indeed, what counts as acquisition is so dependent on the theoretical premises of the research domain that the same measurement data may be interpreted as evidence of acquisition or the lack thereof, depending on the theoretical approach adopted. A good illustration of this point can be found in a well-known study by Trahey and White (1993). Measurement outcomes from this study showed that young francophone learners in intensive ESL programs in Quebec, after a two-week regime of exposure to English input flooded with adverbs, accepted more cases of Subject-Adverb-Verb-Object sentences (ungrammatical in the L1 but grammatical in English) than they had accepted before. However, positive evidence alone (i.e., exposure to only correct SAVO exemplars in the flooded input) did not cause these learners to reject Subject-Verb-Adverb-Object sentences (grammatical in the L1 and ungrammatical in English). From the generativist perspective of the authors, these measurement observations were interpreted to show that acquisition had not occurred, because there was no evidence of parameter resetting, which would require simultaneous acceptance of SAVO and rejection of *SVAO. However, arguments from interactionist SLA, including developmental accounts of L2 learning (e.g., Meisel, Clahsen, and Pienemann, 1981; Mellow et al., 1996) and claims about the role of attention and awareness in L2 learning (e.g., Schmidt, 1993, 1994; Tomlin and Villa, 1994), would call for an alternative interpretation of the same data as evidence for incipient acquisition of adverb placement in L2 English. In fact, in studies of implicit and incidental instructional conditions (i.e., external interventions that do not orient learners to learning with intention; see Schmidt, 1993) researchers have repeatedly found

evidence for acquisition in small post-instructional increases in recognition of or preference for the targeted form (a behavior typically observed in input flood treatments, as in Trahey and White, 1993) and/or in increased, albeit initially unsuccessful, attempts to produce the targeted form (a behavior typically observed in typographical input treatments; see Alanen, 1995; Jourdenais et al., 1995).

To summarize, what counts as acquisition (theoretically defined), as well as the utility of viewing L2 acquisition in particular ways, may be disputed by researchers from differing paradigms. However, such disagreements themselves bear witness to the fact that construct definitions are available. Given theoretical construct definitions, additional conceptual bases for measurement may be evaluated. Therefore, we turn now to an examination of the evidence required for making interpretations about acquisition and the measurement tasks used to provide such evidence.

2.2 *Behaviors and tasks: evidence for acquisition*

As indicated in section 1.2, the major threat to validity during the conceptualization of measurement involves construct underrepresentation. Construct underrepresentation occurs when the complex *link* between a theoretical interpretation and required behavioral evidence is inadequately understood and/or conveyed into practice. In order to avoid underrepresentation of a construct, researchers must carefully define the evidentiary requirements (in the form of behaviors) for their intended interpretations, then link these requirements to empirically, or at least logically, related elicitation tasks or situations, which are themselves understood in terms of the behavior(s) that they elicit. Given the range of measurement tasks actually employed by SLA researchers, from discrete-point recognition items to full-blown spontaneous communicative performance, as well as the range of construct interpretations that are based on them, the possible sources for construct underrepresentation are many. In this section, we address four of the most serious (and most common) conceptual problems: providing evidence for both causal and outcomes interpretations (section 2.2.1); understanding and matching complex interpretations with complex behaviors (section 2.2.2); specifying the variable qualities of behaviors in meaningful units that are sensitive to the levels of interpretation to be made (section 2.2.3); and avoiding the "valid test" fallacy (section 2.2.4).[2]

2.2.1 Evidence for causes and outcomes

Where interpretations are to be made about the relationship between causal or moderating processes (noticing, comprehension, cognitive resources of memory and attention, attentional focus, language aptitude, etc.) and L2 acquisition products, behavioral evidence for such constructs will also need to be specified and associated measurement tasks selected. SLA research frequently employs dependent variable measures which only provide evidence bearing on the linguistic "products" of acquisition (vocabulary recognition items, grammaticality

judgment tasks, elicited imitation, communicative performance, etc.). Such measures do little to inform interpretations about the independent variables to which acquisition-related behavioral patterns are ascribed; the actual construct interpretations (i.e., about the relationship between certain causes and linguistic outcomes in acquisition) will thus be underrepresented within measurement practice.

Two recent cognitive interactionist proposals for task-based second language learning, advanced by Robinson (2001b) and Skehan (1998), provide a good illustration of theories which call on measurement simultaneously to inform both causal and outcomes interpretations. These two theoretical models invoke distinct explanatory processes while predicting very similar changes in L2 behavior. In both theories, the more cognitively complex a task (a meaning-oriented communicative activity), the more likely it will yield increasingly more complex but less fluent language output by learners. Both models posit this relationship on the assumption that cognitive complexity of tasks is positively related to L2 learning. However, Robinson (2001b) argues that the linguistic processing demanded by cognitively more complex tasks entails a mobilization of attentional pools dedicated to language production, and thus pushes the internal system in several ways (i.e., by fostering deeper linguistic processing that promotes rehearsal in short-term memory and eventual reorganization of form/function connections; see also Robinson, 1995). This is essentially an emergentist or functionalist rationale (see N. Ellis, 1998; MacWhinney, 1998; Tomasello, 1998b) that rests on a multiple-resource model of attention and memory (Wickens, 1989). By contrast, Skehan (1998) claims that unmitigated/uncensored cognitive complexity can have the undesirable effect of overloading a learner's limited attentional resources and fostering an easy way out through lexical (as opposed to syntactic) processing of L2 input and output. Therefore, according to Skehan, during competence-expanding L2 performance, it is necessary to orchestrate learner-external interventions to ensure that learners consciously attend to the linguistic code and prioritize accuracy goals during performance. This is in essence an information-processing and skills-acquisition rationale that assumes limited attentional capacity (see Anderson, 1993; McLaughlin, 1987).

Since both Robinson (2001b) and Skehan (1998) predict, as a result of task-based learning, very similar outcomes in terms of L2 performance (with regard to productive complexity and fluency; accuracy is much-debated terrain – see Ortega, 1999), the only way to inform the full range of interpretations that need to (and will) be made in related research is by gathering evidence bearing on the explanatory constructs invoked in each theory in addition to language performance data. For Robinson's predictions to be measurable, this will mean eliciting behaviors that reflect psycholinguistic operations (e.g., deeper processing and rehearsal in short-term memory), which reside beyond conscious control. For Skehan's theory, behaviors must be elicited which reflect metalinguistic operations (e.g., strategic attention to the code and a prioritization of accuracy), which are subject to conscious learner control. Each

type of interpretation calls for distinct, indirect techniques to provide empirical evidence for either psycholinguistic or metalinguistic operations. For instance, introspective methodologies (see Ericsson and Simon, 1993; Sugrue, 1995) seem the best available options for accessing metalinguistic operations, whereas implicit memory tasks (priming tasks, implicit recognition tasks, etc.) may be the most appropriate choices for attempting to tap psycholinguistic, automatic operations (see Bjork and Bjork, 1996; Stadler and Frensch, 1998). Finally, measurement in the service of both theories will also need to provide evidence for interpretations about the so-called cognitive "complexity" of L2 performance tasks (see discussion in Norris, Brown, Hudson, and Yoshioka, 1998; Robinson, 2001a; Skehan, 1998). Of course, establishing a link between the full sets of interrelated constructs (cognitive complexity, linguistic complexity, strategic accuracy-orienting operations, deeper processing operations, complexity/fluency/accuracy in performance) and long-term L2 learning, rather than immediate L2 performance, raises additional questions regarding the timing and frequency of measurement that will be necessary to provide adequate evidence for such complex interpretations.

This example underscores the necessity of defining the evidentiary requirements for *all* construct interpretations to be based on measurement, such that an adequate range of corresponding behaviors may be elicited. Other explanations for SLA which are based on the contribution of causal processes run a similar risk of construct underrepresentation, including: the role of noticing and awareness (e.g., Leow, 1997) and attentional focus (e.g., Williams, 1999); the potential contribution of uptake (e.g., Lyster, 1998; Mackey and Philp, 1998); the moderating influence of aptitude (e.g., Sawyer and Ranta, 2001; Grigorenko, Sternberg, and Ehrman, 2000); and the relationship between interactional modifications and actual L2 learning, via either facilitated comprehension (e.g., Loschky, 1994) or provision of negative feedback (e.g., Iwashita, 1999; Mackey, 1999). For these and other approaches to acquisition research which make reference to cognitive processes, advances in measurement within the cognitive sciences should prove instructive, where, as a rule, a multiplicity of behavioral observations is gathered to inform and triangulate interpretations (see Pellegrino, 1988; Siegler, 1989; Snow and Lohman, 1989; Sugrue, 1995). For example, Royer, Cisero, and Carlo (1993) point out that "cognitive assessment procedures should be able to provide indices of change in knowledge organization and structure and indices of the accuracy, speed, and resource load of the activities being performed" (p. 202). Bennett (1999) also shows how developing technologies will enable researchers simultaneously to capture and measure a much wider array of behavioral evidence bearing on cognitive constructs.

2.2.2 Matching complex interpretations with complex behaviors

Whereas the previous section addressed problems in construct underrepresentation which occur when researchers fail to employ multiple measures for multiple interpretations, this section addresses problems arising from the multidimensional or complex nature of both the evidence required by particular constructs

and the evidence provided by particular behaviors. In the context of correlational and experimental research on child language acquisition, Richards (1994) calls the problem of ignoring or underestimating the complexity of variables at play the *holistic fallacy*. This fallacy arises when the relationship between behaviors and constructs is conceptualized as being "more widely applicable or more uniform than may be the case" (p. 100). The holistic fallacy can take several forms in SLA research. On the one hand, researchers may fail to recognize the complex nature of the behavioral evidence that is required by a given construct interpretation; in such cases, resulting measurement data tend to be overinterpreted because the behaviors selected to be observed do not, in fact, provide sufficient evidence for the full construct interpretation. On the other hand, researchers may fail to recognize the complexity of the behavioral evidence that will be provided by measurement tasks/situations, when the actual sources of variability within the selected behaviors are not understood; in these cases, measurement data tend to be underinterpreted because the observed variations in behavior may really be attributable to factors beyond those found in the construct interpretation.

Nichols and Sugrue (1999) have observed that many educational tests and test items fail adequately to reflect intended constructs because of a mismatch between "the simple cognitive assumptions often embedded in conventional test development practices and the cognitively complex nature of the constructs to be measured" (p. 18). Several measurement examples in SLA research underscore similar problems. In a meta-analytic review of experimental and quasi-experimental studies of L2 instruction, Norris and Ortega (2000) compared the observed magnitude of effects when instructional outcomes were measured using metalinguistic judgments (various kinds of grammaticality judgment tasks), free constructed responses (discourse-level communicative L2 performance), and constrained responses (selecting or producing word- or clause-level linguistic responses). They found that the observed effects associated with constrained response types ranged from half again up to as much as three times the effects associated with metalinguistic judgments and free constructed response types. Obviously, in light of the consistent differences in observed effects, researchers would come to very different conclusions about acquisition if they chose to elicit constrained response behaviors instead of the other evidence types. Indeed, there is good reason to believe that the constrained response type of measure does not adequately reflect the complexity of interpretations being made about L2 acquisition in such studies. Constrained response tests reduce language behavior to the single instance of "ticking the right box" or producing a form out of extended discursive context. Given the disjuncture between such isolated language-like behaviors and either communicative language use or a learner's underlying mental representation of the L2 grammar, the link with complex interpretations about changes in ability for use or grammatical competence is at best tenuous. While not without their own problems, it can be argued that the behaviors elicited in metalinguistic judgments and free constructed response measures better reflect constructs

like "grammatical competence" and "ability for use." Metalinguistic judgments directly ask learners to indicate which aspects of the grammar they find acceptable and which they do not, behaviors which, if carefully planned and elicited (e.g., Sorace, 1996), may provide a much more complete depiction of the learner's internal L2 grammar than the suppliance of "correct" responses to isolated grammar questions. Likewise, free constructed response behaviors offer insights into how a learner actually deploys acquired L2 forms in real-time, meaning-focused communication, as opposed to how a learner responds to selected language forms presented out of context.

A number of other complex construct interpretations in SLA research call for complex behaviors to be elicited. For example, as Sorace (1996) has pointed out, interpretations about grammatical competence which attempt to incorporate inherently variable phenomena (i.e., as opposed to ignoring variable phenomena which "are not representative of a learner's linguistic knowledge," Gass, 1994, p. 308), such as grammatical indeterminacy, optionality, and hierarchies of grammatical acceptability, will be poorly served by grammaticality measures which simply ask learners to judge sentences categorically as either acceptable or not. In order to inform such interpretations, measurement will need to enable a greater range in elicited response behaviors which may better reflect the range of actual interpretations (e.g., magnitude estimation techniques in Bard, Robertson, and Sorace, 1996; Sorace, 1996; Sorace and Robertson, forthcoming; Yuan, 1997). Where interpretations are to be made about dynamic constructs, such as grammatical development along attested routes of acquisition, multiple instances of behaviors will need to be elicited over time, in order to determine what rules or forms may already be present or not within the learner's interlanguage system, and what a change in behavior with a rule or form may indicate (emergence of a rule, U-shaped or omega-shaped developmental behavior, etc.). Where only static behaviors are elicited, as is often the case in cross-sectional research or pre-test/post-test design studies (see Willett, 1988), unidentified baseline trends in behavior may go undetected at a single point of measurement because the dynamic nature of the construct is not reflected (see Mellow et al., 1996; Pienemann, 1998). Finally, because of the accidental statistical structure of an impoverished language corpus, interpretations about the existence or absence of a given rule/form in the IL system may be unwarranted (Bley-Vroman, 1983). For example, where interpretations are to be made about the emergence of linguistic phenomena which exhibit both variational and developmental characteristics (such as emergence of word order rules in L2 German acquisition; Meisel et al., 1981), measurement will need to elicit behaviors across numerous linguistic and communicative contexts in order to show that interpretations are not based on a lack of evidence, as opposed to evidence for the lack of emergence (see discussion in Hudson, 1993; Pienemann, 1998; and potential solutions in Pienemann, 1998; Pienemann, Johnston, and Brindley, 1988).

Although measurement data may often be overinterpreted as SLA researchers attempt to provide evidence for complex constructs, it is likely that measurement

data are more frequently underinterpreted when researchers do not adequately conceptualize the complexities of measurement behaviors that they intend to elicit. Thus, while elicited behaviors may reflect intended constructs in part, no elicitation procedure, regardless of how much control is exercised by the researcher, is immune to variability introduced by the interaction of the human subject with the measurement task or situation. In this regard, an issue raised some time ago by Grotjahn (1986) rings particularly true for measurement in SLA research: "in order to really understand what a (language) test measures [. . .], we first have to understand the individual task-specific cognitive processes on which the observed performance depends" (p. 162). Making warranted interpretations on the basis of elicited performance will depend, then, on understanding to what extent observed behaviors are influenced by the interaction of learner variables with task/situation variables (see Bachman and Cohen, 1998; J. D. Brown, Hudson, Norris, and Bonk, forthcoming; Norris, 2000).

Observed performances on L2 measurement tasks may be influenced by a number of learner variables which may or may not be reflected in intended construct interpretations. For example, undocumented differences in learners' prior L2 knowledge (in terms of overall proficiency; see discussions in Hulstijn, 1997; Thomas, 1994) and/or current interlanguage status (e.g., in terms of developmental readiness to acquire a particular structure; see Pienemann, 1998) will prove problematic for developmental as well as causal interpretations in SLA research. Unless learners have been characterized according to language ability or psycholinguistic readiness vis-à-vis the acquisition construct in focus (Chaudron, 1985), elicited behaviors, especially if they are summarized at the group level, may lead to misinterpretations about L2 development or the lack thereof, the relative effectiveness of a given instructional treatment, etc. Likewise, differences in how learners respond to a measurement task at motivational, cognitive, and metacognitive levels will determine in part the performance behaviors that may be observed (Royer et al., 1993; Sugrue, 1995). For example, Leow (2000) found that learners who became aware of targeted forms during experimental exposure, as opposed to those who remained unaware of them, increased in their ability to recognize and produce the same forms immediately after the experiment (cf. similar findings in Alanen, 1995). In such cases, construct interpretations would need to tease out the learner's state of awareness in terms of the structures being measured in order comprehensively to understand elicited language performance behaviors. A number of additional individual learner differences may also influence the language behaviors elicited during measurement, including language aptitude, memory capabilities, learning backgrounds, first language, linguistic training, and mental state (see Bardovi-Harlig, 1994a; de Graaff, 1997; DeKeyser, 1995; Robinson, 1997; Sorace, 1996; Zobl, 1995).

Observed performances may also be influenced by characteristics of the measurement tasks/situations themselves, which again may or may not be reflected in intended construct interpretations. For example, the linguistic contexts elicited in measurement may vary according to communicative activity

type. Tarone and Parrish (1988) found that a narrative activity was inherently less demanding on learners' abilities to apply English article rules than was an interview activity. Whereas the narrative primarily elicited linguistic contexts for the least difficult type of reference (i.e., reference to an entity already introduced in the narration), the oral interview elicited a balanced mixture of contexts for all three types of reference involving article use (see Huebner, 1983). Obviously, interpretations about learners' abilities with this particular grammatical subsystem would depend largely on an understanding of the particular elicitation tasks selected. Similarly, language performance behaviors may depend in part on the formatting and presentation of measurement tasks. For example, Bley-Vroman and Chaudron (1994) demonstrated that learners' performances on elicited imitation tasks were systematically influenced by stimulus length and serial order effects (see also Chaudron and Russell, 1990). Thus, depending on both the length of the sentence to be repeated and the placement within the sentence of targeted structures, learners would either correctly or incorrectly repeat the structure to be measured. Numerous other characteristics of measurement tasks may introduce systematic variability into the performances elicited from learners, including characteristics of the measurement setting, the communicative or linguistic context, and task instructions and formatting (see extensive treatment in Bachman and Palmer, 1996; R. Ellis, 1994; Loschky and Bley-Vroman, 1993; Norris et al., 1998; Tarone, 1998; Wolfram, 1985; Yule, 1997).

In sum, SLA researchers will need to conceptualize carefully the link between intended construct interpretations and the behaviors selected to provide evidence about them. Recent empirical and theoretical approaches to cognitive task analysis should prove helpful in conceptualizing the cognitive demands made by characteristics of measurement tasks and the ways in which learners deal with such demands during task performance (e.g., Baxter and Glaser, 1998; Mislevy et al., 1999; Nichols and Sugrue, 1999; Royer et al., 1993; Sugrue, 1995). More fundamentally, measurement for SLA research purposes would be well served by adopting an *evidence-centered* approach to the design of instruments and procedures. Bennett (1999) summarizes evidence-centered design as the process of "identifying the evidence needed for decision making in terms of some complex of student characteristics, the behaviors or performances required to reveal those constructs, and the tasks needed to elicit those behaviors" (p. 5). Recent work on the application of evidence-centered design principles to educational and occupational assessment problems offers detailed and useful examples of this process (e.g., Mislevy et al., 1999, forthcoming).

2.2.3 *Specifying meaningful qualities of behavior*

Even if behaviors to be elicited in measurement are carefully selected in order to provide adequate evidence for intended construct interpretations, construct underrepresentation remains a threat unless the variable qualities of behaviors are specified in units of analysis which are sensitive to the intended interpretations. Chaudron (1988) has pointed out, "when we test hypotheses with a

quantitative method, we have derived them from qualitative, conceptual considerations. Before we count, we have to decide what categories to count" (p. 16). In SLA research, meaningful categories may include, among others: (i) frequency or amount of behaviors; (ii) duration of behaviors; (iii) sequences of behaviors; (iv) combinations of behaviors; and (v) comparisons of one sort of behavior with others. Each of these approaches to synthesizing behavioral observations requires a corresponding scale with units that match the scope of intended interpretations (e.g., counting milliseconds, seconds, or minutes will obviously affect the level at which chronometric research findings may be discussed; see related problems in Siegler, 1989). In addition, it may frequently be the case that a single set of scales/units will prove insufficient for capturing the complexity of construct interpretations. For example, while "error" counts may offer evidence for interpretations about the extent of a learner's knowledge, they will do little in the way of informing interpretations about the cognitive resource demands or expertise in performing a task using that knowledge, especially when "improvements in skilled performance continue long after errorless performance is achieved" (Royer et al., 1993, p. 210). Therefore, conceptualizing the variable qualities of elicited behaviors in construct-meaningful ways will prove critical for maintaining construct validity during the scoring and analysis of measurement outcomes.

Interlanguage analysis techniques, typically carried out within interactionist approaches to SLA, offer a useful example of problems which researchers encounter when criterial qualities of behavior do not match the scope of intended interpretations. For example, Pica (1983) found that the application of different levels of analysis to the same interlanguage performance data "resulted in two different interpretations regarding the role of L2 exposure conditions in second language acquisition" (p. 73). Pica compared accuracy results from the measurement of suppliance in obligatory contexts (SOC; see R. Brown, 1973) with results from the measurement of target-like use (TLU), a technique developed to account for oversuppliance errors. She found that the results of TLU analyses, but not of SOC, revealed a marked tendency among instruction-only learners to oversupply certain morphemes, a tendency which was absent in the L2 performance of naturalistic learners. Further TLU analyses based on types (where only different word types were counted for accurate use), but not based on tokens of the same data (where each word token was entered into the accuracy count), revealed that naturalistic learners and instruction-only learners had a smaller expressive vocabulary and used the English plural morpheme with fewer word types than instruction-plus-exposure learners. Had the data been subjected solely to SOC and token-based TLU analyses, these two patterns would have gone undetected. Another illustration of how increased sensitivity of analytical units and procedures may contribute to a better understanding of the behaviors of interest within a given theory is found in Oliver (1995). In her study of the provision of negative feedback during task-based interactions, Oliver observed that only 10 percent of recasts produced by English native-speaking children during interactional exchanges

were incorporated by their ESL interlocutor peers. However, when she introduced a finer level of analysis for NNS third turns in recast episodes, by adding to her coding scheme the category "no opportunity to incorporate" (due to discursive-pragmatic constraints on turn-taking), she found that over one-third of all recasts were incorporated.

A particularly thorny issue in interlanguage research is adjudication of the extent to which accuracy in production of L2 forms should be taken as reflective of IL development. An early caution against accuracy as a viable criterion for L2 acquisition (as traditionally established in L1 acquisition studies by R. Brown, 1973) was advanced by Meisel et al. (1981; see also Pienemann, 1998). These authors argued that emergence, defined as the first documented occasion of productive (i.e., non-formulaic) use of a given form, is the most IL-sensitive approximation for measuring development. Likewise, measures of grammatical accuracy have difficulty accounting for attested IL developmental phenomena, such as threshold and stage-related effects (Meisel et al., 1981), flooding (Huebner, 1983), and U-shaped behavior (Kellerman, 1985), all of which can obscure interpretations. Additional qualities of interlanguage development have been proposed which may further constrain interpretations based on grammatical accuracy. For instance, Wolfe-Quintero, Inagaki, and Kim (1998, pp. 73–4) have suggested a phenomenon that they call omega-shaped behavior, referring to a temporary increase in the frequency of (possibly less-than-accurate) suppliance of a recently emerged form, followed by a normalization in rate of suppliance, once the new form has been worked out by the learner. Another underexplored quality of learner L2 production is the gradual extension of suppliance of a form from a few simple contexts to a wider range of (possibly more complex) contexts (see Richards, 1990, on L1 acquisition; and Pishwa, 1994, on L2 acquisition).

What the existence of such interlanguage processes and phenomena suggests is that curvilinear rather than linear relationships can be expected between accuracy in producing a given L2 form and IL development of that form. These curvilinear relationships need to be taken into account when conceptualizing criteria for behavioral qualities and when planning analyses of L2 performance, as they will certainly affect the interpretations that follow. An IL analytical approach that combines emergence and accuracy (of the same form or of related forms) may prove more informative and useful than an exclusive focus on emergence or, no doubt, on accuracy. For example, by combining analyses of emergence and accuracy in a longitudinal corpus, Bardovi-Harlig (1994b) was able to establish that the emergence of initial instances of past perfect marking in L2 English was dependent upon learners reaching a reasonable level of stability (i.e., productive accuracy of around 85 percent SOC) in the marking of past tense morphology. In the end, the most desirable approach, particularly with longitudinal data, may be to adopt a three-step coding process which gauges: (i) first suppliance (or emergence), (ii) non-target-like but more sustained suppliance (frequency of functional contexts attempted), and (iii) target-like suppliance at optimal ultimate levels of attainment (accuracy).

This multifaceted approach to characterizing qualities of behavior might most precisely reflect the gradual processes in IL development that many SLA researchers are interested in mapping (see Stromswold, 1996, for similar methodological suggestions in L1 acquisition research).

These examples of interpretive problems arising in interlanguage analysis underscore what should be a fundamental concern for SLA researchers who utilize measurement data. That is, for all measures, researchers should be able to demonstrate how a particular type and level of behavioral analysis enable construct-relevant interpretations to be made. What does it mean for a learner to score 60 percent correct on a post-test as compared with 50 percent correct on the pre-test? What does an observed difference in "amount of interaction" have to do with differences in acquisition? How can similar reaction times in sentence-matching tasks from advanced and novice learners be explained? What does an "incorrect" answer on a grammatical acceptability item tell us about the learner's internal grammar? How does frequency of "errors" in a written narrative offer insights into a learner's developing interlanguage? Where basic questions like these about the qualities of observed behaviors cannot be answered, researchers will remain "unenlightened" about the meanings attributable to measurement outcomes (Chaudron, 1988; Schachter, 1998), and construct interpretations will remain unwarranted. It should also be obvious from the examples above that the only source for answers to such questions, and the basis for establishing meaningful qualities of measurement behaviors, resides in empirical knowledge that has been accumulated about the acquisition-related behaviors of interest. In this regard, as has been recommended for measurement in other domains of inquiry (e.g., the measurement of automatization in cognitive processing; Royer et al., 1993), there is an obvious increased role to be played in SLA research by descriptive longitudinal studies which establish norms of performance for particular processes and phenomena in L2 acquisition (e.g., Ortega, 2000). Indeed, attempting to "measure" acquisition without a sound descriptive basis for meaningful differences in particular acquisition-related behaviors would be akin to timing a runner's performance over a mile without knowing how many times around the track a mile happens to be.

2.2.4 The "valid test" fallacy

From time to time, SLA researchers adopt measures employed in previous studies, or in other non-research contexts, for the purposes of their own investigations. In itself, repeated use of identical measurement instruments/procedures for measuring the *same construct(s)* is a fundamentally worthwhile endeavor. As Norris and Ortega (2000) have pointed out for studies of L2 instructional effectiveness, it is only through such exact replication (e.g., by measuring the same dependent variable) across research settings that trustworthy findings about a given variable may begin to accumulate (see also Bangert-Drowns, 1986; Cohen, 1997; Light and Pillemer, 1984; Rosenthal, 1979). However, when SLA researchers adopt pre-existing measures wholesale, simply

because they seemed valid in other studies or measurement contexts, the researchers are guilty of the "valid test" fallacy.

In such cases, researchers or other measurement users mistakenly assume that validity is a property of test instruments and procedures, rather than the uses that are made of them. As the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) make clear, validation is a process of gathering evidence and theoretical arguments supporting the *use* of test scores for particular interpretations and related purposes. As such, the *Standards* emphasize, "When test scores are used or interpreted in more than one way, each intended interpretation must be validated" (p. 9). If SLA researchers assume that a given measure is a "valid" indicator of acquisition (or learning or proficiency or knowledge or aptitude, etc.), then apply that measure to their own situated purposes, without taking the time to establish the link between behavioral evidence provided by the measure and their own intended constructs, the validity of resulting interpretations will be threatened (see related discussion in Messick, 1989; Thompson, 1998).

For example, Shohamy (1994) observed that tests intended for educational decision making, such as the Test of English as a Foreign Language (TOEFL), are frequently utilized by SLA researchers as measures of learning or acquisition, even though such tests were designed as indicators of global academic language abilities. Likewise, holistic proficiency measures, such as the ACTFL (1986) *Guidelines* and related procedures, may be used as a basis for assigning learners to instructional research conditions, even though the scores on such measures may have nothing to do with the particular L2 forms or abilities being investigated (see discussion in Norris, 1996, 1997; Young, 1995b). The "valid test" fallacy applies equally to so-called "objective" measures, such as those used in analyzing spoken or written L2 performance (e.g., accuracy, complexity, and fluency measures; see Polio, 1997; Wolfe-Quintero et al., 1998), when researchers misguidedly assert or hope that such units of analysis will be "valid" for all reasons (Foster, Tonkyn, and Wigglesworth, 1998). We are not suggesting that for every research study new measures need to be developed; this would only serve to limit generalizability of findings and hinder the accumulation of knowledge. We are suggesting that SLA researchers need to conceptualize carefully their constructs and the evidence that will be brought to bear on them, and then match these conceptual bases with corresponding instruments and procedures, in order for each occasion of measurement use to inform warranted interpretations.

3 How Should Acquisition be Counted? Procedural Concerns for Measurement in SLA

Given adequate conceptualization of what counts as acquisition, the mechanics of measurement may take place, following several procedural stages (4–6 in section 1.1): (i) selected tasks/situations are employed to elicit behaviors;

(ii) meaningful qualities in observed behaviors are summarized in the form of scores; and (iii) scores are analyzed to produce evidence for intended interpretations about acquisition. As conceptual decisions are translated into practice, the particular actions that are taken by researchers may influence resulting interpretations. Such unintended or unsystematic sources of variance which issue from the act of measurement itself can be summarized under the heading of measurement error. The fundamental construct validity question for these procedural stages, then, asks to what extent patterns in the behavioral data which are actually elicited, scored, and analyzed can be attributed to the construct interpretations that researchers want to make, as opposed to construct-*irrelevant* variance due to measurement error.

There are numerous approaches to developing and using measures which may help to reduce the influence of measurement error. For practical guides, readers are referred to several sources directly related to applied linguistics (e.g., Bachman, 1990; J. D. Brown, 1996; Hatch and Lazaraton, 1991; Henning, 1987; Scholfield, 1995; Woods, Fletcher, and Hughes, 1986) as well as to the educational and psychological measurement literature (e.g., Anastasi and Urbina, 1997; Gronlund and Linn, 1990; Linn, 1989; Orwin, 1994; Pedhazur and Schmelkin, 1991; Popham, 1981; Traub, 1994). Our purpose in the current section is briefly to address a few of the most critical concerns associated with the proceduralization of measurement in SLA research, and to suggest directions in research practice which might help to reduce the threat of construct-irrelevant variance due to measurement error.

3.1 Reliability in elicitation and scoring

Reliability reflects the extent to which a measure leads to *consistent* interpretations about a particular construct on each measurement occasion. Such consistency is traditionally viewed (e.g., Traub, 1994) as the relationship between an observed score or any quantified outcome of measurement, the amount of that observed score which is attributable to the construct of interest, and the amount of observed score which is attributable to measurement error: observed score = true score + error. As behavioral data are elicited and scored, varying amounts of error may be introduced from a number of sources, including: (i) environmental factors associated with the data-collection or test-administration context; (ii) data-collection or test-administration procedures; (iii) characteristics of items or other components of the measurement instrument; (iv) data-coding or test-scoring procedures; and (v) idiosyncrasies of research participants, such as interest, attention, and motivation (see J. D. Brown, 1996; Traub, 1994). Obviously, the greater the influence of such error types, the less reliable measurement outcomes will be (i.e., the less an observed score on a measure will represent a learner's true score vis-à-vis the construct). In order for measurement-based SLA research to inform warranted interpretations, such sources of measurement error should be reduced where possible. It is also essential to observe, analyze, and report reliability and error

for each use of a measure, as indicated by the American Psychological Association task force on statistical inferencing: "[A]uthors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores" (Wilkinson and the Task Force on Statistical Inference, 1999, p. 596). Furthermore, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) hold that reports of reliability should include discussion of: (i) the operationalization and administration of instruments and procedures; (ii) the development and use of scoring or coding schemes; (iii) the training of coders or raters; (iv) the performance of coders or raters; (v) the characteristics of participants or populations; and (vi) the characteristics of scores.

3.1.1 Error in behavior elicitation

The behavior elicitation and observation stage of measurement is particularly susceptible to the introduction of error, owing to the multitude of factors to be considered in order to maintain procedural consistency (see, e.g., the detailed list in J. D. Brown, 1996, p. 189). On the one hand, researchers must ensure that all critical aspects of tasks or situations are faithfully translated into measurement instruments and procedures as conceptualized, such that the scope of behaviors and behavioral qualities may be fully captured. For example, in research on developmental sequences in L2 syntax and morphology, the design of measurement tasks must reflect a number of considerations in order to elicit consistent behavioral patterns. Because initial emergence of particular syntactic and morphologic forms is posited to be implicationally related with the preceding or subsequent emergence of other forms, behavioral data must be gathered across a variety of linguistic contexts. Furthermore, given the fact that initial emergence of a form may occur in different communicative contexts for different learners, behavioral data must be gathered using a variety of communication tasks (or in a variety of situations). In light of such evidentiary requirements, it is only through the elicitation of extensive amounts and types of L2 behaviors that measurement can show that particular forms have emerged, that implicationally preceding forms have also emerged, and that subsequent forms have not emerged. If measurement tasks fail to provide the range of linguistic and communicative contexts necessary for patterns in emergence to be displayed, then interpretations about learners' developmental stages will remain inconclusive at best (see discussion in Clahsen, Meisel, and Pienemann, 1983; Hudson, 1993; Pienemann, 1998; Pienemann, Johnston, and Brindley, 1988; Pienemann and Mackey, 1993).

On the other hand, researchers must also be wary of potentially unpredictable sources of error that may be associated with features of the measurement context, measurement forms or instructions, individual learners, etc. For example, for SLA research which seeks to make interpretations based on learners' oral L2 discourse, characteristics of the interlocutor as well as particular actions undertaken by the interlocutor may unpredictably influence a

learner's L2 performance. Research on oral interview types of language tests, wherein one or more examinees interact with one or more interlocutors, has demonstrated that such characteristics as gender and age of the interlocutor may substantially affect the amount and quality of language produced by the examinee (e.g., McNamara and Lumley, 1997; O'Sullivan, 2000). Likewise, the particular activities engaged in by interlocutors (especially interviewers), such as discourse accommodation, have been demonstrated to influence what an examinee says and how it is said (e.g., Lazaraton, 1992, 1996; Ross and Berwick, 1990; Young, 1995a; Young and He, 1998; Young and Milanovic, 1992).

In order to reduce the effect of these and many other problems that may emerge during the elicitation of behaviors for measurement purposes, there is much to be said for following systematic methods in the production of tests and other procedures, and especially for careful pilot-testing and revision of instruments, directions, and administration guidelines (see Bachman, 1990; Bachman and Palmer, 1996; J. D. Brown, 1996, forthcoming; Campbell and Reichardt, 1991; Lynch and Davidson, 1994; Popham, 1981). Recent developments in measurement theory and technology may also prove useful in this respect, for example, in the form of computerized item-generation capabilities (e.g., Irvine and Kyllonen, 2001).

3.1.2 Error in scoring

Even if they are consistently elicited and observed, measurement behaviors on their own are typically insufficient for enabling intended interpretations; hence, they are almost always summarized or scored in light of particular qualities which are relevant to the L2 acquisition constructs. Measurement error may also be introduced during this scoring process. First, the particular scoring procedures employed by researchers may serve as sources of error, if they are not consistently carried out. Second, important qualities of measurement behaviors may be distorted or obscured by characteristics of the scores that have been selected to represent them.

The coding of learners' spoken or written L2 production for patterns in interlanguage development offers a good example of the possible sources of error which may be introduced during measurement scoring. Such "interlanguage coding" involves the subjective application of particular criteria by raters or coders in order to identify various attested or predicted phenomena within learner performances, such as: (i) target-like grammatical accuracy of syntactic or morphologic forms; (ii) lexical range, density, and diversity; (iii) rate of speech, number and length of pauses, hesitations, and other features of fluency; (iv) range, length, and suppliance of various clausal types; and (v) length, amount, and frequency of various semantic and/or phonological units (see overviews in Crookes, 1990, 1991; Norris, 1996; Ortega, 1999, 2000; Polio, 1997; Richards and Malvern, 1997; Skehan, 1998; Wolfe-Quintero et al., 1998). Typically, in coding for these and related phenomena, individual coders work through recordings, transcripts, or written products, identifying and marking the phenomena in question as they go. A number of

problems may occur during this coding process which introduce error into the resulting scores. Coders may be insufficiently knowledgeable of, or trained to recognize, the IL phenomenon in the first place, or the phenomenon may be defined so poorly within the research domain as to defy accurate coding of complex data (e.g., the coding of utterances or T-units for spoken discourse, as Crookes, 1990, and Foster et al., 1998, have pointed out). When working with a lengthy corpus, coders may become fatigued, frustrated, or bored. Over time, they may "drift" in their assessments of how a phenomenon is realized in the data. Finally, coders may be biased to identify or ignore the particular IL phenomena that they are investigating. Each of these problems can cause coders to miscode, or simply miss, characteristics of the behavioral data which have been elicited. In order to minimize the impact of such coding problems, a systematic series of error-reduction strategies (Orwin, 1994) can be employed, including the careful development and pilot-testing of coding protocols, the sufficient training of coders, the use of multiple codings of the same data, and the scheduling of coding rounds in a staged fashion to minimize coder drift (e.g., Ortega, 2000). In addition, the periodic and overall calculation of intercoder agreement coefficients will enable the identification and reduction of coder error, as well as provide evidence regarding the extent to which such error influences the final scores attributed to individual learners.

Once codings are completed, they are tallied and converted into numerical scores which represent the interlanguage phenomena in various ways (number of pauses, number of different clause types per total number of clauses, target-like forms supplied in obligatory contexts, etc.). Of course, simple miscounts of the codings or miscalculations of comparisons among them will distort the actual behaviors observed, although the mechanization of counting and calculating can greatly reduce such error (e.g., MacWhinney, 2000). At the same time, the index or scale selected for scoring may itself introduce error into eventual interpretations. For example, a host of reliability problems have been associated with discrepancies between scores, the overall size of a corpus and variable text length within a corpus, and intended interpretations (see discussions in Biber, 1990; Bley-Vroman, 1983; Richards, 1994). Simple raw frequency counts of a phenomenon (e.g., number of relative clauses) can prove problematic when scores are to be compared among different learners' texts, because lengthier texts increase the likelihood that a given phenomenon will be observed more frequently (Richards, 1994). Thus, general learner *productivity* may serve to confound interpretations about a learner's use or knowledge of a given L2 form. In addition, the exclusive reporting of raw frequencies makes it difficult to compare results across studies yielded by total corpora of differing lengths (Wolfe-Quintero et al., 1998).

One solution favored by many researchers is to convert frequency tallies into ratio scores (e.g., words per second, clauses per T-unit, unique lexemes per total lexemes, etc.). However, ratios are not impervious to reliability problems associated with the size of a corpus and the relative size of the texts (or samples) which comprise it. For example, the lexical type–token ratio (number

of lexical types per total number of lexical tokens) has been shown repeatedly to have a non-linear, and often negative, relationship with corpus size and to be a very unstable score when text samples of varying lengths are compared (see Hess, Sefton, and Landry, 1986; Richards, 1987). This instability occurs because closed-class words as well as high-frequency words are likely to be repeated increasingly in extended production by a given learner, while new words are progressively less likely to be used (i.e., relative to the other words). Thus, shorter samples tend to display inflated type–token ratios relative to longer samples. As a solution to this productivity bias, it has been suggested that a minimum standardized length of 300 tokens (e.g., words, T-units, etc.) per sample may be necessary for lexical ratios to stabilize (see Hess et al., 1986). However, perhaps the most accurate, if somewhat more computationally demanding, approach to resolving such problems has been proposed by Richards and Malvern (1997), who have shown that a statistical model of lexical diversity better reflects lexical differences among learners. Such modeling of multiple sources of variance in observed behaviors may be the only means for accurately summarizing interlanguage codings in a way that is adequately consistent and relevant to intended construct interpretations.

These examples underscore the extent to which error may be introduced into measurement through the scoring process. Among other problems (e.g., violation of a cardinal assumption for statistical inferencing), resulting low reliability in measurement scores can cloud outcomes to the point that findings are not interpretable or actual relationships and effects are not detected. As such, it is essential that researchers seek to understand the error involved in each use of a measure. Along these lines, Thompson (1994) has emphasized:

> The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. For example, we may conduct studies that could not possibly yield noteworthy effect sizes, given that score reliability inherently attenuates effect sizes. Or we may not accurately interpret the effect sizes in our studies if we do not consider the reliability of the scores we are actually analyzing. (p. 840)

There are several major theoretical approaches to, and numerous techniques for, estimating the amount and type of error in measurement scoring and scores. In addition to classical test theory approaches and techniques (e.g., J. D. Brown, 1996; Traub, 1994), developments in reliability theory over the past several decades have led to a much more sophisticated understanding of how and to what extent error may be influencing scores. Item response theory and associated computerized analyses (e.g., Linacre, 1998), which focus on the probabilities of various score patterns, not only enable the calculation of learner ability and task difficulty estimates according to a single true interval scale, but also allow for the estimation of error associated with each individual score point, as opposed to the traditional and much less informative single reliability estimate for an entire set of scores (see discussion in Embretson and

Hershberger, 1999; Hambleton, Swaminathan, and Rogers, 1991). An even more thorough understanding of the amount of error contributed to scores by each of any number of different sources (raters, tasks, forms, examinee populations, etc.) may be achieved through the use of generalizability theory and related techniques (e.g., Marcoulides, 1999; Shavelson and Webb, 1991). Of course, while more sophisticated approaches to reliability estimation will help researchers better understand the extent to which error is affecting their measurement scores, it is only through improvements in scoring practices that researchers will be able to reduce the influence of error on their eventual interpretations about acquisition (see discussion of innovations in test scoring methods in Thissen and Wainer, 2001).

3.1.3 Reporting reliability of measurement scores

A major concern which directly influences the interpretability of SLA research findings and the accumulation of trustworthy knowledge about acquisition constructs is the fact that reliability and error in measurement scoring are at best infrequently considered and only inconsistently reported. For example, Norris and Ortega (2000) found that only 16 percent of 77 studies on the effectiveness of L2 instruction, published between 1980 and 1998, reported any kind of reliability information for scores on dependent variable measures. Similarly, in a review of 39 studies of L2 writing research, published between 1974 and 1996, Wolfe-Quintero et al. (1998) observed that only 18 percent reported any information about the reliability of procedures used to measure accuracy, complexity, and fluency in written performance data. In smaller-scale reviews of more recent bodies of L2 research (e.g., 10 planning studies reviewed in Ortega, 1999; 16 writing studies reviewed in Polio, 1997; 10 recent SLA studies surveyed by Shohamy, 2000), findings show that at best only half of the studies addressed reliability, and that most researchers reported only global or averaged reliability estimates without specifying, let alone discussing, the indices employed or the particular sources for error (this is not a phenomenon unique to SLA or applied linguistics research; see Royer et al., 1993; Vacha-Hasse, Ness, Nilsson, and Reetz, 1999; Whittington, 1998).

The failure to estimate, report, and discuss reliability and error may generate several problems for SLA research. First, unless reliability or error estimates are reported, individual study findings will be uninterpretable, because it will remain unclear to what extent measurement outcomes reflect the construct of interest versus other unintended sources of variance. Second, as Hunter and Schmidt (1994) have pointed out, unless reliability estimates are reported in individual studies, the influence of measurement error on a range of findings accumulated from studies which investigate the same variable cannot be understood. As such, syntheses of an overall effect or relationship observed across studies will be less accurate, because correction for overall score attenuation due to error will be impossible. Third, without accurate reporting of the *sources* of error influencing score reliability, as well as the *amount* of error involved, systematic efforts at reducing measurement error in future studies will be

hindered. Where reliability of measurement scores *is* consistently reported within a domain of inquiry, there may be unique possibilities for researching and better understanding the amounts and sources of error associated with particular measures, scoring procedures, learner populations, and features of measurement contexts. Vacha-Hasse et al. (1999) propose the notion of "reliability generalization," a meta-analytic method for combining the reliability results of the use of similar dependent variables across a range of studies in order to make interpretations about sources of measurement error associated with such measures and measurement contexts.

3.2 Analyzing measurement scores

SLA researchers employ a variety of analytic techniques (statistical inference, implicational scaling, correlational analyses, statistical modeling, etc.) to summarize, compare, and interpret scores in light of research questions, hypotheses, and predicted relationships among and between variables, thereby completing the transformation of measurement-based data into evidence. Because appropriate analyses are determined in part by the particular research questions and methods of a study, their selection falls within the scope of overall research design (see Chaudron, this volume) and is not a concern isolated to measurement per se. Nevertheless, it is often the case that measurement data are further manipulated within such analyses; thus, the link between behavioral evidence and intended interpretations is also susceptible to construct-irrelevant variance at this stage in the measurement process. In this section, we highlight a few examples of analytic problems in measurement-based SLA research.

A most basic problem involves the selection of analytic tools which may be inappropriate for the particular kinds of interpretations to be made. For example, Paolillo (2000) demonstrated how response patterns on grammaticality judgment tasks (GJTs) can lead to spurious findings (which then become reified within the research community), owing to the application of statistical analyses which are insufficiently sensitive to the actual range and sources of variance in elicited behavioral data. Paolillo (2000) first showed how a chi-test for independence, which has been recommended as the appropriate statistical approach to analyzing GJTs (Bley-Vroman, Felix, and Ioup, 1988), is incapable of disentangling whether GJT response patterns are due to: (i) a systematic (and UG-predicted) interaction between the correctness of learners' judgments and the grammaticality or ungrammaticality of items (an asymmetry effect); or (ii) simple indeterminacy in learners' responses. Paolillo then employed a multivariate analysis (logistic regression) to reveal a more complex relationship in response patterns than that which had been predicted; namely, in the particular data set he was studying, GJT behaviors were best modeled as an interaction between learner conservatism (i.e., a tendency to judge items as ungrammatical), the types of grammatical constructions being measured, and target grammaticality norms for these items, in addition to the UG-predicted

asymmetry effect. Paolillo concluded by emphasizing that all such potential effects on GJT response patterns "need to be examined and factored into the explanation of the data in order to arrive at the intended UG-based interpretation" (2000, p. 223).

As Paolillo demonstrates, for certain approaches to SLA the application of multivariate statistics and related analyses can help to clarify exactly what measurement data may reveal about constructs. At the same time, in much of the research on L2 acquisition, there is a virtually default practice of utilizing inferential statistics for all analytic purposes. Unfortunately, the "quest" for statistical significance may actually obscure what measurement data have to say, especially when: (i) the use of inferential statistics leads to insufficient reporting of other forms of measurement data; (ii) the results of statistical analyses are inaccurately interpreted; and (iii) studies are not adequately planned to meet the basic assumptions for such techniques. For example, in their review of 77 studies on L2 instructional effectiveness, Norris and Ortega (2000) found that researchers were more likely to report the outcomes of inferential statistical analyses than basic descriptive statistics, such as means, standard deviations, and number of test items, even though the latter provide the only direct indication of the behavioral patterns that were actually observed on measures. Norris and Ortega also found that researchers frequently interpreted the results of statistical significance tests to be indicative of the *magnitude* of effects or relationships observed via measurement, as opposed to the probability levels associated with particular observations, and that research designs and measurement data types often violated the assumptions of the statistics being used. One consequence of these problems in the reporting and interpretation of inferential statistics is that meaningful patterns in measurement scores may be obscured to the point that accurate interpretations about intended constructs are no longer feasible (see related discussion in Carver, 1978; Cohen, 1988, 1990; Cooper, 1998; Cooper and Hedges, 1994; Harlow, Mulaik, and Steiger, 1997; Light and Pillemer, 1984; Rosenthal, Rosnow, and Rubin, 2000; Rosnow and Rosenthal, 1989; Wilkinson and the Task Force on Statistical Inference, 1999).

In order for researchers to understand what measurement data have to say about their research questions and hypotheses, they will need to know what analyses are available, what kinds of analyses are appropriate for what kinds of data, and how to interpret and report the outcomes of these analyses. In this regard, and in light of the propensity of SLA researchers to employ inferential statistics, any of the available treatments of standard univariate and multivariate statistical analyses would be a good place to start (e.g., Tabachnik and Fiddell, 1996; Woods et al., 1986). At the same time, the potential role to be played by alternative analytic tools should be further explored, as these may offer more direct and appropriate means for summarizing and understanding what measurement data have to say. For example, analytic approaches to research designs which have inherently small data sets (Hoyle, 1999), as well as analyses appropriate for longitudinal, multiwave studies (Willett, 1988), may prove

particularly useful for many SLA research studies. Furthermore, the potential analytic role to be played by simple effect sizes, confidence intervals, and graphic displays should not be overlooked (e.g., Cooper, 1998; Light and Pillemer, 1984; Rosenthal et al., 2000). Finally, it will be critical for researchers to pay closer attention to the nature of measurement scores and the ways in which various score types may interact with particular analytic tools. For example, problems with the use of raw scores from tests and other measures in parametric statistical analyses have begun to be widely discussed, in light of the fact that raw scores never provide the true interval data, or equal reliabilities for all score points, that are assumed by such analyses (see related discussions in Embretson and Hershberger, 1999).

4 Making it Count: Accumulating Measurement-Based Knowledge

As a concluding stage in the measurement process, final construct interpretations are made on the basis of the evidence provided, and research findings are integrated by primary and secondary researchers into the cumulative knowledge of the domain of inquiry. The extent to which these construct interpretations will contribute warranted and relevant knowledge to theories of SLA will depend on how well researchers have countered threats to construct validity at each of the stages in measurement practice (see figure 21.1). In particular, we have raised several fundamental weaknesses in the conceptualization and proceduralization of measurement in SLA which will demand attention. First, SLA researchers must acknowledge that a single measure will not provide sufficient evidence for informing the range of interpretations typically sought in most SLA studies and that theories which posit cognitive constructs will need to incorporate means for observing the full range of these constructs, not simply the language performance outcomes attributed to them. Second, serious efforts will need to be made by SLA researchers in order to develop the empirical knowledge bases required for understanding what observed behaviors may tell us about acquisition in the first place; this implies much broader implementation of descriptive, longitudinal studies of various L2 acquisitional phenomena. Third, measurement error will continue to play an unknown role in most measurement-based SLA research until researchers begin to report appropriate reliability estimates and to consider the various sources for error in their measures. Fourth, SLA researchers need to recognize that inferential statistics do not provide the only, and in many cases do not provide the appropriate, analytic tools for understanding measurement scores and incorporating scores into research findings. Fifth, it will be crucial for SLA researchers who intend to utilize measurement as a primary research tool to be trained in the fundamentals of measurement, so that they may attend to advances within measurement theory and practice which are of direct relevance to their own methods. Finally, researchers and editors alike will need to recognize that

much more explicit and thorough reporting of all phases of measurement practice will be necessary for the accumulation of scientifically worthwhile knowledge about SLA to be possible.

Within the language testing field, it has been suggested for some time now that a research priority should be the development of comprehensive programs of validation for the various intended uses of language ability tests (Bachman, 1989; Bachman and Clark, 1987; Bachman and Cohen, 1998). We would suggest that validity *generalization* of this sort (see also Wiley, 1991) should also be a priority for measurement used within SLA research and should constitute the site of true collaboration between language testers or measurement specialists and measurement-informed SLA researchers. It should not be incumbent on the individual researcher alone to pursue a comprehensive program of measurement development, use, and validation for each construct interpretation to be made (indeed, as Messick, 1989, has suggested, this would be virtually impossible). Rather, we believe that where entire SLA research communities engage in a comprehensive approach to all of the stages in the measurement process, the field will find itself much better able to make theoretically meaningful interpretations about its constructs and to pursue the accumulation of scientifically worthwhile knowledge.

NOTES

1 A particular cognitive theory within interactionist SLA must be singled out here because of some notable differences. Skills acquisition theories (e.g., DeKeyser, 1997) argue that fast, accurate, and effortless application of L2 knowledge to novel cases provides evidence of true learning. Further, interpretations about automatization are central to this type of theory, and automatization is thought to be typically reflected in "gradual drop-offs in reaction time and error rates, and diminished interference from and with simultaneous tasks" (DeKeyser, 1997, p. 196). Thus, skills acquisition studies are more likely than other interactionist studies to include measures of reaction times and nativelike accuracy over multiple trials in order to document changes in speed and accuracy of rule application to novel cases. From this theoretical perspective, *acquired* means fast, accurate, and effortless performance that reflects automatized production and/or comprehension resulting from sufficient practice guided by declarative knowledge (i.e., conceptually driven learning). It is important to note that, although the similarities with emergentist-connectionist theories are striking, the theoretical models of learning that are assumed in skills acquisition theory and in emergentism are radically different.

2 Readers will note that many of the measurement examples we employ throughout this chapter are typically associated with interactionist approaches to SLA research. This unbalanced treatment simply reflects

our own research backgrounds and training; we do not wish to suggest that measurement in interactionist SLA is either particularly problematic or particularly effective relative to other epistemologies and associated measures. Naturally, we hope that readers will be able to generalize from our examples to their own measurement applications and problems.

REFERENCES

Alanen, R. 1995: Input enhancement and rule presentation in second language acquisition. In R. Schmidt (ed.), *Attention and Awareness in Foreign Language Learning and Teaching*. Technical Report No. 9. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, 259–302.

ACTFL (American Council on the Teaching of Foreign Languages) 1986: *Proficiency Guidelines*. Yonkers, NY: ACTFL.

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

Anastasi, A. and Urbina, S. 1997: *Psychological Testing*. Upper Saddle River, NJ: Prentice-Hall.

Andersen, R. 1984: The one-to-one principle of interlanguage construction. *Language Learning*, 34, 77–95.

Anderson, R. J. 1993: *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Angoff, W. H. 1984: *Scales, Norms and Equivalent Scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H. 1988: Validity: an evolving concept. In H. Wainer and H. I. Braun (eds), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 19–32.

Bachman, L. F. 1989: Language testing–SLA research interfaces. *Annual Review of Applied Linguistics*, 9, 193–209.

Bachman, L. F. 1990: *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Clark, J. L. D. 1987: The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science*, 490, 20–33.

Bachman, L. F. and Cohen, A. D. 1998: Language testing–SLA interfaces: an update. In L. F. Bachman and A. D. Cohen (eds), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 1–31.

Bachman, L. F. and Palmer, A. S. 1996: *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bangert-Drowns, R. L. 1986: Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388–99.

Bard, E. G., Robertson, D., and Sorace, A. 1996: Magnitude estimation of linguistic acceptability. *Language*, 72, 32–68.

Bardovi-Harlig, K. 1994a: Anecdote or evidence? Evaluating support for hypotheses concerning the development of tense and aspect. In

E. Tarone, S. Gass, and A. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 41–60.

Bardovi-Harlig, K. 1994b: Reverse-order reports and the acquisition of tense: beyond the principle of chronological order. *Language Learning*, 44, 243–82.

Baxter, G. and Glaser, R. 1998: Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17 (3), 37–45.

Bennett, R. E. 1999: Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, 18 (3), 5–12.

Beretta, A. 1991: Theory construction in SLA: complementary and opposition. *Studies in Second Language Acquisition*, 13, 493–511.

Bialystok, E. 1991: Achieving proficiency in a second language: a processing description. In R. Philipson, E. Kellerman, L. Selinker, M. S. Smith, and M. Swain (eds), *Foreign/Second Language Pedagogy Research*. Clevedon: Multilingual Matters, 63–78.

Biber, D. 1990: Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5, 257–69.

Birdsong, D. (ed.) 1999: *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bjork, E. and Bjork, R. (eds) 1996: *Memory: Handbook of Perception and Cognition*. 2nd edition. New York: Academic Press.

Bley-Vroman, R. 1983: The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning*, 33, 1–17.

Bley-Vroman, R. 1989: What is the logical problem of foreign language acquisition? In S. M. Gass and J. Schachter (eds), *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press, 41–68.

Bley-Vroman, R. and Chaudron, C. 1994: Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass, and A. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 245–61.

Bley-Vroman, R., Felix, S., and Ioup, G. 1988: The accessibility of Universal Grammar in adult language learning. *Second Language Research*, 4, 1–32.

Brindley, G. 1998: Describing language development? Rating scales and SLA. In L. Bachman and A. Cohen (eds), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 112–40.

Brown, G., Malmkjaer, K., and Williams, J. (eds) 1996: *Performance and Competence in Second Language Acquisition*. Cambridge: Cambridge University Press.

Brown, J. D. 1988: *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research Design*. London: Heinemann.

Brown, J. D. 1996: *Testing in Language Programs*. Englewood Cliffs, NJ: Prentice-Hall.

Brown, J. D. forthcoming: *Using Surveys in Language Programs*. Englewood Cliffs, NJ: Prentice-Hall.

Brown, J. D., Hudson, T. D., Norris, J. M., and Bonk, W. forthcoming: *Investigating Task-Based Second Language Performance Assessment*. Honolulu: University of Hawai'i Press.

Brown, R. 1973: *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Campbell, D. T. and Reichardt, C. S. 1991: Problems in assuming the comparability of pretest and posttest in autoregressive and growth models. In C. E. Snow and D. E. Wiley (eds), *Improving Inquiry in Social Science*.

Hillsdale, NJ: Lawrence Erlbaum Associates, 201–19.
Carver, R. 1978: The case against statistical significance testing. *Harvard Educational Review*, 48, 389–99.
Chapelle, C. A. 1998: Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds), *Interfaces between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press, 32–70.
Chaudron, C. 1985: Intake: on models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition*, 7, 1–14.
Chaudron, C. 1998: *Second Language Classrooms: Research on Teaching and Learning*. Cambridge: Cambridge University Press.
Chaudron, C. and Russell, G. 1990: The validity of elicited imitation as a measure of second language competence. Ms. University of Hawai'i.
Clahsen, H., Meisel, J., and Pienemann, M. 1983: *Deutsch als Zweitsprache: Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr.
Cohen, J. 1988: *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates.
Cohen, J. 1990: Things I have learned so far. *American Psychologist*, 45, 1304–12.
Cohen, J. 1997: The earth is round ($p < .05$) In L. Harlow, S. Mulaik, and J. Steiger (eds), *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates, 21–36.
Cooper, H. 1998: *Synthesizing Research: A Guide for Literature Reviews*. 3rd edition. Thousand Oaks, CA: Sage.
Cooper, H. and Hedges, L. V. (eds) 1994: *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.
Cronbach, L. J. 1980: Validity on patrol: how can we go straight? In *New Directions for Testing and Measurement: Measuring Achievement, Progress Over a Decade, No. 5*. San Francisco: Jossey-Bass, 99–108.
Cronbach, L. J. 1988: Five perspectives on validity argument. In H. Wainer and H. I. Braun (eds), *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum Associates, 3–45.
Cronbach, L. J. and Meehl, P. E. 1955: Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
Crookes, G. 1990: The utterance and other basic units for second language discourse. *Applied Linguistics*, 11, 183–99.
Crookes, G. 1991: Second language speech production research. *Studies in Second Language Acquisition*, 13, 113–32.
Crookes, G. 1992: Theory formation and SLA theory. *Studies in Second Language Acquisition*, 14, 425–99.
de Graaff, R. 1997: *Differential Effects of Explicit Instruction on Second Language Acquisition*. The Hague: Holland Institute of Generative Linguistics.
DeKeyser, R. 1995: Learning second language grammar rules: an experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, 17, 379–410.
DeKeyser, R. 1997: Beyond explicit rule learning: automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19, 195–221.
Ellis, N. C. 1998: Emergentism, connectionism and language learning. *Language Learning*, 48, 631–64.
Ellis, N. C. 1999: Cognitive approaches to SLA. *Annual Review of Applied Linguistics*, 19, 22–42.
Ellis, N. C. and Schmidt, R. 1997: Morphology and longer-distance dependencies: laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition*, 19, 145–71.
Ellis, R. 1985: A variable competence model of second language acquisition.

International Review of Applied Linguistics, 23, 47–59.

Ellis, R. 1994: *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Embretson, S. E. 1999: Issues in the measurement of cognitive abilities. In S. E. Embretson and S. L. Hershberger (eds), *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates, 1–15.

Embretson, S. E. and Hershberger, S. L. (eds) 1999: *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A. and Simon, H. A. 1993: *Protocol Analysis: Verbal Reports as Data*. Revised edition. Cambridge, MA: MIT Press.

Eubank, L. (ed.) 1991: *Point Counterpoint: Universal Grammar in the Second Language*. Philadelphia: John Benjamins.

Feldt, L. S. and Brennan, R. L. 1989: Reliability. In R. L. Linn (ed.), *Educational Measurement*. 3rd edition. New York: Macmillan, 105–46.

Ferguson, C. A. and Huebner, T. 1991: Foreign language instruction and second language acquisition research in the United States. In K. De Bot, R. B. Ginsberg, and C. Kramsch (eds), *Foreign Language Research in Cross-Cultural Perspective*. Philadelphia: John Benjamins, 3–19.

Foster, P., Tonkyn, A., and Wigglesworth, G. 1998: Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21, 354–75.

Gardner, R. 1979: Social psychological aspects of second language acquisition. In H. Giles and R. S. Clair (eds), *Language and Social Psychology*. Oxford: Blackwell.

Gass, S. M. 1988: Integrating research areas: a framework for second language studies. *Applied Linguistics*, 9, 198–217.

Gass, S. M. 1994: The reliability of second-language grammaticality judgments. In E. Tarone, S. Gass, and A. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 303–22.

Gass, S. M. and Schachter, J. (eds) 1989: *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press.

Givón, T. 1979: *On Understanding Grammar*. London: Academic Press.

Gregg, K. 1990: The variable competence model of second language acquisition, and why it isn't. *Applied Linguistics*, 11, 364–83.

Gregg, K. R. 1993: Taking explanation seriously: or, let a couple of flowers bloom. *Applied Linguistics*, 14, 276–94.

Gregg, K. R. 1996: The logical and developmental problems of second language acquisition. In W. Ritchie and T. Bhatia (eds), *Handbook of Second Language Acquisition*. New York: Academic Press, 49–81.

Grigorenko, E. L., Sternberg, R. J., and Ehrman, M. E. 2000: A theory-based approach to the measurement of foreign language learning ability: the CANAL-F theory and test. *Modern Language Journal*, 84, 390–405.

Gronlund, N. E. and Linn, R. L. 1990: *Measurement and Evaluation in Teaching*. 6th edition. New York: Macmillan.

Grotjahn, R. 1986: Test validation and cognitive psychology: some methodological considerations. *Language Testing*, 3, 159–85.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. 1991: *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds) 1997: *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hatch, E. and Lazaraton, A. 1991: *The Research Manual: Design and Statistics for Applied Linguistics*. New York: HarperCollins and Newbury House.

Henning, G. 1987: *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MA: Newbury House.

Hess, C. W., Sefton, K. M., and Landry, R. G. 1986: Sample size and type–token ratios for oral language of preschool children. *Journal of Speech and Hearing Research*, 29, 129–34.

Hoyle, R. H. (ed.) 1999: *Statistical Strategies for Small Sample Research*. Thousand Oaks, CA: Sage.

Hudson, T. 1993: Nothing does not equal zero: problems with applying developmental sequences findings to assessment and pedagogy. *Studies in Second Language Acquisition*, 15, 461–593.

Huebner, T. 1983: *A Longitudinal Analysis of the Acquisition of English*. Ann Arbor, MI: Karoma.

Huebner, T. 1991: Second language acquisition: litmus test for linguistic theory? In T. Huebner and C. A. Ferguson (eds), *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam and Philadelphia: John Benjamins, 3–22.

Huebner, T. and Ferguson, C. A. (eds) 1991: *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam and Philadelphia: John Benjamins.

Hulstijn, J. H. 1997: Second language acquisition research in the laboratory: possibilities and limitations. *Studies in Second Language Acquisition*, 19, 131–44.

Hunter, J. E. and Schmidt, F. L. 1994: Correcting for sources of artificial variation across studies. In H. Cooper and L. V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 323–36.

Hymes, D. H. 1972: On communicative competence. In J. B. Price and J. Holmes (eds), *Sociolinguistics*. Baltimore: Penguin, 269–93.

Irvine, S. and Kyllonen, P. (eds) 2001: *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Iwashita, N. 1999: The role of task-based conversation in the acquisition of Japanese grammar and vocabulary. Doctoral dissertation. University of Melbourne.

Jourdenais, R., Ota, M., Stauffer, S., Boyson, B., and Doughty, C. (1995): Does textual enhancement promote noticing? A think-aloud protocol analysis. In R. Schmidt (ed.), *Attention and Awareness in Foreign Language Learning*. Technical Report No. 9. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, 183–216.

Kane, M. T. 1992: An argument-based approach to validity. *Psychological Bulletin*, 112, 527–35.

Kellerman, E. 1985: If at first you do succeed . . . In S. M. Gass and C. Madden (eds), *Input in Second Language Acquisition*. Rowley, MA: Newbury House, 345–53.

Krashen, S. 1981: *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press.

Lambert, R. D. 1991: Pros, cons, and limits to quantitative approaches in foreign language acquisition research. In K. De Bot, R. B. Ginsberg, and C. Kramsch (eds), *Foreign Language Research in Cross-Cultural Perspective*. Philadelphia: John Benjamins.

Lantolf, J. (ed.) 1994: *Sociocultural Theory and Second Language Learning*. Special issue of *Modern Language Journal*, 78, 4.

Larsen-Freeman, D. 2000: Second language acquisition and applied linguistics. *Annual Review of Applied Linguistics*, 20, 165–81.

Lazaraton, A. 1992: The structural organisation of a language interview: a conversational analytic perspective. *System*, 20 (3), 373–86.

Lazaraton, A. 1996: Interlocutor support in Oral Proficiency Interviews: the case of CASE. *Language Testing*, 13 (2), 151–72.

Leow, R. P. 1997: Attention, awareness, and foreign language behavior. *Language Learning*, 47, 467–506.

Leow, R. P. 2000: A study of the role of awareness in foreign language behavior: aware vs. unaware learners. *Studies in Second Language Acquisition*, 22, 557–84.

Light, R. and Pillemer, D. 1984: *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Lightbown, P. and White, L. 1987: The influence of linguistic theories on language acquisition research: description and explanation. *Language Learning*, 37, 483–510.

Linacre, J. M. 1998: *Facets 3.17*. Computer program. Chicago: MESA Press.

Linn, R. L. (ed.) 1989: *Educational Measurement*. 3rd edition. New York: American Council on Education and Macmillan.

Linn, R. L. 1997: Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice*, 16 (2), 14–16.

Loevinger, J. 1957: Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–94.

Long, M. H. 1980: Input, interaction and second language acquisition. Doctoral dissertation. University of California at Los Angeles.

Long, M. H. 1990: The least a second language acquisition theory needs to explain. *TESOL Quarterly*, 24, 649–66.

Long, M. H. 1993: Assessment strategies for second language acquisition theories. *Applied Linguistics*, 14, 225–49.

Long, M. H. 1996: The role of the linguistic environment in second language acquisition. In W. C. Ritchie and T. K. Bahtia (eds), *Handbook of Second Language Acquisition.* New York: Academic Press, 413–68.

Loschky, L. 1994: Comprehensible input and second language acquisition: what is the relationship? *Studies in Second Language Acquisition*, 16, 303–23.

Loschky, L. and Bley-Vroman, R. 1993: Grammar and task-based methodology. In G. Crookes and S. Gass (eds), *Tasks and Language Learning: Integrating Theory and Practice.* Philadelphia: Multilingual Matters, 123–67.

Lynch, B. and Davidson, F. 1994: Criterion-referenced language test development: linking curricula, teachers, and tests. *TESOL Quarterly*, 28, 727–43.

Lyster, R. 1998: Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, 48, 183–218.

Mackey, A. 1999: Input, interaction, and second language development: an empirical study of question formation in ESL. *Studies in Second Language Acquisition*, 21, 557–87.

Mackey, A. and Philp, J. 1998: Conversational interaction and second language development: recasts, responses, and red herrings? *Modern Language Journal*, 82, 338–56.

MacWhinney, B. (ed.) 1998: *The Emergence of Language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 2000: *The CHILDES project: Tools for Analyzing Talk. Vol. I: Transcription Format and Programs*. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Marcoulides, G. A. 1999: Generalizability theory: picking up where the Rasch IRT model leaves off? In S. Embretson and S. Hershberger (eds), *The New*

Rules of Measurement: What Every Psychologist and Educator Should Know. Mahwah, NJ: Lawrence Erlbaum Associates, 129–52.

McLaughlin, B. 1987: *Theories of Second Language Learning*. London: Edward Arnold.

McLaughlin, B. 1990: Restructuring. *Applied Linguistics*, 11, 1–16.

McNamara, T. 1996: *Measuring Second Language Performance*. New York: Longman.

McNamara, T. F. and Lumley, T. 1997: The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, 14 (2), 140–56.

Meisel, J., Clahsen, H., and Pienemann, M. 1981: On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition*, 3, 109–35.

Mellow, D., Reeder, K., and Forster, E. 1996: Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition*, 18, 325–50.

Messick, S. 1975: The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–66.

Messick, S. 1989: Validity. In R. L. Linn (ed.), *Educational Measurement*. 3rd edition. New York: American Council on Education and Macmillan, 13–103.

Messick, S. 1994: The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23 (2), 13–23.

Mislevy, R. J. 1994: Evidence and inference in educational assessment. Presidential address to the Psychometric Society. *Psychometrika*, 59, 439–83.

Mislevy, R. J. 1995: Test theory and language-learning assessment. *Language Testing*, 12, 341–69.

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. 1999: *On the Roles of Task Model Variables in Assessment Design*. CSE Technical Report 500. Los Angeles, CA: Center for the Study of Evaluation, Graduate School of Education and Information Studies at the University of California, Los Angeles.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., and Penuel, W. R. forthcoming: Leverage points for improving educational assessment. In B. Means and G. D. Haertel (eds), *Designs for Evaluating the Effects of Technology in Education*.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., and Johnson, L. 1999: A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior*, 15, 335–74.

Moss, P. A. 1992: Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, 62 (3), 229–58.

Nichols, P. and Sugrue, B. 1999: The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice*, 18 (2), 18–29.

Norris, J. M. 1996: A validation study of the ACTFL guidelines and the German Speaking Test. Master's thesis. University of Hawai'i.

Norris, J. M. 1997: The German Speaking Test: utility and caveats. *Die Unterrichtspraxis*, 30 (2), 148–58.

Norris, J. M. 2000: *Tasks and Language Assessment*. Paper presented in the colloquium "Key issues in empirical research on task-based instruction" at the annual American Association for Applied Linguistics conference (AAAL). Vancouver, British Columbia, Canada, March 14.

Norris, J. M. and Ortega, L. 2000: Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417–528.

Norris, J. M., Brown, J. D., Hudson, T., and Yoshioka, J. 1998: *Designing Second Language Performance Assessments*. Technical Report No. 17. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Oliver, R. 1995: Negative feedback in child NS/NNS conversation. *Studies in Second Language Acquisition*, 17, 459–81.

Ortega, L. 1999: Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 109–48.

Ortega, L. 2000: Understanding syntactic complexity: the measurement of change in the syntax of instructed L2 Spanish learners. Doctoral dissertation. University of Hawai'i at Manoa.

Orwin, R. G. 1994: Evaluating coding decisions. In H. Cooper and L. V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 139–62.

O'Sullivan, B. 2000: Exploring gender and oral proficiency interview performance. *System*, 28 (3): 373–86.

Paolillo, J. C. 2000: Asymmetries in Universal Grammar: the role of methods and statistics. *Studies in Second Language Acquisition*, 22, 209–28.

Pedhazur, E. J. and Schmelkin, L. P. 1991: *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pellegrino, J. W. 1988: Mental models and mental tests. In H. Wainer and H. Braun (eds), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 49–60.

Pica, T. 1983: Methods of morpheme quantification: their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, 6, 69–79.

Pienemann, M. 1984: Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition*, 6, 186–214.

Pienemann, M. 1998: *Language Processing and Second Language Development: Processability Theory*. Philadelphia: John Benjamins.

Pienemann, M. and Mackey, A. 1993: An empirical study of children's ESL development and Rapid Profile. In P. McKay (ed.), *ESL Development: Language and Literacy in Schools, vol. 2*. Commonwealth of Australia and National Languages and Literacy Institute of Australia, 115–259.

Pienemann, M., Johnston, M., and Brindley, G. 1988: Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, 10, 217–43.

Pishwa, H. 1994: Abrupt restructuring versus gradual acquisition. In C. A. Blackshire-Belay (ed.), *Current Issues in Second Language Acquisition and Development*. New York: University Press of America, 143–66.

Polio, C. G. 1997: Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101–43.

Polio, C. and Gass, S. 1997: Replication and reporting: a commentary. *Studies in Second Language Acquisition*, 19, 499–508.

Popham, W. J. 1981: *Modern Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Preston, D. R. 1989: *Sociolinguistics and Second Language Acquisition*. New York: Blackwell.

Richards, B. 1987: Type/token ratios: what do they really tell us? *Journal of Child Language*, 14, 201–9.

Richards, B. J. 1990: *Language Development and Individual Differences:*

A Study of Auxiliary Verb Learning. New York: Cambridge University Press.

Richards, B. J. 1994: Child-directed speech and influences on language acquisition: methodology and interpretation. In C. Gallaway and B. J. Richards (eds), *Input and Interaction in Language Acquisition*. Cambridge: Cambridge University Press, 74–106.

Richards, B. J. and Gallaway, C. (eds) 1994: *Input and Interaction in Language Acquisition*. Cambridge: Cambridge University Press.

Richards, B. J. and Malvern, D. D. 1997: *Quantifying Lexical Diversity in the Study of Language Development*. Reading: University of Reading, New Bulmershe Papers.

Robinson, P. 1995: Attention, memory, and the "noticing" hypothesis. *Language Learning*, 45, 283–331.

Robinson, P. 1997: Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning*, 47, 45–99.

Robinson, P. 2001a: Task complexity, cognition and second language syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (ed.), *Cognition and Second Language Instruction*. New York: Cambridge University Press, 287–318.

Robinson, P. 2001b: Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22, 27–57.

Rosenthal, R. 1979: Replications and their relative utility. *Replications in Social Psychology*, 1, 15–23.

Rosenthal, R., Rosnow, R. L., and Rubin, D. B. 2000: *Contrasts and Effect Sizes in Behavioral Research*. New York: Cambridge University Press.

Rosnow, R. L. and Rosenthal, R. 1989: Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–84.

Ross, S. and Berwick, R. 1990: The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 159–76.

Royer, J. M., Cisero, C. A., and Carlo, M. S. 1993: Techniques and procedures for assessing cognitive skills. *Review of Educational Research*, 63 (2), 201–43.

Rutherford, W. 1984: Description and explanation in interlanguage syntax: the state of the art. *Language Learning*, 34, 127–55.

Saito, H. 1999: Dependence and interaction in frequency data analysis in SLA research. *Studies in Second Language Acquisition*, 21, 453–75.

Sawyer, M. and Ranta, L. 2001: Aptitude, individual differences, and instructional design. In P. Robinson (ed.), *Cognition and Second Language Instruction*. New York: Cambridge University Press, 424–69.

Schachter, J. 1998: Recent research in language learning studies: promises and problems. *Language Learning*, 48, 557–83.

Schmidt, R. 1993: Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206–26.

Schmidt, R. 1994: Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 11, 11–26.

Scholfield, P. 1995: *Quantifying Language: A Researcher's and Teacher's Guide to Gathering Language Data and Reducing it to Figures*. Bristol, PA: Multilingual Matters.

Schumann, J. 1978: The acculturation model for second-language acquisition. In R. C. Gringas (ed.), *Second Language Acquisition and Foreign Language Teaching*. Washington, DC: Center for Applied Linguistics, 27–50.

Schwartz, B. 1992: Testing between UG-based and problem-solving

models of L2A: developmental sequence data. *Language Acquisition*, 2, 1–19.

Schwartz, B. D. 1993: On explicit and negative data effecting and affecting competence and linguistic behaviour. *Studies in Second Language Acquisition*, 15, 147–63.

Seliger, H. W. and Shohamy, E. 1989: *Second Language Research Methods*. Oxford: Oxford University Press.

Shavelson, R. J. and Webb, N. M. 1991: *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.

Shepard, L. A. 1993: Evaluating test validity. *Review of Research in Education*, 19, 405–50.

Shepard, L. 1997: The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5–13.

Shohamy, E. 1994: The role of language tests in the construction and validation of second-language acquisition theories. In E. Tarone, S. Gass, and A. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 133–42.

Shohamy, E. 2000: The relationship between language testing and second language acquisition, revisited. *System*, 28, 541–53.

Siegler, R. S. 1989: Strategy diversity and cognitive assessment. *Educational Researcher*, 18 (9), 15–20.

Sinclair, J. M. and Coulthard, R. M. 1975: *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. London: Oxford University Press.

Skehan, P. 1998: *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Snow, R. and Lohman, D. 1989: Implications of cognitive psychology for educational measurement. In R. L. Linn (ed.), *Educational Measurement*. 3rd edition. Washington, DC: American Council on Education and National Council on Measurement in Education, 263–332.

Sokolik, M. E. 1990: Learning without rules: PDP and a resolution of the adult language learning paradox. *TESOL Quarterly*, 24, 685–96.

Sorace, A. 1993: Incomplete vs. divergent representations of unaccusativity in near-native grammars of Italian. *Second Language Research*, 9, 22–48.

Sorace, A. 1996: The use of acceptability judgments in L2 acquisition research. In W. Ritchie and T. Bhatia (eds), *Handbook of Second Language Acquisition*. New York: Academic Press, 375–409.

Sorace, A. and Robertson, D. 2001: Measuring development and ultimate attainment in non-native grammars. In C. Elder, A. Brown, N. Iwashita, E. Grove, K. Hill, and T. Lumley (eds), *Experimenting with Uncertainty: Essays in Honour of Alan Davies*. Cambridge: Cambridge University Press, 264–74.

Stadler, M. A. and Frensch, P. A. (eds) 1998: *Implicit Learning Handbook*. Thousand Oaks, CA: Sage.

Stromswold, K. 1996: Analyzing children's spontaneous speech. In D. McDaniel, C. McKee, and H. S. Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 23–53.

Sugrue, B. 1995: A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice*, 14 (3), 29–36.

Swain, M. 1985: Communicative competence: some roles of comprehensible input and comprehensible output in its development. In S. M. Gass and C. G. Madden (eds), *Input in Second Language Acquisition*. Rowley, MA: Newbury House, 235–53.

Swain, M. 1995: Three functions of output in second language learning. In G. Cook and B. Seidhofer (eds),

Principles and Practice in the Study of Language. Oxford: Oxford University Press, 125–44.

Tabachnick, B. G. and Fidell, L. S. 1996: *Using Multivariate Statistics.* 3rd edition. New York: HarperCollins.

Tarone, E. 1988: *Variation in Interlanguage*. London: Edward Arnold.

Tarone, E. 1998: Research on interlanguage variation: implications for language testing. In L. F. Bachman and A. D. Cohen (eds), *Interfaces Between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press, 71–89.

Tarone, E. and Parrish, B. 1988: Task-related variation in interlanguage: the case of articles. *Language Learning*, 38, 21–44.

Thissen, D. and Wainer, H. (eds) 2001: *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Thomas, M. 1994: Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44, 307–36.

Thompson, B. 1994: Guidelines for authors. *Educational and Psychological Measurement*, 54, 837–47.

Thompson, B. 1998: Five methodology errors in educational research: the pantheon of statistical significance and other faux pas. Presentation at the American Educational Research Association annual conference. San Diego, April 15. Available at: <http://acs.tamu.edu/~bbt6147/>

Tomasello, M. 1998a: Introduction: a cognitive-functional perspective on language structure. In M. Tomasello (ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure.* Mahwah, NJ: Lawrence Erlbaum Associates, vii–xxiii.

Tomasello, M. (ed.) 1998b: *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tomlin, R. 1990: Functionalism in second language acquisition. *Studies in Second Language Acquisition*, 12, 155–77.

Tomlin, R. S. and Villa, V. 1994: Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, 16, 183–203.

Trahey, M. and White, L. 1993: Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition*, 15, 181–204.

Traub, R. E. 1994: *Reliability for the Social Sciences: Theory and Applications*. Thousand Oaks, CA: Sage.

Vacha-Haase, T., Ness, C., Nilsson, J., and Reetz, D. 1999: Practices regarding reporting of reliability coefficients: a review of three journals. *Journal of Experimental Education*, 67 (4), 335–41.

Vygotsky, L. 1986: *Thought and Language*. Translation newly rev. and ed. Alex Kozulin. Cambridge, MA: MIT Press.

White, L. 1991: Second language competence versus second language performance: UG or processing strategies. In L. Eubank (ed.), *Point Counterpoint: Universal Grammar in the Second Language.* Amsterdam: John Benjamins, 167–89.

White, L. 1996: Universal Grammar and second language acquisition: current trends and new directions. In W. C. Ritchie and T. K. Bhatia (eds), *Handbook of Second Language Acquisition.* San Diego: Academic Press, 85–120.

White, L. 2000: Second language acquisition: from initial to final state. In J. Archibald (ed.), *Second Language Acquisition and Linguistic Theory.* New York: Blackwell, 130–55.

White, L. and Genesee, F. 1996: How native is near-native? The issue of ultimate attainment in adult second

language acquisition. *Second Language Research*, 12, 233–65.

Whittington, D. 1998: How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, 58, 21–37.

Wickens, C. D. 1989: Attention and skilled performance. In D. H. Holding (eds), *Human Skills*. Chichester: John Wiley, 71–104.

Wiley, D. E. 1991: Test validity and invalidity reconsidered. In R. E. Snow and D. E. Wiley (eds), *Improving Inquiry in Social Science*. Hillsdale, NJ: Lawrence Erlbaum Associates, 75–107.

Wilkinson, L. and the Task Force on Statistical Inference 1999: Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, 54 (8), 594–604.

Willett, J. B. 1988: Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.

Williams, J. N. 1999: Memory, attention, and inductive learning. *Studies in Second Language Acquisition*, 21, 1–48.

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. 1998: *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Technical Report No. 17. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Wolfram, W. 1985: Variability in tense marking: a case for the obvious. *Language Learning*, 35, 229–53.

Woods, A., Fletcher, P., and Hughes, A. 1986: *Statistics in Language Studies*. New York: Cambridge University Press.

Wright, B. D. 1999: Fundamental measurement for psychology. In S. E. Embretson and S. L. Hershberger (eds), *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates, 65–104.

Young, R. 1995a: Conversational styles in language proficiency interviews. *Language Learning*, 54, 3–42.

Young, R. 1995b: Discontinuous interlanguage development and its implications for oral proficiency rating scales. *Applied Linguistics*, 6, 13–26.

Young, R. and He, A. W. 1998: *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Studies in Bilingualism, 14. Philadelphia: John Benjamins.

Young, R. and Milanovic, M. 1992: Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 403–24.

Yuan, B. 1997: Asymmetry of null subjects and null objects in Chinese speakers' L2 English. *Studies in Second Language Acquisition*, 19, 467–97.

Yule, G. 1997: *Referential Communication Tasks*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zobl, H. 1995: Converging evidence for the "acquisition–learning" distinction. *Applied Linguistics*, 16, 35–56.