22 Data Collection in SLA Research

CRAIG CHAUDRON

1 Introduction

Although professionals working in second/foreign language pedagogy have always displayed interest in assessing learners' achievement, the approach to doing so has typically involved tests developed within an instructional/achievement/normative paradigm (see Chaudron, 2001, for a review of foreign language research), so that a record of individuals' or groups of learners' L2 development was not the principal focus of research. In contrast to this tradition, research on second language acquisition (SLA), deriving from studies of child language acquisition of the 1950s and 1960s, began in earnest in the late 1960s with the advent of the notion of interlanguage development (see Hatch, 1978b, for a review).

The application of research methods in second language acquisition since the early 1970s has intensified and become more refined in substantial ways, by adopting or expanding upon methodologies employed especially in psychology, sociology, anthropology, and linguistics, all of which were themselves fledgling fields with hybrid research approaches barely a century ago. *Research design*, as in experimental studies, case studies, and correlational studies, and many other approaches and techniques for *data analysis* are equally critical to successful, informative research (see overview texts such as Hatch and Lazaraton, 1991; D. Johnson, 1992; Nunan, 1992; see also, for example, the treatment of issues with respect to time-series designs in Mellow, Reeder, and Forster, 1996). However, because the principles underlying reliable and valid adoption and refinement of these approaches are largely common to research in the broader disciplines, and can, therefore, be reviewed and accessed through that literature, this chapter will adopt a narrower focus – on the core of research procedures in SLA, namely *data-collection procedures* as employed since the early 1980s in SLA research (see descriptions of instrumentation and procedures in Doughty and Long, 2000; Larsen-Freeman and Long, 1991, pp. 26–38; Seliger and Shohamy, 1989, pp. 158–80).[1] It has become clear in the development of the SLA research

tradition that, regardless of the particular approach or design adopted by the researcher, a variety of data-collection procedures is feasible,[2] if not desired, in order for the researcher to obtain the best sample of learners' performance potential. The various approaches to the elicitation and collection of data on learner performance in second languages will therefore be outlined, with a concern for the appropriate methodologies that ensure valid SLA analysis.

This overview will focus on the following primary questions:

i   What procedures have been applied in SLA research, and what are the conditions and characteristics of each?
ii   How reliable and valid are the procedures and methods for collecting SLA data?
iii   As a specific outcome of the question of validity, what are the limits to interpretation from data collected by each procedure as reflections of underlying SLA performance/competence?
iv   As a matter of the generalizability of procedures, to what extent can procedures adopted for L1 research be applied in L2 research?
v   What new or additional procedures or adaptations are possible or needed for L2 research?

A wide range of procedures and methods will be described, each illustrated by a number of exemplary studies, and overviews of SLA research using each procedure, when they exist, will be identified for further study.

1.1   General dimensions of data collection

First language data-collection methodologies have been characterized by Bennett-Kastor (1988) either as "naturally observed" or as elicited under "controlled observation" (p. 26). Of the naturally observed data sources, she includes three types: "indirect or anecdotal evidence; native speaker 'intuitions', especially as judgments of the acceptability of utterances; and 'raw' data actually manifested in conversational and other naturally occurring forms" (p. 26). Bennett-Kastor acknowledges, however, that data may also be experimentally elicited through controlled procedures, which are necessarily used to obtain judgment data, but which involve as well "manipulation of objects, pictures or even the child's postures . . . and the elicitation of descriptions by the child of object configurations," and even imitation or other verbal manipulation tasks (p. 29ff). This naturalistic/experimental dimension for classifying research methodologies (recapitulated by Nunan, 1996, with reference to research design and general methodology) has been applied under various guises by many L2 theorists, as in Cook's (1986) "authentic" vs. "non-authentic" distinction, van Lier's (1988) "± interventionist" dichotomy, Seliger and Shohamy's (1989, pp. 158ff) "low explicitness/heuristic" vs. "high explicitness/deductive," and Larsen-Freeman and Long's (1991, pp. 14–15) "qualitative/quantitative continuum." These characterizations, at times, include the entire perspective

Table 22.1 Data collection methods classified on the naturalistic to experimental continuum

Naturalistic (contextualized)	Elicited production	Experimental (decontextualized)
Learner speech production		
Recorded natural (spontaneous) speech	Role play	On-line processing measures (sentence matching/verification, signal detection, word recognition, decision)
Communication task (meaning-focused)	Story (video) retelling	
Unstructured interviews	Structured interview	Utterance completion
		Elicited imitation
	Picture description/picture prompts	
	Instruction giving	Word association
Classroom observation	Discourse completion	Elicited translation
		Sentence manipulation
Communication task (structure-focused)	Structured questionnaire	
	OPI	
	SOPI	
	Stimulated recall (verbal reports, etc.)	Discrete-point tests
Diary (self)	Integrated (e.g., cloze) tests	Metalinguistic tests (card sorting, grammaticality judgments, magnitude scaling, paired comparisons, sentence correction, rule expression)
Observational notes		
Reflection on production: perception/interpretation		

undertaken in the research (distinguished as in Grotjahn, 1987, on a continuum from "nomothetic" to "heuristic"), rather than the specific data-collection procedures, any one of which might be applicable to a diverse array of designs and research purposes. In this chapter, however, we will avoid entering into the question of general research orientations, restricting the overview to the description of the procedures and their specific characteristics and capacity for obtaining valid data.[3]

Given a dimension, as in Bennett-Kastor, from "naturalistic" to "experimental," where the degree of *contextualization for meaningful and purposeful language use* is a key criterion to be used to place a method on the continuum, the data-collection techniques to be reviewed here can be laid out as in table 22.1. Here, a vertical dimension has been added to suggest the *extent of direct vs. indirect elicitation* of L2 linguistic forms or competence, that is, the extent to which the subjects/learners are led to produce without further reflection, or to express their reflections and interpretations of the language they have produced or are presented with. In a related view, this dimension is referred to by Kasper and Dahl (1991, p. 217) as "modality of language use: perception/comprehension → production." (See also Chaudron, 1985b, for the input–intake relationship.)

To some extent, the naturalistic–experimental continuum may seem related, as well, to the "interlanguage continuum" distinction made by Tarone (1979, 1983; following Labov, 1969), with spontaneous free speech ("vernacular" is a term that is often used, but may be inappropriate for L2 speech) on one end and careful controlled speech on the other. SLA researchers are cautious about disentangling L2 production effects that are based on implicitly acquired knowledge from those that derive from explicit knowledge and its influence in monitored speech (see discussion in Doughty and Long, 2000, pp. 154–6). But this important factor of speech-style monitoring that can increase intra-subject variability is somewhat independent of the specific elicitation methodology chosen, and more contingent on certain contextual conditions that can be manipulated or altered in any given methodology, such as the amount of time allowed to the subject to plan or reflect on his or her performance; the degree of social demand for more polished performance, which occurs in interaction with larger groups or higher-status interlocutors; and the extent of feedback.

After an introductory note on reliability and validity (see also Norris and Ortega, this volume), a topic which will be reprised at the end of this review, the methods reviewed will be grouped according to three primary types, following the horizontal dimension on table 22.1 from left to right: *naturalistic*, *elicited production*, and *experimental*, or more decontextualized, performance stimuli.

1.2   *Reliability and validity*

A key concern of this review is whether the particular methods employed by researchers have proven to be reliable and valid, that is, whether they have consistently led to successful elicitation of learners' language performance and, possibly, competence, and whether the analysis of this performance matches

other, independent measures of or expectations for the learners' production. It is frequently impossible to tell from a given collection of data whether the forms produced are simply an artifact of the method. This is why many researchers today employ multiple measures, in order to "triangulate" their findings, and to differentiate the possible effects of the method employed from the stable or developing traits of the learners' underlying language capacity. The use of multiple measures is exemplified in three procedures discussed by Doughty and Long (2000) in their review of data elicitation, all of which employ various tasks to elicit a range of language abilities: the *developmental linguistic profiling* procedures of Pienemann and his associates (Pienemann, 1998); the *descriptive linguistic profiling* employed by both the European Science Foundation (ESF) project (Dietrich, Klein, and Noyau, 1995; W. Klein and Perdue, 1992; Perdue, 1982) and the FLIRT research project at Georgetown University (see the web page at <http://cfdev.georgetown.edu/flirt/docs/advlearn.htm>); and the research of Pica and her colleagues (Pica, Kanagy, and Falodun, 1993) and Swain and her colleagues (Swain and Lapkin, 1998) on *production via interaction on tasks*, which involves performance on multiple or sequential communication tasks. At the end of the following review of procedures, we will attempt to determine which comparisons among distinct methods and tasks appear to confirm or to question the validity of particular data elicitation procedures.

2   Naturalistic Data Collection Procedures

Some of the earliest approaches to and sources of productive data on SLA involved *observation of children's language use in play and normal interaction* with parents and others (Leopold, 1939, 1948; Ravem, 1968, 1970). Well-known early SLA research of this sort also includes studies by Hakuta (1974, 1976) and Huang and Hatch (1978); see the collection of studies in Hatch (1978a). Procedures for observing and recording children's (or adults') speech in such naturalistic settings have gradually been elaborated and are outlined and critiqued in some detail regarding L1 research in Milroy (1987), Bennett-Kastor (1988, pp. 55–73), and Demuth (1996), where recommendations for the use of video- and audio-recording devices are made.

2.1   *Advantages and disadvantages*

Naturalistic observation must be systematic, and details of the physical and verbal context in which observations and recordings are made must be recorded, in order to retrieve sources of deictic reference, restrictive/non-restrictive relative clause contexts, and possible social-interactive events that would influence pragmatic meanings and interpretations, and, whenever possible, to supplement potentially unintelligible linguistic information in the recordings. Assuming a context for observation in a naturally occurring social event, the advantage of data collected in this way is that the learner's production will be a true sample

of his or her L2 speech, possibly in open communication with familiar colleagues or friends, and uninfluenced by artifactual aspects of an elicitation method, as well as potentially less influenced by the learner's careful monitoring or application of learned rules of production. Also, extended research studies can collect very large amounts of learner production data in this manner, without substantially preparing new materials or altering any procedures for data collection. As will be clear in later sections, the preparation of quality materials for eliciting more controlled speech forms is a highly complicated, intensive process.

The disadvantages of naturalistic observation include a number of well-known limitations. Obviously, the quality of recordings made can be a significant factor in reducing or enhancing the validity of the data. Mechanical failures in recording equipment (or failure to plug it in or turn it on!) are frequent sources of difficulty for researchers (cf. Swain and Cumming, 1989), and the observer's written record may omit contextual features or distort speech production data if no independent recording is available. But more significant is the major concern of researchers that particular target structures or competences of the L2 may be avoided or underrepresented and thus not assessed: it is difficult to interpret the absence of structures in the learners' performance, and comprehension is typically not easily evaluated. Especially in records of younger children interacting, whether or not a particular form is productive in use cannot be determined. A further limitation is that naturalistic observation is highly labor-intensive, owing to the need for a trained observer at all data-collection moments and later during elaborate transcription of the data; as a consequence, only smaller numbers of subjects can be studied. This is why naturalistic observation is typically used in case studies of only a few children or a single classroom. In such research, the value of the data will depend more on the variability that arises through the natural interaction that might occur, or possible differences between the few subjects, or because the procedure is used in a longitudinal design to discover trends in acquisition.

2.2 Observation of child and adult learning in context

Among a number of studies using naturalistic observation are those displayed in table 22.2, which include recent examples of the use of naturalistic observation in studies of young bilingual children (ages from less than 2 to 4) engaged in free play with their parents or a TL-speaking research assistant (in most cases, each parent is a native speaker of a different language) over a period of from one to two years (Meisel, 1994a, 1994b; Nicolaidis and Genesee, 1996; Paradis and Genesee, 1996; Schlyter, 1993).

In such studies, the number of children observed or analyzed is small (from two to six – though Meisel's group, cf. Meisel, 1994b, has collected data from 13 children), but the large number of longitudinal data obtained can allow for a very wide range of analyses. Typically, the researchers will only sample from the large amount of data collected (from one-third to half of the data) for the

Table 22.2 Studies of natural language production in bilingual children

Study	Languages	N	Ages	Context and period	Frequency	Focus
Studies involving natural play with parents or interlocutors:						
Swain and Wesche (1973)	French English	1	3	$1\frac{1}{2}$-hour play with two TL researchers for 9 months	Every 2 weeks	Language mixing and switching elicited translation
Schlyter (1993)	French Swedish	6	2–4	$\frac{1}{2}$-hour play with each parent for 2 years	Every 4 months	Constraints on finiteness, negation, word order, pronouns in stronger/weaker language
Meisel (1994a)	French German	2 (out of larger n = 13)	1:3–3:0	1-hour play with TL speakers/ parents for 3+ years	Every 2 weeks	Mixing of L1/L2, acquisition of structural constraints on mixing/switching
Paradis and Genesee (1996)	French English	3	1:11–3:3	1-hour play with each parent and both for 1 year	Every 4 months	Development of finiteness, negation, pronominals

Study	Languages	N	Age	Data type	Frequency	Focus
Nicolaidis and Genesee (1996)	French English	4	1:5–3:8	$\approx$ 1-hour play with each parent for $1\frac{1}{2}$ years	Every 2 months	Language choice with parents, use of translation
Studies involving guided conversations, play or interviews:						
Pfaff (1992)	Turkish German	3 (out of larger n = 31)	1–5	$\frac{1}{2}$-hour play with researchers over 3 years	Twice a month (once each language)	Grammaticalization, pragmatic/syntactic categories
Tomiyama (2000)	Japanese English	1	8–11	1+-hour English conversation + elicited productions over almost 3 years	Every month, then every 2 months	Attrition of English morphology and lexis
Haznedar (2001)	Turkish English	1	4–5:6	40–90-minute play interaction in English	Three times per month	Development of copula, aux., tense agreement, subject pronouns
Bardovi-Harlig and Hartford (1996)	English Various	16	Adults	$\approx$ 15–20-minute academic interviews	Twice an academic year	Pragmatic competence development

sake of transcription and detailed analysis. While Nicolaidis and Genesee focus on the developmental changes in the children's code switching and appropriateness of language choice with their parent interlocutors, the others – Meisel, Paradis and Genesee, and Schlyter – attempt to compare the children's patterns of structural constraints in the dual acquisition of French with German, English, and Swedish, respectively, to expectations for universal L1 patterns of acquisition in each (e.g., structural constraints on language mixing within utterances, acquisition of finite marking and pronominal subjects), or to assess transfer effects from the dominant into the weaker language and vice versa.

Researchers employing naturalistic observation and recording with older subjects are less likely to use this technique without some form of intervention or control, minimally an *interview* or *conversation* with the subject, whether unstructured or structured. (However, see research on code switching in adult bilinguals by, for example, Poplack and her colleagues – Poplack, Wheeler, and Westwood 1989 – which employs "standard social network techniques" to record and analyze free conversations. Also, Bardovi-Harlig and Hartford, 1993, 1996, used recordings of authentic academic interviews with non-native students in order to evaluate the development of their pragmatic competence.) Thus, even when studying a subject as young as 3 years of age, in her groundbreaking research on early childhood bilingualism, Swain (1972; Swain and Wesche, 1973) and her co-researcher interacted with their subject, each in a different language, in order to elicit the child's abilities, occasionally asking the child to translate for the other language speaker. Likewise, while researching the same topic as Paradis and Genesee, Haznedar (2001) chose to engage in play-like interaction with a 4-year-old Turkish child learning L2 English, primarily in order to elicit L2 use. And Tomiyama (2000) employed English conversation interviews for 33 months with an 8-year-old Japanese native child returning from a seven-year residence in the United States.

A number of European studies of L2 development in adults and children, such as the ESF project, obtained data from the natural "free conversation" of informants and occasionally observation of their speech with others. Generally, these studies have relied more on *participant observation* records, informant self-recorded *diaries*, and unstructured and structured *conversational interviews* (Perdue, 1982, pp. 56–61), along with more controlled and experimental procedures (see, e.g., studies by Perdue and Klein, Pfaff, Giacalone Ramat, and Skiba and Dittmar, all in Dittmar, 1992; chapters in Dietrich et al., 1995; and all articles in Perdue, 2000). In the US, Park (2000) reports an analysis of data collected on Korean children acquiring L2 English, as part of an early 1980s project of the National Center for Bilingual Research in Los Angeles. The data were collected with individualized tape-recorders both at home and in school, with some semi-structured and structured elicitation. Finally, an innovative approach to the use of "diaries" with adults is the study by Brown, Sagers, and LaPorte (1999), who examined oral and written *dialog journals* (interactions between EFL learners and their teachers) as a basis for the analysis of vocabulary acquisition.

It deserves mention that the use of *self-report diaries* maintained by researchers on their own language-learning experiences has not seen a significant continued use in the investigation of L2 production and development. Aside from reports of such studies, primarily in the 1980s, whose focus was on affective and cognitive effects of learning or teaching in uninstructed and classroom contexts (e.g., Bailey, 1983, 2001; F. Schumann, 1980; see Bailey and Nunan, 1996, for reports of classroom use of student and teacher diary-keeping), one diary study that gained notoriety for its analysis of linguistic features in the acquisition of Portuguese by Schmidt (Schmidt and Frota, 1986) remains a model for too few subsequent studies (see Jones, 1995, for a self-study of Hungarian vocabulary acquisition). Schmidt's detailed diaries were also compared with analyses of his Portuguese use in monthly recordings of unstructured conversations with his co-author. In this way, the researcher was especially able to compare the relationship between his sometimes unconscious production and use in interaction and his development in Portuguese grammar.

2.3 Classroom observation

As a special case of naturalistic observation, techniques for the *observation of learning in classrooms* have been developed since the mid-1960s (as reviewed in Allwright, 1988; Allwright and Bailey, 1991; Chaudron, 1988; van Lier, 1988; and many more). Such approaches have not typically been used to examine SLA development directly, as they usually involve global observation schedules or checklists, and focus on teacher behavior or classroom interaction processes. However, through analyses of additional measures of interaction and learning opportunities arising from classroom behaviors, researchers have pointed to contexts and processes that might influence SLA. (See, e.g., Markee, 1994, who analyzed vocabulary awareness from transcripts of lessons during which all participants recorded separately on their own lapel microphones; Lyster, 1998a, 1998b, and Lyster and Ranta, 1997, who evaluated student performance relative to teacher feedback treatments; and Williams, 1999, who evaluated the nature of student interaction during language form-related activities.) It deserves note, however, that a few such studies have made use of the analysis of classroom transcripts or direct observation of learner production in order to characterize L2 progress. An early such study was that of Adams (1978, originally conducted in 1971), who maintained a systematic record of Spanish-speaking ESL learners' production of a variety of morphemes and syntactic structures over a two-year period. These results were compared with elicited imitation and translation tasks involving similar structures. Likewise, Ellis (1992) studied learners' classroom performance, although his analysis was based on his hand-recorded notes on two learners' productions and interaction, with audio-recordings used only to confirm analyses. Both of these researchers lamented the high degree of background noise that often affects classroom recordings, which is an argument for individually assigned microphones and recorders. For the most part, however, classroom-based observations

of learner development have made use of more controlled tasks that will be described in later sections.

3   Elicited Production Procedures

Although the data obtained from naturalistic observation are demonstrably extensive and informative, especially insofar as intra- and inter-subject longitudinal development can be analyzed, the disadvantages of such data, noted above, have been a concern for many researchers interested in more specific L2 learning targets. Therefore, a number of techniques have been developed that are designed to elicit learners' productive language performance in a more concentrated and focused fashion, by providing some initial verbal or physical context selected by the researcher. These include *structured interviews* intending to elicit particular target forms (e.g., past tense, hypotheticals, adjectives, and relative clauses) or topics (e.g., family relationships), a very wide range of *communication tasks* with greater or lesser attention to meaning vs. form (as suggested in table 22.1, the more attention is directed to form, the more reflective the task), *role plays*, *picture descriptions* and *instruction-giving*, *story-retelling*, *discourse completion*, *stimulated recall* and other *structured questionnaires*, and combinations of these as used in the *Oral Proficiency Interview* and *Simulated Oral Proficiency Interview*. For most of these procedures, first language researchers have created many more extensive uses, although owing to the lesser cognitive maturity of the subjects of such studies, a number of more meaningful communication tasks (such as role plays and debates) have not normally been used to elicit L1 linguistic performance per se. (For reviews of L1 methods, see chapters 5, 6, and 9–13 in Menn and Bernstein Ratner, 2000; see Crookes, 1991, for L2 research methodology.)

3.1   *Advantages and disadvantages*

The advantages in general of such elicitation tasks relative to naturalistic observation are that, first, they can be tailored to specific points of L2 learning that are the theoretical focus of the research, especially certain communication tasks or picture descriptions that can be designed to promote specific productions, known as "task essentialness" (cf. Loschky and Bley-Vroman, 1993). Second, they can be employed in a more mechanical fashion using recording instruments, so that the researcher and assistants can elicit more subjects' data with less concern for observational reliability. A third advantage is that they lend themselves to use with learners of virtually any level of L2 competence, because translations of instructions or materials can be provided. Finally, fourth, they tend to be more easily analyzed and scored, although transcription and coding of protocols do require reliability assessments.

Relative to more decontextualized or receptive experimental measures, these tasks also have the following advantage: depending on the volume of language

elicited and the extent of naturalistic context provided (e.g., as in a role play), they can elicit an extensive range of potentially natural, unmonitored learner performance appropriate to a given genre of speech behavior or style. For the above reasons, most researchers have tended to elicit data using more than one such method, thereby allowing for better cross-task validation of their findings.

   Among some possible disadvantages are ones similar to those for naturalistic observation. First, there remains sufficient lack of control over linguistic context that subjects may still avoid targeted structures or language use. Second, as with more decontextualized, experimental techniques, the social and psychological demands of the task, which can interact with norms of the L1 culture or personal anxieties, may lead to hesitance on the part of subjects to comply with the task demands. Finally, with very small children, some procedures require thoughtful adjustment to make the expectations for execution of the task clear and to stimulate the children to perform.

   Several recent reviews of SLA have cited the use of many of these data-collection methods (and other more controlled, experimental ones that we will review in the following main section), frequently with several used concurrently, in the study of speech acts (Ellis, 1994, pp. 169–71), tense and aspect development (Bardovi-Harlig, 1999), Swedish L2 morphology and word order (Pienemann and Håkansson, 1999), and the effects of pre-task planning on L2 performance (Ortega, 1999). Regrettably, however, those reviews did not propose to evaluate the methods per se. But, as in the case of naturalistic observation, if findings from the use of different procedures lead to results that are consistent with specific theoretical interpretations, then this amounts to a cross-task validation of the procedures. Table 22.3 summarizes a number of representative studies that have employed multiple tasks; when comparable, or divergent, results are obtained with their procedures, we can thereby draw conclusions about their validity.

3.2   (Un)structured interviews

Many researchers use (un)structured interviews in order to obtain what is often referred to as "free conversation," even though the data do not constitute "overheard" speech, rather speech by the informant(s) following sequences of questions by the researcher or other interviewer. Surprisingly, in most of this literature, there is little description of exactly what sort of protocols, guidelines, questions, or procedures are employed (for some guidelines, see Seliger and Shohamy, 1989, pp. 166–8, or qualitative research manuals such as Glesne and Peshkin, 1992, especially ch. 4). The reason appears to be that SLA researchers use interview data, unlike more naturalistic research, as a context within which to elicit particular target structures. The target linguistic focus intended can be word order and negation (Bolander, 1989), past time reference and propositional encoding (Sato, 1990), lexical development (Viberg, 1993), subject/topic prominence and pro-drop (Hendriks, 2000; Jin, 1994; Liceras and Díaz, 1999), existentials (Duff, 1993), tense and aspect (Bayley, 1994; and Klein

Table 22.3 Example studies using multiple elicited production procedures and tasks

| Study | Languages | | N | Procedures used | Topic focus | Design |
	Target	L1				
Bolander (1989)	Swedish Swedish Swedish	Finnish Polish Spanish	20 20 20	15-minute interview Picture description	Inversion Negation	4-month longitudinal pre-/post-test
Rintell and Mitchell (1989)	English (NS comparison group)	Various	34 37	Discourse completion: role play	Speech acts: requests, apologies	Comparative experiment
Sato (1985, 1990)	English	Vietnamese	2	Conversation with researcher Communication games Picture description, sequences Story translation Conversational translation Elicited imitation Oral reading, spelling	Past time reference Propositional encoding	Longitudinal sampling (every week for 10 months, less often over 2 more years)
Duff (1993)	English	Cambodian	1	Free conversation Personal narrative Picture description Story-telling	Possessive and existential	Longitudinal sampling (3 years)

Study	L1	L2	N	Tasks	Focus	Notes
Viberg (1993)	Swedish	13 languages	23	Free conversation Story-retelling (video) Play with flannel board	Lexical development: verb semantics	Cross-sectional and longitudinal development
Jin (1994)	Chinese	English	46	Oral interview Story-retelling Free composition	Subject- and topic-prominence in transfer	Lumped data from all tasks
Ioup, Boustagui, El Tigi, and Moselle (1994)	Arabic	English	2	Recipe description Accent identification Sentence translation Grammatical judgment Anaphora referent choice	NS-level accuracy in accent, syntax	NS judges, multi-task
W. Klein, Coenen, van Helvert, and Hendriks (1995)	Dutch	Turkish Moroccan	2 2	Conversation Personal narrative Story-retelling	Tense and aspect	Cross-sectional and longitudinal
Flanigan (1995)	English	8 language groups	23	Written answers to anaphora questions Story-telling L1 story-telling Sentence combining	Anaphora in relative clauses, reflexives	Correlations with proficiency measures
House (1996)	English	German	32	Interviews Pair interactions Role play Self-perception	Pragmatic gambits and competence	Instructional experiment over 1 term

Table 22.3 (cont'd)

| Study | Languages | | N | Procedures used | Topic focus | Design |
	Target	L1				
Myles, Mitchell, and Hooper (1999)	French	English	16	Picture description Story retelling Information gap (one-way) Problem-solving (two-way) Role play (group) Vocabulary test	Formulaic phrases, interrogatives	6 samples over 2+ years
Mackey (1999)	English	9 languages	34	Picture sequencing Picture description Picture differences Story completion	Question formation	Experimental task orders Output pooled across tasks
Rosa and O'Neill (1999)	Spanish	English	67	Multiple-choice Think-aloud	Past conditional (subjunctive)	Between-group Pre-/post-test
Helms-Park (2001)	English	Hindi-Urdu Vietnamese	47 45	Picture description-written (1 sentence, w/verb prompt) Multiple-choice response to question Grammatical judgment task	Lexis Causative verbs Picture context	Cross-sectional comparison

et al., 1995, which is representative of many of the ESF studies, as also found in Dietrich et al., 1995, and Dittmar, 1992), and speech acts and pragmatics (House, 1996; other studies in Kasper, 1996). More often than not, additional more specific tasks are included in the course of the interviews, such as picture descriptions (Bolander, 1989; M. Carroll, Murcia-Serra, Watorek, and Bendiscioli, 2000; Holmen, 1993; Sato, 1990), experimental elicitation tasks (Sato, 1990), personal narratives (Duff, 1993; Klein et al., 1995), story-retelling (Jin, 1994; Klein et al., 1995; Sato, 1990; Viberg, 1993), instruction-giving (Ahrenholz, 2000), or focused interactions and self-reflections (House, 1996).

With such varied procedures being employed, the same data can be analyzed for many different features, as is evident in the second stage of analysis of the ESF data (Perdue, 2000), where the data that had been originally analyzed for basic morphology and phrase structure, as in Dittmar (1992) and Dietrich et al. (1995), was then analyzed for negation, interclausal relations, and discourse structure. In few of the studies just cited, however, have researchers made a clear comparison between the analyses of the data gathered from the basic interviews and of those elicited by means of more controlled production tasks. Many primarily report on the more spontaneous conversation portions, and Jin (1994) claims (without displaying their separate frequencies) that since no significant differences occurred in the results for different tasks, the data are collapsed across tasks. Typically, of course, the nature of the specific target analysis is such that the key data are derived from the more focused task. A notable exception to this lack of comparison is Sato's (1985, 1990) longitudinal study of two Vietnamese boys, whose speech was analyzed after conducting various tasks – conversation, reading aloud, elicited imitation, and prepared text recitation. Sato's (1985) analysis of one of the boys' target-like final consonants and final consonant clusters revealed considerable cross-task consistency on the former by the end of the 10-month period. However, on final consonant clusters, on which his performance was overall much worse (generally lower than 30 percent accuracy, compared to nearly 70 percent for single consonants), there was substantial cross-task differentiation. The imitation task resulted in most target-like production, while conversation was the lowest (with consequent greater variability) for all but the final sample. Oral reading and text recitation led to gradual lower accuracy in later elicitations, possibly due to the competing demands on the boy to attend to other linguistic features than phonology. It would appear to be important for researchers to make more such comparisons across tasks, in order to assess learners' underlying competence with respect to their performance in less structured conversation. That is to say, more controlled elicited production procedures like the ones following in this section need to be used to complement data from interviews and personal narratives.[4]

3.3   Communication tasks

Although a large number of communication tasks have been developed and employed for the purpose of direct L2 instruction since the mid-1960s, their

use as a tool for eliciting L2 speech production in research has only gradually developed since the late 1970s. Much of the focus of early research making use of such tasks has been on the study of the effects of interaction on SLA (see the articles collected in Gass and Mackey, 1998; recent studies such as Shehadeh, 1999, using *picture dictation* and *opinion exchange*, and Van den Branden, 1997, using *picture description*; and previous reviews by Long, 1983; Pica, 1994). Yet such tasks have also been used to elicit learner production data for SLA analysis (some examples are given in table 22.3 above).

   Communication tasks vary considerably, from map reading, real-world sales exchanges and *information-getting* tasks, and *problem-solving* discussions, to narrower searching for differences in pictures, *picture description*, or sorting out the order of unordered picture sequences (see Yule, 1997, for representative types; and Norris, Brown, Hudson, and Yoshioka, 1998, for sample tasks within a schematic approach to the development of task-based performance assessment). Despite their variety, Skehan's (1998, p. 95) itemization of characteristics typical of communication tasks is essentially adequate:

  i   meaning is primary;
 ii   there is some communication problem to solve;
 iii there is some sort of relationship to comparable real-world activities;
 iv   task completion has some priority; and
  v   the assessment of the task is in terms of outcome.

There have furthermore been various proposals for describing and designing tasks, such as the following ones for the principal factors or dimensions of variation among tasks:

* *Nunan (1989)*: goals, input and materials, activity, teacher and learner roles, and context;
* *Pica et al. (1993)*: interactional activity (relationship and requirements) and communicative goal (orientation and outcome options);
* *Skehan (1998, p. 99)*: code complexity, cognitive complexity (familiarity and processing), and communicative stress;
* *Robinson (2001, p. 30)*: complexity (resource-directing and resource-depleting), conditions (participation and participant variables), and difficulty (affective and ability variables).

Choices from among the options for any given dimension can lead to differences in the complexity and linguistic demands on the subjects. Examples of the use of communicative tasks to elicit learners' speech for basic SLA analyses include Bolander (1989) and Ellis (1989) – picture descriptions for negation and/or word order; Ioup et al. (1994) – *description* of a favorite recipe in order to assess nativelikeness; Myles et al. (1999; see also Myles, Hooper, and Mitchell, 1998) – several one-way and two-way *information gap tasks* administered to French L2 secondary school children to elicit interrogatives; Wode (1999) – a

complex *problem-solving* task to examine grade seven English FL learners' lexical development; and Robertson (2000) – *one-way information gap* task to elicit English article use by Chinese L1 speakers. Also, from the earliest period of SLA research on communication strategies (see especially the collection of reprinted studies in Færch and Kasper, 1983; and more recent research in Kasper and Kellerman, 1997; as well as discussion on methodology by Bialystok, 1990, pp. 50–1), picture descriptions and many information-exchange tasks have been widely used to elicit learners' use of communication strategies.

As noted before, however, and as Doughty and Long (2000) exemplify, communicative tasks have been employed in SLA research in order to explore the influence of instruction and interaction on learners' acquisition, in which only certain narrowly defined target and interlanguage forms have been analyzed – along with more global measures of fluency, complexity, and accuracy – with respect to their improvement following task performance. Thus for example, studies employing communication tasks for the evaluation of effects of task dimensions on learner speech have included those on the amount of *planning time* allowed before performance: Crookes (1989), using Lego construction and *map description* tasks; Mehnert (1998), using instructions in telephone messages; Foster and Skehan (1996) and Skehan and Foster (1997), using information exchange and *decision-making* tasks; and Ortega (1999), using *picture/story-retelling*. Despite the differences among task types, relatively consistent effects of planning on these measures have been found. These include the result that, while fluency and complexity improve, accuracy does not change in any dramatic direction, except for some measures for which planning enhances accuracy (e.g., target-like use of Spanish noun modification in Ortega's study). In addition, some results point to increased effects of planning on some of these measures when the tasks vary on a dimension such as *complexity* (as in Foster and Skehan, 1996). Further, using a two-way interactive map-direction task of varying complexity, Robinson (2001) demonstrates, for instance, that task complexity tends to exert a favorable effect on accuracy and lexical variety, while diminishing fluency.

3.4 Story-retelling

The use of a pre-selected story presented to a learner in either written, picture, or video mode has been a widely used technique in (narrative) language elicitation. As early as Perkins and Larsen-Freeman (1975), non-dialog films have been used to elicit L2 English production. Perkins and Larsen-Freeman had to alter their elicitation, however, by interspersing the viewing of the film with questions, because their subjects were avoiding the morphological features they were investigating. More recently, this procedure has been used to elicit L2 lexical development (Hyltenstam, 1988, 1992, who used segments of Charlie Chaplin's *Modern Times*), and anaphoric reference and evidence of topic- or subject-prominence by several researchers, for example, Flanigan (1995; see table 22.3), who used a wordless picture story to elicit children's L1 and

L2 story-retellings in order to assess anaphoric reference; Jin (1994; also in table 22.3), who used silent cartoon films retold in L2 Chinese; and Polio (1995), who used the *Pear Film* to elicit L2 Chinese reference (see considerable L1 research using the "Pear Stories," starting with Chafe, 1980). Also, recently Rose (2000), who had earlier questioned the validity of discourse completion tasks (see the next section) in eliciting speech acts cross-culturally (Rose, 1994; cf. Rose and Ono, 1995, for an L1 Japanese comparison), employed cartoon strips to elicit politeness speech acts, and Gass, Mackey, Alvarez-Torres, and Fernández-García (1999) used video story-retelling (in fact, on-line simultaneous description of events) in a complex design to investigate the degree of improvement in Spanish L2 production (measured in terms of holistically rated speech, morphosyntax – *ser/estar* distinction – and complexity of vocabulary), after repeated viewings of the same video or repeated performances on different videos.

   Retelling has been used especially to elicit L2 tense and aspect features (as in the ESF studies – cf. Perdue, 1982, and others' use of *Modern Times*). An early L2 report of this sort analyzed in terms of foreground–background relationships in ESL was Tomlin (1984). And Bardovi-Harlig (2000, pp. 199–202) reviews this methodology especially with respect to the use of silent films and variations in the context and frequency with which they are presented, noting several advantages of this procedure: "[i] the sequence of events is known to the researcher . . . [ii] such narratives can be compared across learners . . . [iii] retell tasks may encourage some learners to produce longer samples than they would otherwise . . . [iv] the content of stories . . . may be manipulated" (pp. 199–200). To this one might add that, if a particular source for retelling is long, omissions of events are likely, unless the recipient of the retelling assists the subject. Among the disadvantages Bardovi-Harlig points to are: "there is still noteworthy variation in number of the tokens that learners produce . . . [and] certain types of predicates appear more frequently than others" (p. 201). The result of these disadvantages is that researchers have had to resort to narrower elicitation measures such as cloze passages in conjunction with retellings. A study that attempted to alleviate some of the disadvantages noted is that of Salaberry (1999) (who previously – 1998 – had used the *Pear Film* to elicit L2 French tense/aspect marking), as he presented segments of *Modern Times* individually in private to college students of Spanish L2. For purposes of motivation and authenticity, the students were then asked to relate the story to another student playing the role of a detective, who would, in turn, retell the events to a native or near-native listener.

3.5   *Role plays and discourse completion*

Some techniques have been developed primarily in order to elicit only certain domains of target structure. This is the case with *role plays* and *discourse-completion tests* (DCT), which have been used predominantly in L2 research to elicit data on pragmatic abilities in a variety of speech acts, with requests and

apologies being the most commonly studied. An anthology of such research is presented in Blum-Kulka, House, and Kasper (1989), in which a particular DCT the editors developed is employed in various studies. Further, Hudson, Detmer, and Brown (1995) provide a model for the development of DCTs, and Kasper and Dahl (1991) and Kasper (1998) provide extensive reviews of research methodology in L2 pragmatics.

In an early effort to make cross-procedure comparisons, Rintell and Mitchell (1989) explicitly contrasted results from oral role plays and written discourse-completion responses by ESL learners and native controls, with respect to length, variety, and directness of apologies and requests. They found that, while oral interaction led to less direct and longer speech acts than written responses, there were mainly similarities in the range of strategies used. Aside from the added length that indirectness leads to, the writer's opportunity to reflect on appropriate forms in the written response resulted in more concise, less hesitant responses. A quite similar result was found in a within-subject comparison in a recent study by M. Sasaki (1998) on Japanese L1 EFL learners. These findings suggest that while DCTs may provide legitimate data, they may also elicit a narrower range of pragmatic performance than learners are actually capable of.

Kormos (1999) conducted a slightly different but consistent comparison between role-play results and typical language-testing "interview" procedures for EFL in Hungary. In intra-subject comparisons of 30 learners being tested via a conversational interview and guided role play with the same interviewer, Kormos found that the role plays allowed much more opportunity for the subjects to display their conversational competence, in topic management in particular. That study is but a more recent investigation into a long-standing issue, that is, the adequacy of a structured or unstructured oral proficiency interview to assess a learner's ability.

3.6   (Simulated) oral proficiency interview

Research has also been conducted on the use of (simulated) oral proficiency interviews to elicit overall learner proficiency or more specific L2 performance. Space does not permit a more extensive review of the history of the OPI (and ACTFL scales – see Clark and Clifford, 1988), or the British-based Cambridge Certificate Examination (UCLES, 1990), but, as these formats include a combination of the sorts of semi-structured tasks that we have been reviewing, it seems reasonable to consider combinations of such procedures as legitimate measures as well.

It is sufficient to note that such formal examinations, besides playing a major role in educational and occupational placement, have been used by researchers, as well, as a comparative standard in evaluating learners' development in second languages. Yet, criticism of their validity as specific measures of SLA has mounted over the years, with particularly strong arguments arising from SLA-oriented analyses of topic- and conversational management of the interaction during such

interviews (as seen above in Kormos, 1999; see also articles in M. Johnson, 2000; S. Ross and Berwick, 1992; Valdman, 1988; Young, 1995; Young and He, 1998; and Young and Milanovic, 1992). The upshot of these critiques is that, as the OPI/UCLES procedures involve interaction between a status-dominant and guiding interviewer, the nature of the language performance that results is less than optimal, especially with respect to the elicitation of a full range of verbal and pragmatic competence on the part of the learner. Similarly, because a SOPI does not involve a real interlocutor, performance on it in comparison to an OPI results in different discourse and strategic use of the L2 (Koike, 1998).

3.7   *Stimulated recall*

The final general type of elicitation that falls within the group of more naturalistic, less decontextualized procedures is that of *stimulated recall,* which is a cover term for what are described in the literature as *think-aloud* (protocols), *introspective/retrospective interviews*, *verbal reports*, and *cued recall*, among other terms. These all have in common the elicitation from L2 subjects not of a direct linguistic performance, but of a more reflective, metalinguistic analysis or description of their language use, and internal representations or reconstructions of what they have said and how they arrived at their performance. These methods are reviewed most currently in Gass and Mackey (2000), but L2 researchers' interest in evaluating them dates back over 20 years to seminal L1 research, such as Ericsson and Simon (1980), and L2 surveys, such as Cohen and Hosenfeld (1981), and Færch and Kasper (1987a, 1987b). Cohen and Olshtain (1994, p. 148) suggest that verbal reports are vital to the validation and interpretation of learner behavior at each stage in the cycle of research on speech acts, that is, from (i) observation, to (ii) role play, to (iii) discourse completion, to (iv) learner acceptability checks.

Færch and Kasper's (1978a, p. 11) classification of the methodological framework that can be used to analyze most forms of stimulated recall is useful here (see table 22.4). Stimulated recall has been used to elicit learners' awareness and explanations of such phenomena as their tense and aspect use (Liskin-Gasparro, 1998); general grammar rule awareness and correction (P. Green and Hecht, 1992); vocabulary knowledge and decoding or translation processes (Dechert and Sandrock, 1986; Fraser, 1999 – who used both retrospection about processing and cued recall as a test; and Paribakht and Wesche, 1999 – who used both concurrent and retrospective introspection, along with training tasks and other measures of vocabulary comprehension; cf. also Wolter, 2001); and metalinguistic processes and awareness of noticing forms in input (e.g., Ellis, 1991, on learners' processing of grammaticality decisions; Leow, 1997; Rosa and O'Neill, 1999; and most of the research collected in Schmidt, 1995a; for example, Jourdenais et al., 1995, who provide a detailed description of their procedures and analytical coding), and reflections on planning processes (Ortega, 1999).

While introspective methods have gained a high degree of acceptability in L2 research, as most of the above studies have displayed findings that shed

Table 22.4   Abridged framework for classification of introspective methods

1   Object of introspection:
    linguistic, cognitive, affective, social
    declarative vs. procedural knowledge
    modality (spoken vs. written, receptive vs. productive, etc.)
    continuous process vs. specific aspect
2   Relation to concrete action:
    concrete/specific to non-specific/abstract
3   Temporal relation to action:
    simultaneous, immediately consecutive to delayed.
4   Participant training:
    ± instructions, ± training
5   Elicitation procedure:
    degree of structure
    ± media support
    self-initiated vs. other-initiated
    ± interaction between informant and experimenter
    ± integration with action
    ± interference with action
6   ± Combination of methods

Source: Færch and Kasper (1978a)

additional light on concurrent measures (such as vocabulary comprehension measures and other evidence of learner processing in uptake of input, for example), there remain concerns about the extent of application of these methods. Cohen (1991), an advocate of verbal reporting, following the work of Ericsson and Simon (1984), recognizes a number of criticisms of the method, which he summarizes (pp. 136–8): data from verbal reports (i) do not access unconscious skill learning; (ii) may simplify a particular process just by virtue of accessing it; (iii) may distort the original process by virtue of constructing a verbal report, and, therefore, become distant from the underlying events; (iv) may repress information (e.g., for social-affective reasons); (v) may intrude on the process investigated; and (vi) may also distort the reality of the process by using a different (target or native) language for the report. Jourdenais (1998, 2001) has documented in an SLA study influences of type (v).

4   Experimental Procedures and Tasks

The final principal group of data-collection procedures (as in the procedures in table 22.1 in the right-hand column) will be referred to here as "experimental" procedures and tasks, as they tend to be employed under more controlled

conditions, with elicitation of L2 production or performance on perceptual-receptive tasks, with less communicatively driven and decontextualized constraints. Although some of these may also be considered broadly as "elicited production," as in the previous section, we are making the distinction here with respect to the more meaningful context and generally lengthier nature of the language production process or result that obtains in those procedures discussed previously, compared to the shorter, controlled productions typical of experimental procedures.

The current set of methods (see table 22.5, which lists representative studies for each grouping of them) includes, first, a variety of "on-line" (often computer-managed) target language processing tasks like those frequently used in L1 psycholinguistic research, such as *signal detection* tasks (of target phonemes, lexis – *word recognition*, or semantic content), *sentence-* (or other linguistic form) *matching, verification, reading* and *discrimination*, and certain *decision* (yes/no identification) tasks. For a review of several such tasks in L1 psycholinguistics, see Murdock (1982); Olson and Clark (1976); and more recent methods in McKee (1996); cf. L2 discussion by Simard and Wong (2001).

Second, many tasks that involve more cognitive processing of stimuli, and possibly learning, have been used in many SLA experiments, such as *utterance completion, elicited imitation, word association, word/list memorization, elicited translation, sentence manipulation* (*combining, transformation*), *recall* of linguistic elements following exposure to them, *rule learning, act-out* of presented structures (as a measure of comprehension) and a large variety of *discrete-point* and *cloze-type tests* measuring L2 productive knowledge of morphology, syntax, discourse, and lexis.

Finally, a very large range of methods that involve more extensive use of reflective capacities and access to higher levels of L2 knowledge have been used to elicit learners' *metalinguistic knowledge* (*grammaticality judgment tests – GJT*), *magnitude scaling* and other *judgments* (ratings), *paired comparisons, card sorting*, and ability to *express* (grammatical) *rules* or *lexical definitions* and apply them in *correction* or *editing* tasks. A number of chapters in McDaniel, McKee, and Cairns (1996) provide extensive discussions on some of these techniques as employed in child L1 studies, especially on elicited production (of the experimental sort), elicited imitation, picture selection, act-out, and judgments (including grammaticality).

It is helpful to bear several methodological issues in mind in this section. First, for the sake of distinguishing SLA linguistic performance/competence from many other topics dealt with in the L2 literature, in this review we are distinguishing the linguistic from the non-linguistic focused targets of the above procedures or methods. This is to say that many of the above procedures have been used to elicit data from subjects on such matters as their attitudes (as in many attitude surveys or scales and the use of the matched guise technique), anxiety, reflections on learning strategies, and general cognitive measures, such as aptitude or memory functioning (e.g., digit span, musical memory, analogical ability, cognitive styles). While much of this research has been vital to

Table 22.5 Example studies using experimental procedures and tasks

Study	Languages Target	L1	N	Procedures used	Topic focus	Design analysis
On-line tasks:						
Juffs and Harrington (1995)	English	Chinese	26	Word-by-word reading Grammaticality judgments: (whole and paced sentences) ["Moving window"]	± grammaticality of *wh*-extraction	Cross-task comparisons Reaction time and accuracy measures
Juffs and Harrington (1996)	English	Chinese	(25?)	Same procedures as in Juffs and Harrington (1995)	± grammaticality of *wh*-extraction Garden Path sentences: ± infinitive, ± inside theta	Same as in Juffs and Harrington (1995)
Verification, decision, and reading tasks:						
Trahey and White (1993)	English	French	54	Sentence manipulation Correctness of sentence pairs Grammaticality judgment Cartoon and verbal stimulus: sentence production	Effects of instruction Verb-movement (adverb placement)	Between-group comparisons Accuracy and rate of adverb placement
Y. Sasaki (1994)	Japanese	English	30	Choice of first noun as subject/actor	Competition Model: animacy, case, word order	Mixed ANOVA on choice proportions

Table 22.5 (cont'd)

Study	Languages		N	Procedures used	Topic focus	Design analysis
	Target	L1				
Matsumura (1994)	English	Japanese	110	Choice of referent for reflexive	Reflexive binding Proficiency differences	Between-group comparisons Accuracy
VanPatten and Oikkenon (1996)	Spanish	English	59	Picture selection Picture description: prompted completion	Object word order Effects of instruction	2 experimental, 1 control group comparisons
DeKeyser and Sokalski (1996)	Spanish	English	82	Comprehension: object selection Production: sentence completion, translation, answering questions	Direct object clitics Conditional Effects of instruction	2 experimental groups, pre-test/post-test
Ying (1996)	English	Various	45	Preference selection of NP or VP	Ambiguous NP/VP attachment ± referential context, lexical context type, aural vs. reading input	Within-subject correlations Choice means
Beck (1998)	German	English	48	Sentence-matching Elicited oral translation	Verb-raising ± grammaticality ± inversion	Within-subject correlations Reaction time and grouping by translation performance

Study		N	Tasks	Variables	Measures
Salaberry and López-Ortega (1998)	Spanish / English	74	Cloze completion text, Multiple choice completion, Written narrative from picture stimulus	Cross-task performance, ± communicative pressure, Attention to form, Articles, pro-drop, aspect	Between-groups comparisons, Accuracy
Moyer (1999)	German / English	24	Reading: word lists, sentences, paragraph text, Topical conversation	Motivation effects on high-level pronunciation	Rating of nativeness, confidence of judge
Bongaerts (1999) [3 studies]	English / Dutch, English / Dutch, French / Dutch	22, 21, 27	Reading aloud: sentences, words, Personal narratives	Pronunciation: high- vs. low-proficiency learners, age of acquisition	Accent ratings, Comparability with native speaker ratings
Ju (2000)	English / Chinese	31	Preference for active/passive setence	Active/passive, Lexical (± transitive), Cognitive effects (± external), + Context	Error rate within-group
EI:					
Grigg (1986)	English / Various	18	EI ± stimulus repetition, Dictation, grammar correction, rule statement	Plural -S, article, Relative pronoun marker	Within-subject correlations, Mean accuracy
Munnich et al. (1994)	English / Japanese	12	EI oral vs. taped, GJT: oral vs. taped	Relative clause structure, ± Grammatical	Within-subject correlations, Accuracy, error analysis

Table 22.5 (cont'd)

Study	Languages		N	Procedures used	Topic focus	Design analysis
	Target	L1				
Yang and Givón (1997)	Keki (artificial)	English	29	Word recognition Lexical decision-priming Word translation EI ("recall") ± delay GJT, Picture description Narrative translation and comprehension	Simplified (pidgin) vs. grammatical input	Longitudinal accuracy Reaction time Lexical/grammar recall
Ortega (2000)	Spanish	English	16	EI, GJT Spanish simulated oral proficiency interview Compositions	Cross-task comparisons SLA measures Syntactic/lexical complexity	Longitudinal (pre-test/post-test) Accuracy, ratings
Metalinguistic tasks (GJT, magnitude estimation):						
White (1986)	English	Spanish/ Italian	34	Sentence transformation (questions from statements)	Pro-drop ± grammaticality: VS word order, subject pronoun omission *Wh*-extraction	Between-group comparisons Accuracy
		French	37	GJT		
Bley-Vroman, Felix, and Ioup (1988)	English	Korean	92	GJT	*Wh*-movement: ± grammaticality	Group % accuracy Consistency within-subject
Robinson (1994)	Samoan	English	29	GJT	Noun incorporation ± Grammaticality Empty category principle	Within-subject correlations Reaction time, Accuracy, certainty

Study	L1	L2	N	Task	Features	Analysis
Yuan (1995)	Chinese	English	102	GJT with magnitude estimation	Proficiency level Base-generated topics ± Gapped NPs	Between-group comparisons Accuracy
Yuan (1997)	English	Chinese	159	GJT with magnitude estimation Animacy	Proficiency level Null subjects and objects	Between-group comparisons Accuracy
Mandell (1999)	Spanish	English	91	GJT Sentence construction ("dehydrated" sentences)	Proficiency, S-V inversion ± *Wh*-questions Adverb placement	Between-group comparisons Accuracy

Complex tasks (paired comparisons, rule statement, editing):

Study	L1	L2	N	Task	Features	Analysis
White (1989)	English French	French English	95 151	GJT (paced, scaled) GJT (multiple choice) Sentence comparisons (preference)	Proficiency/age levels Adjacency (adverb placement)	Between-group comparisons Accuracy Cross-task comparisons
Liceras (1989)	Spanish	French English	32 30	GJT Correction Translation	Proficiency Pro-drop parameter: inversion, that-trace, complementizer	Cross-sentence type Between-group comparisons
Alanen (1995)	Finnish	English	36	Sentence completion Comprehension test Word translation GJT with correction or explanation Rule statement Think-aloud	Input enhancement Rule presentation Locative suffixes Consonant alternation	Between-group comparisons Pre-test/post-test

Note: ANOVA = analysis of variance; EI = elicited imitation; GJT = grammaticality judgment task; L1 = first language; N = number of subjects; NP = noun phrase; S = subject; V = verb; VP = verb phrase.

understanding the psychological processes of language acquisition, these measures do not provide direct information about learners' L2 development. Second, we separate the procedures here from various *analytical measures of performance* that might be associated with them (see Norris and Ortega, this volume). Aside from more detailed linguistic analyses, many of these procedures are typically assessed with respect to normative *accuracy*, or speed of access and degree of neurological activity in responses (*reaction time*, as in Robinson and Ha, 1993, and *event-related potentials – ERP*), and, depending on the particular measure, different results may ensue, but how these measures are applied will not be our main concern here.[5]

   Finally, as in any experimental design and methodology, there are many *conditions* for the elicitation, such as those described previously for communicative task dimensions (e.g., time constraints, such as planning time, repetition of stimuli, aural or visual context, nature of the instructions for the task, prior learning tasks, and structuring of input). Only insofar as there have been notable effects associated with a given measure or condition when used in a procedure will we call attention to it.

4.1   Advantages and disadvantages

As Cook (1986, p. 13) put it succinctly: "Controlled data has the advantage that it yields the information we are looking for. It has the disadvantage of artificiality. . . . [T]he behaviour that is studied must correspond with something outside the laboratory if it is to have any ultimate relevance. There is, then, a continual tension between 'internal validity' . . . and 'external validity.'" Following the outline of advantages and disadvantages already seen for the previous two principal groups of data-collection methods, it should be evident here that experimental tasks are very different from naturalistic methods, but they share some of the advantages of elicited production. With respect to the four advantages noted above for elicited production – more tailored targets, more mechanical administration, a wider range of access to subjects, and ease of scoring – these most controlled tasks tend to achieve even more or better outcomes. Moreover, because of the degree of control over target forms, they alleviate one of the disadvantages of elicited production: well designed and analyzed experimental techniques should be able to determine whether target forms were avoided or not. At the same time they elicit less contextualized and less extensive L2 production, thereby also exaggerating at least two of the disadvantages of elicited production: that subjects, especially young children, may find experimental techniques too alien or off-putting, and performance will be avoided.

4.2   On-line methods/word (etc.) recognition

In first language psycholinguistic research, the use of a large number of on-line language processing procedures is very common, yet such procedures have

rarely been employed in SLA research. Technologically sophisticated methods in SLA have principally been incorporated in research on L2 phonetics and phonology perception and production (see Cebrian, 2000; Hardison, 1996; Leather, 1999; Major, 1998; Watson, 1991), and especially psycholinguistic research on bilinguals (see, e.g., the special issue on bilingualism in *Language and Cognition*, 4 (1), 2001; De Bot et al., 1995; De Groot and Kroll, 1997; Schreuder and Weltens, 1993; Weber-Fox and Neville, 1999). The greater availability of computer hardware and software for such research is now allowing more researchers to conduct experiments in speech processing and memory, but many of the L1 methods and procedures described in McKee (1996), such as on-line search and cross-modal priming, have not been widely adopted in the SLA literature.

An early example of *word recognition* (signal detection) in a Spanish L2 and English L1 lexical search task is Meara's (1986) study of learners' progress (determined by reaction time in recognition) over time in acquiring lexis during an instructional program. As an example of *morphological* recognition, Leow (1993) used a multiple choice (M/C) recognition question to assess whether learners in a ± simplified input contrast had recognized either the present perfect or present subjunctive L2 Spanish forms in a reading passage. Recently, Yang and Givón (1997) used word recognition and lexical decision priming to assess learners' acquisition of an artificial language taught via both simplified and normal input procedures.

Juffs and Harrington (1995, 1996) conducted an on-line processing study of *wh*-extraction and Garden Path sentences presented by computer, with both accuracy and reaction time as dependent measures. The target decision for Chinese learners of English was whether or not sentence strings were grammatical ("possible/impossible"), but Juffs and Harrington compared learners' decisions both after reading the sentences as whole sentences, and after being presented the sentences in a "moving window" mode (see Just and Carpenter, 1987; Rayner, 1983), a word-by-word presentation according to the subjects' own pace of reading. The researchers then assessed reading times through the key grammatical segments, as well as decision times on their judgments. (See also Hoover and Dwivedi's, 1998, study of advanced French learners' window-paced reading of causative sentences.)

4.3   Decision tasks: sentence matching, verification, discrimination

It is likely that the largest number of researchers in SLA experimental studies have employed some variant of *decision tasks*, in which subjects are instructed to make some decision among options (categories, pictures, sentences, multiple choices, preferences for appropriateness, referents, and so on). A basic on-line task, for example, is that of *sentence-matching* (proposed for L2 by Bley-Vroman and Masterson, 1989; see also Eubank, 1993), which involves a time-controlled presentation of two sentences (simultaneously or in sequence), with the subject having to decide whether the two are the same or different. Analyses are

based on the reaction time to decide on matching ("same") sentences, so that eventual contrasts are made only for the grammaticality differences between whatever syntactic phenomena are studied. For example, Beck (1998) used such a procedure to test German L2 verb-raising. For this, however, because she predicted a developmental effect dependent on subjects' knowledge of German agreement and S-V inversion rules, she used scores on an *elicited translation* task to group subjects developmentally. Some SLA researchers have made use of a *preference choice* on pairs of sentences (a form of *discrimination* – and also *metalinguistic judgment* – see below) illustrating target features in order to elicit (receptive) comparative judgments of learners' developing sensitivity. This procedure avoids forcing learners to make absolute judgments (of, say, grammaticality), because in many cases, it is the relative contrast between items that is of interest, rather than a judgment of them one at a time on an absolute scale. Trahey and White (1993), White (1989), and White, Spada, Lightbown, and Ranta (1991), for instance, used this task to elicit French (and English FSL in White, 1989) learners' ratings of ESL target sentences involving adjacency conditions (in adverb placement), and question formation. Lakshmanan and Teranishi (1994) used the task to elicit judgments on reflexive binding in L2 English, and Duffield and White (1999) used it in combination with grammaticality judgments to assess L2 knowledge of Spanish clitic placement.

Probably the most common passive/receptive form of experimental decision task is the *M/C response selection* (among picture options, responses to questions, categorical choices, etc.). This is illustrated in many studies exploring the Competition Model of language learning/processing (cf. Gass, 1987; MacWhinney, 1987, 1997; cross-linguistic and L2 studies in *Applied Psycholinguistics*, 8 (3); and MacWhinney and Bates, 1989). In the most frequent application of this paradigm in L2 studies, where word order preferences (e.g., SVO vs. VSO), case markings, and animacy preferences differ across the typologically distinct languages, simple lexical combinations of grammatical cues and word orders are presented aurally to learners, who are to respond by selecting the actor or subject of the "sentences." Ungrammatical strings are also used to force learners to demonstrate their preferences for specific cues, and learner proficiency level is typically included as an independent variable. Examples of this procedure are Harrington (1987), with Japanese L1 learners of English; Kempe and MacWhinney (1998), with Russian and German L2; Rounds and Kanagy (1998), with child L2 immersion learners of Japanese; Y. Sasaki (1994), with adult English L1 learners of Japanese; and Su (2001), with both English and Chinese L1 learners of the other group's language. These studies lend themselves especially to comparisons between native speaker and learner performance in the source and target languages, as the contrasting high and low percentages of choice of "first noun" as agent/subject reveals the degree of interpretation based on processing strategies derived from the L1, L2, or the learner's interlanguage. Kempe and MacWhinney (1998) provide a very detailed accounting of the quantitative analysis that is possible with both the proportion of first noun choices and reaction time measures.

Other examples of (M/C) response selection (often by way of a picture choice) are choice of referent or definition for lexical items (Ellis and He, 1999; Rott, 1999); phrasal verbs (Laufer and Eliasson, 1993); anaphora (and reflexives – Eckman, 1994); reflexives (Matsumura, 1994; Thomas, 1992); object pronouns (DeKeyser and Sokalski, 1996; VanPatten and Oikkenon, 1996); phrasal prosody (Harley, Howard, and Hart, 1995); verb tense/aspect (Salaberry, 1998; Salaberry and López-Ortega, 1998); and NP/VP attachment (Ying, 1996). Some researchers present lengthier response alternatives, such as Tanaka and Kawade (1982) with politeness strategies, Ju (2000) with active/passive sentences, and Helms-Park (2001) with description of causative situations.

4.4   *Reading aloud, utterance completion, elicited productions*

A further wide range of tasks has been employed in SLA experiments that prompts learners to produce forms in a more active and sometimes extended fashion. These tasks include *reading aloud* tasks to assess all levels of speech production, but particularly phonological and lexical abilities; *utterance completion* to assess syntactic and lexical competences; and other *elicited production* (from picture stimuli, questions, or other prompts) with constrained options, but which allow for free access to the learners' knowledge base and more extended speech routines. Some of the earliest SLA research typically adopted such procedures from L1 research, as in the case of d'Anglejan and Tucker's (1975) use of Chomsky's (1969) research methods on questions about infinitive complements, with and without *picture choice*, and Fathman's (1975) use of Berko's (1958) type of materials for *picture prompts* for morphology and syntax. Recent examples of these are *text*, *sentence*, and/or *word list reading* for phonological/lexical evaluation (Bongaerts et al., 1997; Flege et al., 1998; Major and Faudree, 1996; Moyer, 1999; Riney and Flege, 1998) and syntactic speed of access to French L2 syntax (Hoover and Dwivedi, 1998); *sentence completion* or *blank filling (cloze)* prompting French aspect (Harley, 1989);[6] phrasal verbs (Laufer and Eliasson, 1993); Spanish conditionals and object pronouns (DeKeyser and Sokalski, 1996); instrument attachments to either NPs or VPs in ambiguous sentences (Ying, 1996); *picture cued descriptions* (with occasional verbal prompt) eliciting lexical items (Snodgrass, 1993); head NPs of relatives or *wh*-questions (Wolfe-Quintero, 1992); English *adverb placement* (Trahey and White, 1993); Spanish clitic objects (VanPatten and Cadierno, 1993; VanPatten and Oikkenon, 1996); causatives (Helms-Park, 2001); and past hypothetical conditionals (Izumi, et al., 1999).

4.5   *Elicited imitation*

Elicited imitation (EI) is a special sort of elicited production that has gained considerable research attention since its early use in the child language research of the 1960s (Fraser, Bellugi, and Brown, 1963; Slobin and Welsh, 1973; see

Bernstein Ratner, 2000, and Lust, Flynn, and Foley 1996, for reviews of L1 research; see J. Schumann, 1978, and Swain et al., 1974, for early L2 research; and see Bley-Vroman and Chaudron, 1994, for review of L2 research). The procedure involves preparing a stimulus string (usually a sentence, although lower- or higher-order texts have been used to control length and discourse context) that illustrates some grammatical feature (only occasionally has phonology been studied in this fashion), and subjects are instructed to repeat exactly what they hear. The assumption of the procedure is that success at exact imitation demonstrates the subject's possession of the grammatical (or lexical, etc.) feature in her or his knowledge store, unless the stimulus is too short and thereby allows for echoic repetition. Failure to repeat exactly, and any modifications or transformations of the stimulus, can be taken to represent the limits or other representations of the subject's grammatical competence. Thus, the procedure, being apparently relatively simple to prepare, can be used to elicit potentially a very wide range of target structures. Both grammatical and ungrammatical structures can also be tested with the method (e.g., Masterson, 1992), in order to detect greater subject difficulty or ability to reconstruct grammatical forms.

There have been numerous applications of this technique in L2 research: Grigg (1986) used a single and a repeated stimulus to elicit ESL morphology for comparisons with other measures; Flynn (1987) compared ESL imitation of various structures with L1 Spanish and Japanese adults; Verhoeven (1994) elicited L2 Dutch and L1 Turkish sentence imitations in bilingual children in order to establish their bilinguality; Munnich et al. (1994) compared different procedures to elicit repetitions with grammaticality judgments on relative clause structures; Scott (1994) compared Spanish L2 EI performance with other auditory and fluency measures in order to differentiate age differences among learners; Yang and Givón (1997) employed EI to test learning of their artificial Keki language; Roebuck, Martínez-Arbelaiz, and Pérez-Silva (1999) provide their elicitation stimuli in a study of English L2 complementizer phrase production ("filled CP") by Spanish and Chinese L1 speakers; and Ortega (2000) evaluated Spanish L2 EI as a pre- and post-instructional period measure of SLA change. Chaudron (forthcoming) reviews this L1 and L2 research in detail.

4.6   Elicited translation

An alternative manner in which to attempt to constrain the elicitation of specific target structures is to use *translation* sentences. An early use of translation was the study by Perkins and Larsen-Freeman (1975) for Spanish L1 to English L2 morpheme production. Among various measures for eliciting L2 production of Arabic from their near-native learner, Ioup et al. (1994) used an English-to-Arabic translation of selected syntactic structures. Other researchers include Snodgrass (1993), who used *word translation* in comparison with *picture naming*; de Graaff (1997), who used this technique among his measures of learning from instruction in the artificial language eXperanto; and similarly, Yang and

Givón (1997) employed an L2-to-L1 translation task for both words and an entire narrative (sentence by sentence). Beck (1998) used a translation task as a grouping (developmental stage) variable when analyzing her results for sentence-matching in L2 German, and Pérez-Leroux and Glass (1999) used *sentence translation* following picture stories to elicit L2 Spanish ± anaphora. See Malakoff and Hakuta (1991) for a discussion of the role of metalinguistic skill in translation, where they found that translation skills were distinct from source or target language proficiency in helping to predict translation accuracy for words and sentences.

4.7   Word (paired) association, and lexical assessment

A wide range of measures of vocabulary knowledge has been employed in SLA research, following the extensive research of this nature in the L1 psycholinguistic literature (see De Groot and Keijzer, 2000; Kroll and De Groot, 1997; P. Nation, 2001; Smith, 1997; and the special issue of *SSLA*, 21 (2), 1999, on vocabulary acquisition). Wolter (2001), for example, illustrates a number of approaches to assessing word knowledge in bilinguals via *word association* and depth of knowledge (cf. Wesche and Paribakht, 1996); Schmitt and Meara (1997) assess word knowledge via both receptive and productive measures of Japanese L1 learners' knowledge of English suffixation; and Scarcella and Zimmerman (1998) use a knowledge/familiarity rating to determine learners' knowledge of academic vocabulary.

4.8   Discrete-point tests

It is natural that many SLA researchers would employ tests with discrete-point assessment of knowledge of particular target forms, for these are easier to develop and standardize. This overview cannot, however, begin to examine the extensive number of tests and test batteries that have been developed and applied to such purposes, as such a review belongs broadly to the domain of language testing. Such tests do fortunately tend to be subjected to more rigorous analytical procedures than many of the measures described thus far, with respect to their reliability and validity.

4.9   Sentence manipulation (combining, transformation)

A number of researchers have employed *sentence manipulation (combining, transforming)* tasks, which frequently resemble the typical sort of classroom construction exercise used to guide learners in creating interclausal links, anaphora, and many other targets. Such tasks have been used to elicit learners' production of relative clauses, by providing two separate clauses which are to be

combined into one (see Doughty, 1991; Eckman, Bell, and Nelson, 1988; Gass, 1980; Hamilton, 1994), and adverb placement, for example, has also been a target (Trahey and White, 1993). Doughty (1991) also elicited similar relative clauses with a more constrained *sentence completion* task in which a portion of the targeted structure was used as a prompt to elicit the embedded relative clause.

4.10   Act-out

One method that has been very widely employed in child L1 acquisition research but very rarely in SLA research is the *act-out* task (see Goodluck, 1996). This method for assessing comprehension has been used with children primarily because of their lower capacity in productive language relative to receptive abilities, but also because it can more accurately assess matters relative to anaphora, missing subject or object constructions, relative clauses, passives, or *that*-trace in *wh*-questions. SLA researchers, however, have not widely availed themselves of such a measure, with the exception of Flynn (1987), who conducted research on anaphora in comprehension, and recently Finney (1997), who assessed various factors that influence the interpretation of gaps and referents in purpose clauses ("[in order] to" clauses). Flynn (1987), for instance, used sentences with temporal relations and physical movements expressed via main and adverbial subordinate clauses, in order to determine which referents subjects were interpreting for pronouns occurring in either the main or subordinate clause. Learners' actions with objects named in the sentences demonstrated their choice, whereas EI productions of similar sentences could not provide any such indication of contextualized interpretation.

4.11   Metalinguistic knowledge tasks

A final major group of elicitation techniques can be treated together as involving the expression, application, or invoking of learners' "metalinguistic knowledge." That is, instead of direct elicitations of language produced or interpreted, learners are presented with linguistic stimuli, in reaction to which they must make other active decisions, ratings, comparisons, and revisions about the form or meaning of the stimuli. We have seen one variant of this method in the simple "preference" decision task, which does invoke a judgment of acceptability or "correctness." Others include: grammaticality judgments, ratings, paired comparisons and card sorting, rule expression and definitions, and editing and correcting. These procedures have been widely used in the SLA research, particularly because adult learners are thought to be more readily able to carry out the often more complex tasks and decisions involved. As proposed by Bialystok and Ryan (1985), several of these tasks might be viewed as falling on a continuum of increasing use of "analyzed knowledge," from grammaticality judgments to locating ungrammaticality, to correcting ungrammaticality, then explaining ungrammaticality, and stating a rule that is violated. Ellis (1991) employs this notion in a table (p. 163) which suggests the

differential nature of the possible responses for each type of metalinguistic operation, mainly in terms of the extent of verbalization of judgment – from "discrimination" of well-formedness, to "location," then "correction," then "description" of errors. These features thus can be used to differentiate (i) intuitive from analytical responses, (ii) non-verbalizable from verbalizable knowledge, and (iii) recognition from production.

The most familiar and widely used form of metalinguistic knowledge elicitation is the *grammaticality judgment test (GJT)*, which was adopted by theoretical linguists in the 1960s and afterwards as a standard, albeit solitary and individualized, assessment of the acceptability of surface syntax (see studies in Greenbaum, 1977; J. Ross, 1979; a review by McDaniel and Cairns, 1996; and a complete review of this research, also L2-related, in Schütze, 1996). The second language acquisition literature was reviewed by Birdsong (1989), Chaudron (1983), Ellis (1991), and Sorace (1996); see also a mini-analysis of UG-based grammaticality studies by Zobl (1992). As in the analysis of communicative tasks (see the previous section), there are many features of GJT that can influence their effectiveness, of which Ellis (1991) mentions at least the following:

- *target items*: syntactic, lexical, phonological;
- *order of presentation*: ± randomized;
- *distractors*: ± other target structures;[7]
- *medium of presentation*: written/aural;
- *complexity*: controlled or not;
- *contextualized*: in discourse or not;
- *nature of response*: binary, multiple choice, preference, ranking;
- *immediacy of response*;
- *timed response*;
- *familiarization with task requirements.*

Under "familiarization," one would note factors such as training and task instructions, not to mention inter-subject differences in their interpretation of the notions "grammaticality," "acceptability," "correctness," and so on. See the discussion of the acceptable/grammatical contrast in Birdsong (1989), Chaudron (1983), Gass (1994), and Sorace (1996), as well as the very helpful provision of their complete instructions and test items in Bley-Vroman et al. (1998).

With respect especially to the nature of the response, the scale or options used have varied considerably across studies. As Gass (1994) demonstrates, the differences in response format can lead to differences in the interpretation of findings.[8] In order to enhance the sensitivity of grammaticality judgments, researchers have tended to avoid singular or dichotomous decisions. Nonetheless, researchers such as Munnich et al. (1994, p. 231) state that they only "recorded whether or not [the subject] believed the sentence to be grammatical," while most researchers elicit other actions or decisions: for example, Gass (1979, 1980) asked subjects to correct ungrammatical sentences; Lightbown,

Spada, and Wallace (1980) had them label sentences as "C[orrect]," or provide error *corrections* of them if they were judged not correct; and a similar correction-only procedure was used by Liceras (1985). These researchers favor at least three-point decisions to allow for a "not sure"/"I don't know" choice, which is sometimes: (i) counted categorically, occasionally with a correction procedure in order to verify the subjects' attention to and conception of the source of error[9] (Ellis, 1991; Mandell, 1999; Seliger, 1989; Towell, et al., 1993; White, 1986); (ii) considered as "incorrect" (Bley-Vroman et al., 1988); (iii) eliminated from analysis (Felix and Weigl, 1991; Shirai and Kurono, 1998, who had subjects judge each one from among four possible responses using Japanese L2 tense and aspect – a multiple-choice procedure also adopted by White, 1989); (iv) viewed as correct (Coppetiers, 1987); or (v) counted independently as a measure of certainty of response (Robinson, 1994). For the most part, indications of uncertainty in these studies have been limited to less than 3 percent of the responses (but as high as 29 percent in Robinson's study), so that researchers have been confident in ignoring such responses. Likert scale ratings of four or more, and up to even ten points (Gass, 1994; Inagaki, 1997; Papp, 2000; Schachter and Yip, 1990; Wang and Lee, 1999), have been used more in recent years, and White (1989) employed an unscaled line with ± correct polarity, on which subjects were to place a stroke wherever they preferred.

An alternative to a judgment on each sentence, that is, *preference comparisons* between pairs of sentences, was used in early studies, as already mentioned (e.g., studies by White and her colleagues). This procedure, which is a variation on the psychometric method of *paired comparisons* (Edwards, 1957; B. Green, 1954) has been used partly in order to avoid forcing absolute judgments, which researchers have suspected of being less reliable, as different subjects may use very different standards, but also because in many cases, the point of comparison is to determine subjects' *relative* sensitivity to variant structures of the same sort, and within-subject reliability can be increased thereby. These are among Sorace's (1996) arguments in favor of various mechanisms for rank ordering of L2 stimuli by learners. See Chaudron's (1985a) application of paired comparisons to ranking of the comprehensibility of a range of topic reinstatement devices, as well as Mohan (1977) and Walters (1979). This method can be expanded to include more than just pairs of items. The principle of ranking or rating a group of items with respect to some criterion was used, for instance, by Carrell and Konneker (1981) for judgments of politeness, White (1989) for a correctness decision/ranking of four adjacency condition sentences, and Cowan and Hatasa (1994) for sets of four similar grammatical Japanese relative clause sentences to be ranked with respect to their difficulty (targeting complexity of relatives).

Several other procedures have made use of learners' metalinguistic judgments. An interesting one, which has not been used (to my knowledge) since its early applications by Kellerman (1978) and Carrell and Konneker (1981), is *card sorting* (or "Q-sort"; see Miller, 1969), or a type of scaling/grouping according to

judgments of same/different. In order to determine learners' perception of the literal or figurative meanings of English L2 "break," Kellerman had subjects place sentences using "break" in a number of its meanings into distinct groups or piles. Just as with paired comparisons, accumulation of the responses of a number of subjects allows for a pooled rank order of preferences and clusters in perceptions about meaning (in this case – see research on cluster analysis, such as Skehan, 1986). A second method is *magnitude scaling* or *estimation*, a technique derived from psychophysical research traditions, which allows subjects to set their own standard or scale for comparing judgments (see Bard, Robertson, and Sorace, 1996, on L1 acceptability judgments evaluated using magnitude scaling, as well as Sorace, 1990, 1996). Yuan (1995, 1997), for instance, used magnitude estimation in eliciting subjects' "acceptability" judgments on Chinese topic structure by English-speaking learners of Chinese L2, and on English (ungrammatical and grammatical) ± subject or object constructions by Chinese-speaking learners of L2 English. In this procedure, subjects are presented sentences, as in a GJT, but they are asked to create their own value for the degree of acceptability of the first sentence. Then, for each successive sentence, they provide a value relative to the first that represents their judgment. Subjects' scores must then be standardized in order to carry out further analysis, but this procedure is deemed to provide a more sensitive within-subject (and comparative) measure of learners' perceptions and intuitions. Recently, Gass et al. (1999) used magnitude estimation for Spanish native-speaker raters' evaluations of L2 Spanish learners' film-retelling narratives, and Dube (2000) provides the instructions for a magnitude estimation study of Zulu L2 acquisition by learners of a large range of proficiency levels, as measured by a (apparently general) cloze test.

A third method, and arguably the one that accesses the highest level of metalinguistic knowledge, consists of procedures that elicit learners' expression of grammatical or other rule knowledge, vocabulary definitions (e.g., Snow, Cancino, de Temple, and Schley, 1991), or other verbalized intuitions about (e.g., pragmatic) acceptability (see also Ellis's 1991 analysis of learners' think-aloud reports about the strategies and deliberations they went through when making grammaticality judgments). An early study that compared such rule expression with other tasks (elicited imitation, dictation, and grammar correction) was that of Grigg (1986), who asked ESL learners to write out the rules for the phonology of plural *-s*, articles, and relative pronouns ("who," "which," "that"). He evaluated the adequacy of their responses according to a fixed target description, using a 10-point scale. A larger-scale study by P. Green and Hecht (1992) had young German learners of English express the rules for certain grammatical errors of morphology and syntax, while correcting them, as well. The students evaluated the relationship between rule knowledge and ability, in comparison with native English-speaking school children. Green and Hecht provide a description of their scheme for counting various types of rule descriptions, but they do not discriminate among them in any scaled way, with only a binary "correct/incorrect" score calculated.

One of several dependent measures used by Alanen (1995) was rule descriptions for two types of rules in L2 Finnish, as she compared four experimental groups receiving various sorts of input to process. Two of her groups were also given the Finnish rules, so that she could immediately determine the extent of learning of the rules. Her simultaneous elicitation of grammaticality judgments with explanations and think-aloud protocols allowed her to triangulate subjects' extent of awareness of rules with their performance. A similar comparison of rule description and think-aloud protocol analysis allowed Rosa and O'Neill (1999) to assess the degree of noticing or awareness of L2 Spanish learners related to their learning achievement.

5   Reliability and Validity

5.1   Reliability

Although researchers should ultimately be concerned about the validity of their data and conclusions, the reliability of the data-collection procedure or instrument needs to be determined first. In most of the studies involving the measures discussed above, the researchers failed to establish one or more of the following: inter-rater reliability, "test–retest" reliability, or internal consistency reliability; or to apply other such measures traditionally used in the domain of psychometrics (for standard psychometric measures, see, e.g., Bachman, 1990, ch. 6; also see Chaudron, Crookes, and Long, 1988, regarding observational and linguistic coding judgments; and see Norris and Ortega, this volume). In the case of naturalistic observations, for example, most researchers state that their transcriptions were verified by at least one other researcher, but as is typical in linguistic analysis of this sort, quantitative (inter-rater) reliability measures have generally not been reported, and regrettably, there is little mention in these studies of concerns over the issue. Similarly, and somewhat surprisingly, data collected even in most of the more experimental studies have not frequently been assessed for internal consistency or other measures of reliability. Most researchers using forms of elicited production appear to take it as given that the linguistically focused analyses they conduct on the data elicited by their instruments are inherently reliable, and that the researcher, or a pair or team of researchers in consultation with one another, is or are competent to judge reliability without submitting their analyses to objective or independent comparison. It is true, in fact, that many of the results obtained using such instruments involve relatively low-inference structures. So concern for validity and reliability tends only to appear in L2 research using more experimental tasks, in particular procedures such as standardized tests. In contrast, as perusal of virtually any journal will demonstrate, researchers in child L1 acquisition, and especially in language impairment studies, adopt more rigorous measures of reliability and validity for many linguistically oriented analyses (e.g., Bucks, Singh, Cuerden, and Wilcock, 2000; Damico, Oller, and Tetnowski, 1999; Fenson et al., 2000).

5.2   *Validity*

Assuming reliable measurement, learner data elicited by any of the above procedures may be valid information on SLA, depending on the degree of conformity and consistency of such data with one or more of the following (see Norris and Ortega, this volume, for specific discussion of forms of validation in SLA research):

i   Theoretical proposals (*prediction*): researchers investigate underlying linguistic and developmental phenomena in search of differences in performance among learners, so that outcomes that confirm their predictions tend to "validate" their theories, but at the same time confirm that the measures used are effective, and thus "valid." This occurs whether the predictions arise out of comparative linguistic analyses, or assumptions about differences among learners based on proficiency levels, learning experiments, or development over time via maturation. (See, however, the cautions expressed by Thomas, 1994, concerning the lack of use of proficiency measures in L2 research.)
ii   Comparable data from other studies of a similar nature (*replication*, in the case of intentional retesting of a prior finding, or *convergence*, if two researchers happen to have carried out comparable studies): to the extent that outcomes on the same or related measures with respect to similar target features point consistently to the same fundamental conclusion, those measures may be judged valid.
iii   Simultaneous measures within a study using other techniques (*triangulation*): only slight differences in outcomes on different tasks can frequently accentuate the consistency of predictions of learner production; on the other hand, if specific dimensions on which the techniques are expected to differ result in distinct outcomes, the limits of generalizability of the methods can be determined (a form of validation).

5.3   *Theoretical predictions*

As an example of comparability of data across naturalistic studies (not strictly *replication*, as the researchers have operated independently of one another), Meisel (1994a), Paradis and Genesee (1996), Schlyter (1993), and Swain and Wesche (1973) are all concerned with the appearance of comparable structural constraints on the acquisition of the children's two languages. On the whole, they arrive at similar findings with respect to a natural sequence of acquisition according to L1 norms of at least the stronger of the children's languages, but there is evidence of greater variability, non-L1 type errors, and failure to attain structural differentiation in the weaker of the children's languages, even when linguistic developmental level is controlled (mean length of utterance in Schlyter, 1993). Similarly, Meisel (1994a) and Swain and Wesche (1973) both find language mixing (especially of lexical items) occurring early in development,

but code switching begins to be constrained by structural factors once functional categories such as agreement and tense appear. Also, apart from some degree of early use of one language's lexis within the other's syntax (as also in Swain and Wesche, 1973), Paradis and Genesee's (1996) data illustrate that neither of the two simultaneously acquired (and more equally balanced) languages has either a strong negative or a facilitating influence on the syntactic development of the other, as each one follows an L1 developmental sequence. Likewise, Polio (1995) and Jin (1994), independently investigating, among other targets, evidence for zero pronoun production by L2 Chinese learners doing story-retelling, found no evidence in speech by beginning learners, regardless of whether the L1 of the subjects was a subject- or a topic-prominent language (Polio's subjects were both Japanese and English speakers). Finally, as already noted in the section on elicited production, the feature of +planning, as an aspect of communicative task design, had proved to result in consistent findings using various tasks and measures across studies.

Many studies involve theoretical predictions of L2-influenced performance as shown by comparison with data from native speakers. For example, in their on-line sentence-reading task, Juffs and Harrington (1995, 1996) found L2 GJT error rates for *wh*-extraction (18–40 percent) higher than L1 rates (especially problematic were *wh*-extraction of subject from infinitives); however, error rates were lower and closer to NS performance on garden path sentences, by which NSs also were predicted to be misled. There was a corresponding variability in response time to judgments of grammaticality relative to NS performance. Also, in reading time per word on the garden path sentences, the NNSs were slower at comparable critical points, although in a more exaggerated pattern. Likewise, among the numerous measures employed by Ioup et al. (1994) in order to distinguish their advanced NNSs of Arabic from NSs, while most measures did not clearly discriminate between them, an anaphora-interpretation task involving discourse semantics clearly distinguished Julie, their advanced subject, from NSs. Of interest here was that this was one of the few tasks which showed a weaker performance for Julie than for another advanced L2 speaker who was an instructed learner of Arabic, and who may thus have benefited from instruction in learning the relevant structures.

Very frequently, researchers who investigate the effects of instruction on acquisition will predict that a particular instructional treatment will lead to changes in L2 performance (cf. Norris and Ortega's, 2000, review of the overall effects of instruction). Although design problems may lead to questionable results, and some instruction fails to achieve an effect, numerous tasks have been successful in detecting the effects of instruction. Many of the studies by White and her colleagues (e.g., Trahey and White, 1993; White, 1991; White, Spada, Lightbown, and Ranta, 1991), have measured the effects of theoretically motivated instruction with sentence-comparison preference tasks, picture descriptions or responses, judgment tasks, and others. In most of these, clear effects of instruction (on question formation and adverb order) have been revealed. For instance, Trahey and White (1993) showed subjects' sensitivity to

training or input on the target form (adverb order in L2 English), with a differential effect between the training and control groups on most measures, according to predictions of the value of negative input for French L1 learners to unlearn incorrect SAV order. Of interest was their finding that the preference task was sensitive to the learning effects of explicit training on English adverb order, while it was not as sensitive in distinguishing between their control training groups (question formation and input flood). Trahey (1996) further shows that these results endure over a full year's time. Other training studies, such as Rosa and O'Neill's (1999) comparison of formal instruction and rule-search procedures in the acquisition of Spanish L2 past hypotheticals, and Alanen's (1995) study of explicit rule presentation and input enhancement with Finnish L2 locative morphology, found positive effects on a variety of measures: multiple-choice recognition selection of the correct verb form in the former, and rule statements and sentence completion in the latter.[10] And DeKeyser (1997) showed that comprehension and production training in an artificial language, as measured by multiple-choice and metalinguistic tests of the same modality format, resulted in linear improvement in reaction time, but that "students with comprehension practice only . . . improve more in comprehension skills than students with production practice only and vice versa, whereas students with an equal amount of practice in both skills . . . perform at an intermediate level in both skills" (p. 213).

Another type of prediction of effects on tasks is that they would be sensitive to basic differences in subjects' level of proficiency (as determined by external factors, such as amount of prior instruction or natural developmental growth – correlations between proficiency measures and tasks will be addressed in a later section). For example, Salaberry and López-Ortega (1998) compared two (placement) levels of L2 Spanish learners' abilities on several measures of article and subject pronoun use, and aspect: multiple-choice sentence completion, open-ended fill-in-the-blank, and a written narration based on a picture stimulus. They report all measures as being sensitive to the increased level of proficiency, with the fill-in-the-blank task, being the more difficult task, showing the greatest sensitivity. Similarly, among a number of elicited production and experimental tasks employed by Ortega (2000) to assess theoretically expected developmental changes in complexity and accuracy in the course of a semester of advanced L2 Spanish – student journal writing, academic paper writing, written and oral picture narratives, as well as GJT and elicited imitation – the oral narratives showed consistent change in complexity, while the GJT and EI showed significant improvement in accuracy.[11]

5.4   Replication

It will be recalled that there were convergent findings by both Rintell and Mitchell (1989) and M. Sasaki (1998) in which role plays were superior to DCTs in their elicitation of more complex speech act behavior, and a finding also by Kormos (1999) with respect to learners' greater conversational competence

in a guided role play compared to an interviewer conducting a guided interview (a general finding against OPI-type procedures). Such findings justify other researchers' efforts to employ role plays as a more sensitive measure of speech acts. Another method that has been found to produce similar results across tasks is EI, or sentence repetition, in particular when the stimulus is repeated or there is a delay prior to the prompt to repeat. Among his several measures, Grigg (1986) employed two different versions of the EI task, one with the stimuli repeated twice. These intercorrelated highly ($r = .76$), and both correlated with a dictation task (.68), but while the EI task with one repetition of the stimulus showed little relation to his grammar rule knowledge task ($r = .19$), the EI task with a repetition, and thus more adequate time for responding, correlated significantly with the rule knowledge task ($r = .51$, p < .05). Just as in the L1 study by McDade, Simpson, and Lamb (1982), which assessed the effects of a delay before repeating, the implication is that, with time to process the stimuli, L2 subjects will best be able to call upon their rule-based competence in production. A different experimental adjustment with much the same conclusion was used by Yang and Givón (1997), who injected a distracter task for 15 seconds, so that subjects who were better able to process the grammatical and lexical information in the stimuli proved more successful in eventual repetition: the delay resulted in a better discrimination of ability between their training groups than the immediate repetition. The authors caution, however, that "these findings suggest that elicited sentence recall may not be a valid measure of grammatical competence for aspects of grammar that are subject to surface information loss" (p. 186), such as morphological or phonological information.

5.5   *Triangulation*

The final major means by which researchers can validate the findings of their elicitation measures is by confirming that their results are consistent across different intra-study and intra-subject tasks. Yang and Givón (1997), for instance, who used a large number of measures to attempt to discriminate between their full-instruction vs. simplified ("pidgin") input of an artificial language, found most measures discriminating consistently between the treatments, although vocabulary measures did so less, since the input in the two treatments was virtually the same. Likewise, almost all of Alanen's various measures tended to discriminate in favor of the rule-based over the non-rule-based treatments.

   Among naturalistic examples of intra-subject, cross-task validation are Schmidt and Frota's (1986) comparison of data from Schmidt's self-observations with those elicited in their conversations, and Tomiyama's (2000) use of other elicitation measures for the purpose of triangulation. In the former, discrepancies between observations of his awareness and his productions led, in particular, to Schmidt's theorizing on the importance of conscious awareness and attention in learning (cf. Schmidt, 1995b). In Tomiyama's study, there was concurrence

in the findings based on objective measures of differential rates of attrition in syntax and lexis and those based on the subject's spontaneous speech. In the same way, Rosa and O'Neill (1999) compared groups who explicitly mentioned rules for L2 Spanish contrary-to-fact conditionals, and were able to formulate them, with those who did not, and the rule formulators were significantly better in performance.

Very many comparisons between different tasks demonstrate that some may be more sensitive relative to other criterion performances, and these differences are important in order for future researchers to exploit particular differences in elicitation procedures. For example, Helms-Park (2001) used picture production and multiple-choice picture selection (comprehension) tasks to investigate English L2 causative acquisition by Vietnamese and Hindi-Urdu learners. She found the subjects to perform similarly on production, but they were distinguished according to L1 transfer predictions on the comprehension measure, with some effects revealed by level of proficiency. Similarly, Flanigan's (1995) study of relative clause structure showed weaker success on production than comprehension, and the more difficult production measure revealed a predicted order of OS-OO-SS-SO, while the easier comprehension task did not. Flanigan also noted a significant correlation between scores on the Bilingual Syntax Measure and her question-answering task.

5.6   *Grammaticality judgments and validity*

Perhaps the most widely used measure in SLA research has been the GJT, which we have already noted in comparison with other measures and predicted outcomes. Owing to its experimental and quantifiable nature, as well as the widespread application of GJT for diverse target structures, many researchers have shown concern for the absolute (i.e., theoretically distinct and predicted) and concurrent validity of this method, so we have saved a discussion of this method until the end of this section. That GJTs do exhibit concurrent (triangulated) validity has been demonstrated not only in several of the studies cited above, but, for example, in consistency between it and a (SOPI-like) Spanish Speaking Test ($r = .87$; Ortega, 2000), on-task performance in dative movement training (S. Carroll and Swain, 1993), introspective evidence (R. Nation and McLaughlin, 1986), a paragraph story-task (White et al., 1997), "Dehydrated Sentence" reconstruction of several Spanish L2 structures (Mandell, 1999), and sentence-matching examining L2 Spanish clitic placement (Duffield and White, 1999). Also, as a matter of "predicted" performance, comparing GJTs presented with audio and written input, D. J. Johnson (1992; see also Slavoff and Johnson, 1995) found that age of onset of L2 exposure, thus earlier initiation of target language literacy, tended to favor more accurate performance on the written version.

However, GJTs have not always resulted in findings consistent with other measures (such as the comparison of picture-cued description using English hypothetical past and GJTs on the same targets, in Izumi, Bigelow, Fujiwara,

and Fearnow, 1999; see early comments on discrepant results in Ellis, 1991; Gass, 1983). It is clear from much discussion on this procedure that a number of factors must be considered in interpreting results from its use.[12]

In an early review of this procedure, Chaudron (1983) found generally that there was a correspondence between GJT findings and other measures. However, Gass (1983) conducted a study on a set of English L2 learners in order to determine their ability to correct their own and peers' productions. On the basis of high variability in performance, which included inconsistent and erroneous judgments, Gass argued that learners do not progress in a natural way to improve their judgments, but instead rely on an increasing degree of analytical knowledge in order to evaluate and correct errorful target structures. Thus, if task conditions do not promote access to such knowledge, the measure may not succeed in eliciting performance comparable to subjects' other productive capacities. The question of absolute validity arises, then, of what the relationship is between performance on a GJT and underlying competence. The problem of variability in learner performance, not only in general interlanguage (IL) production, but also in access to metalinguistic knowledge, is therefore an important one for the study of the GJT as a valid measure. Sorace (1996) refers to this as "indeterminacy":

> First, native judgments themselves can be indeterminate, particularly when the objects of investigation are highly marked or very subtle syntactic properties . . . At the most basic level, [target language] constructions are indeterminate because the learners do not have any knowledge of them . . . [IL indeterminacy due to ignorance] characterizes nonnative grammars throughout the acquisition process, although it is more conspicuous at the initial and intermediate stages of IL development. At more advanced stages, constructions may become indeterminate (after a period of relative stability) because of the increased amount and sophistication of the learner's knowledge.   (pp. 385–6)

Sorace then suggests "that the UG-driven specification of [SLA] core properties is narrower in scope and strength than in native grammars . . . The result is a wider periphery and consequently more room for permeability and variation" (p. 387).[13] Sorace takes this point to the extreme end of comparison between NS judgments and those of near-natives (as in Birdsong, 1992; Coppetiers, 1987; Davies and Kaplan, 1998; Ioup et al., 1994), where she notes: "near-native grammars may also be indeterminate in the same sense as native grammars . . . [which] leads to inconsistent and variable judgments" (p. 390). She uses this observation eventually to argue in favor of various mechanisms for rank ordering of L2 stimuli.

At the same time, other researchers have noted that the basis for non-native judgments of grammaticality (or acceptability) lies not strictly in linguistic knowledge, but in varying degrees of application of other sources of intuitions and cognitive operations (Birdsong, 1989; Davies and Kaplan, 1998; Ellis, 1991; Gass, 1989; Goss, Zhang, and Lantolf, 1994; Schütze, 1996; and see debate between Birdsong, 1994; and Felix and Zobl, 1994). Cowan and Hatasa (1994),

for example, comment on the concurrent validity of the GJT with a scalar judgment task on a criterion measure:

> Our results indicate that no matter how delicate the scale, native speaker and L2-learner judgment data will, at best, reflect sensitivity to only some structural characteristics that affect processing, and that they will always vary with on-line data, which are far more indicative of complexity . . . processing research must employ some on-line task that elicits reaction or reading time plus some measure that provides an indication of the extent to which the stimuli used in the on-line task were comprehended.   (p. 297)

Gass (1994), however, who scrutinizes the variability in reliability of judgments by learners on the Noun Phrase Accessibility Hierarchy, is more cautious: "there is evidence to suggest that low reliability occurs in just those areas where greater indeterminacy is predicted . . . [I]t has been shown that judgment data can, when used properly and appropriately, be useful in second language acquisition research" (p. 320). Her evidence showed that NNS variability in automaticity of L2 perception and processing interferes with access to L2 knowledge representation. Many other studies have found similar results; for example, Ortega (2000) found that with familiarity with the target language, L2 Spanish learners' certainty in judgments and their eventual consistency increased; and in Juffs and Harrington (1996) those with longer reading times were more accurate,

6   Adaptability and Innovation of Research Methods between L1 and L2 Research

We are now left with the last two questions posed at the beginning of this chapter:

iv   To what extent can procedures adopted for L1 research be applied to L2 research?
 v   What new or additional procedures or adaptations are possible or needed for L2 research?

To the first of which it now seems appropriate to add the reverse question:

vi   What L2 procedures can be applied to L1 research?

As for the first question, it would seem as if, since most of the procedures employed in L2 research have indeed arisen in some form from prior L1 research, we have only to look to L1 research for guidance as to the next stages. To some extent this is true. Among technologically sophisticated techniques, L2 research has clearly recognized and taken advantage of computerized, on-line

experiments and data collection (e.g., Beck, 1998; Hagen, 1994; Hulstijn and DeKeyser, 1997; Juffs, 1998; Juffs and Harrington, 1995, 1996; Robinson, 1997). These studies all used reading time or reaction time to judgments. Also, the use of event-related potentials and magnetic resonance imaging (Weber-Fox and Neville, 1999) is beginning to find a place in SLA research. What remains to be examined with such procedures, of course, is how much the data derived from them are subject to greater within- and between-subject error variability, owing to the higher degree of indeterminacy and alternative sources of know-ledge in L2 learners' grammars, as was pointed out about GJTs. Hulstijn (1997) raises just such a cautious note. Yet Kempe and MacWhinney (1998) illustrate in some detail how decision latencies (reaction time) can be analyzed fruitfully to reveal predictable tendencies in acquisition (comparing case-marking process-ing between learners of L2 German and Russian in a competition-model word-order experiment). A wider application of such procedures is warranted.

Despite the evidence of greater overall variability in L2 than L1 learners' grammars and performance, it is difficult to imagine that there could not be an application for L1-associated research methodology and data-elicitation proce-dures of the most sophisticated sort, if only because L2 learners are human, as well, and they possess all the potential abilities of any native speaker. Certainly, all manner of direct naturalistic observation and standard elicited production measures is already well within the repertoire of SLA researchers. But a number of on-line procedures used in the psycholinguistic literature rely on full and highly automatized access by subjects to their mature grammars and perceptual processing, so that subjects with too high variability in reaction time and accuracy are more likely to be excluded from L1 research, whereas they are more of the natural population in SLA studies.[14] Thus, L2 learners may exhibit too much variability and uncertainty for some of these procedures to be applied. But beyond this not insurmountable problem, whatever limitations there may be lie more in the question of access to facilities and financial resources, and possibly the incidental problems of accessing and putting through the rigors of laboratory training L2 learners who may at times be more reluctant than the captive audience of first-year psychology students.

As for the final question, concerning which L2 procedures can be applied in L1 research, it is fair to say that it has been SLA research, rather than L1 research, which has developed most thoroughly a number of methods and measures of language performance and use. Examples of these include: (i) obser-vational measures of classroom language use; (ii) description of pragmatic language abilities in social interaction; (iii) detailed methods for tapping into learners' introspective, metalinguistic knowledge; (iv) elicited production tech-niques and variables in design of picture sequences and tasks (as indicated in section 3.3); and (v) many other procedures for which the focus is the learners' manipulation of linguistic form versus meaning. Some reasons for this greater development in SLA research are that researchers have had the benefit of subjects with greater maturity than infants and young children, who could employ world knowledge in their operations with tasks. Also, the greater concern in SLA

studies with developing grammars during later stages of cognitive maturation has motivated more elaborate designs, methodologies, and tests, in order to examine complex linguistic performance. Therefore, it will be incumbent on L1 researchers to have a closer look at SLA research in order to discover some of the implications and expansions of their own methods and interpretations in the realm of language acquisition and processing, which they ought to perceive as the natural testing ground for claims about the uniformity of human linguistic experience.

NOTES

1   For reasons of length, this review focuses on L2 oral production; the quite extensive research on L2 writing or reading, for example, cannot be included. See Wolfe-Quintero, Inagaki, and Kim (1998) for an extensive review of analytical procedures in the writing domain.

2   Throughout this review, in one effort to avoid stylistic boredom, the terms "procedure," "task," "method," and "technique" will be used interchangeably to refer to types of data collection, although at times each of these may refer to a more specific or distinct referent (as when we refer to "communicative" tasks as one type of technique).

3   See other chapters in this volume, especially Norris and Ortega, regarding approaches to SLA and measurement.

4   As space is limited, and as we are unaware of a primary source of research about personal narratives and their elicitation in second language research, we can only point to the vast literature on the elicitation and analysis of narrative in L1 research. See the extensive, partly annotated bibliography by Handorf, Watson-Gegeo, and Sato (1993).

5   Reaction time and ERP are among the psycholinguistic processing measures commonly employed in L1 research, but they have been used very little in L2 research, a point to be discussed in the concluding section of this review.

6   Cloze tests as employed in SLA research are highly varied, and can be constructed with many distinct targets. A basic distinction is between "random" deletion of cloze items for more global assessment and "rational" deletion for specific target form assessment. It is beyond the scope of this review to address these variations.

7   It is not uncommon for researchers to include as distracters other sets of items that are to be used for a different investigation, but the nature of which is unmentioned in the study. The issue of fatigue and other task-internal effects (e.g., perseverance errors) on the part of the subject in cases of large sets of items is rarely discussed.

8   See discussion in Schütze (1996, pp. 62–77) on L1 theorists' concerns about the dichotomous or scalar nature of grammaticality.

9   Correction is often used to confirm that subjects' grounds for rejection are justified. It has been used in the above studies, as well as E. Klein (1995), Salaberry (1998), Schachter

(1989), and Trahey and White (1993).

10 Surprisingly, Alanen found that on the GJT for locatives, the Rule group was best and the Rule+Enhance condition was the weakest. Her analysis suggested that the latter group were systematically rejecting one of the correct alternative forms.

11 Ortega also reports very high Cronbach's alpha reliability of the GJT and EI, at both the beginning and end of the semester – all but the GJT at the beginning, which was .86, were .95–6. The lack of a control group leaves open the possibility that improvement in these accuracy measures could be due to test–retest experience, although that is an unlikely explanation, given the 14 weeks that intervened.

12 Note that the sociolinguist Milroy (1987, ch. 7, p. 146 ff), discussing "elicitation techniques" for L1 data collection to supplement naturally occurring data, refers principally to acceptability ratings, but questions the reliability of linguistically naive informants' judgments.

13 This notion fits with Zobl's (1992) proposal of two main sources of metalinguistic (grammatical) knowledge – from "input-independent" (e.g., core UG) and "input-dependent" knowledge (language-particular parameters and elements outside core grammar) – which lead to more homogeneity for judgments based on core principles, but more heterogeneity for decisions about input-dependent, L1-specific structures.

14 Except perhaps in research literature on the language-impaired population, for which, see research by, for example, Gathercole and Baddeley (1993) for examples of L2-appropriate methods and research questions.

REFERENCES

Adams, M. 1978: Methodology for examining second language acquisition. In E. Hatch (ed.), *Second Language Acquisition: A Book of Readings.* Rowley, MA: Newbury House, 278–96.

Ahrenholz, B. 2000: Modality and referential movement in instructional discourse: comparing the production of Italian learners of German with native German and native Italian production. *Studies in Second Language Acquisition*, 22 (3), 337–68.

Alanen, R. 1995: Input enhancement and rule presentation in second language acquisition. In R. Schmidt (ed.), *Attention and Awareness in Foreign Language Learning*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center, 259–302.

Allwright, D. 1988: *Observation in the Language Classroom*. London: Longman.

Allwright, D. and Bailey, K. M. 1991: *Focus on the Language Classroom: An Introduction to Classroom Research for Language Teachers*. New York: Cambridge University Press.

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

Bailey, K. M. 1983: Competitiveness and anxiety in adult second language learning: looking *at* and *through* the diary studies. In H. W. Seliger and M. H. Long (eds), *Classroom Oriented*

Research in Second Language Acquisition. Rowley, MA: Newbury House, 67–103.

Bailey, K. M. 2001: Action research, teacher research, and classroom research in language teaching. In M. Celce-Murcia (ed.), *Teaching English as a Second or Foreign Language*. 3rd edition. Boston, MA: Heinle and Heinle, 489–98.

Bailey, K. M. and Nunan, D. (eds) 1996: *Voices from the Language Classroom*. New York: Cambridge University Press.

Bard, E. G., Robertson, D., and Sorace, A. 1996: Magnitude estimation of linguistic acceptability. *Language*, 72 (1), 32–68.

Bardovi-Harlig, K. 1999: From morpheme studies to temporal semantics: tense-aspect research in SLA. *Studies in Second Language Acquisition*, 21 (3), 241–82.

Bardovi-Harlig, K. 2000: Tense and aspect in second language acquisition: form, meaning, and use. *Language Learning 50: Supplement 1*. Ann Arbor: Blackwell.

Bardovi-Harlig, K. and Hartford, B. S. 1993: Learning the rules of academic talk: A longitudinal study of pragmatic change. *Studies in Second Language Acquisition*, 15 (3), 279–304.

Bardovi-Harlig, K. and Hartford, B. S. 1996: Input in an institutional setting. *Studies in Second Language Acquisition*, 18 (2), 171–88.

Bayley, R. 1994: Interlanguage variation and the quantitative paradigm: past tense marking in Chinese-English. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 157–81.

Beck, M.-L. 1998: L2 acquisition and obligatory head movement: English-speaking learners of German and the Local Impairment Hypothesis. *Studies*

in Second Language Acquisition, 20 (3), 311–48.

Bennett-Kastor, T. 1988: *Analyzing Children's Language: Methods and Theories*. Oxford: Blackwell.

Berko, J. 1958: The child's learning of English morphology. *Word*, 14, 150–77.

Bernstein Ratner, N. 2000: Elicited imitation and other methods for the analysis of trade-offs between speech and language skills in children. In L. Menn and N. Bernstein Ratner (eds), *Methods for Studying Language Production*. Mahwah, NJ: Lawrence Erlbaum Associates, 291–311.

Bialystok, E. 1990: *Communication Strategies: A Psychological Analysis of Second-Language Use*. Oxford: Blackwell.

Bialystok, E. and Ryan, E. B. 1985: A metacognitive framework for the development of first and second language skills. In D. Forrest-Presley, G. E. MacKinnon, and T. G. Waller (eds), *Metacognition, Cognition, and Human Performance. Vol. 1: Theoretical Perspectives*. Orlando: Academic Press, 207–52.

Birdsong, D. 1989: *Metalinguistic Performance and Interlinguistic Competence*. Berlin: Springer-Verlag.

Birdsong, D. 1992: Ultimate attainment in second language acquisition. *Language*, 68, 706–55.

Birdsong, D. 1994: Asymmetrical knowledge of ungrammaticality in SLA theory. *Studies in Second Language Acquisition*, 16 (4), 463–73.

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

Bley-Vroman, R. and Masterson, D. 1989: Reaction time as a supplement to grammaticality judgments in the investigation of second language

learners' competence. *University of Hawai'i Working Papers in ESL*, 8 (2), 207–45.

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

Blum-Kulka, S., House, J., and Kasper, G. (eds) 1989: *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.

Bolander, M. 1989: Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. In K. Hyltenstam and L. K. Obler (eds), *Bilingualism Across the Lifespan*. Cambridge: Cambridge University Press, 73–86.

Bongaerts, T. 1999: Ultimate attainment in L2 pronunciation: the case of very advanced late L2 learners. In D. Birdsong (ed.), *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum Associates, 133–59.

Bongaerts, T., van Summeren, C., Planken, B., and Schils, E. 1997: Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, 19 (4), 447–65.

De Bot, K., Cox, A., Ralston, S., Schaufeli, A., and Weltens, B. 1995: Lexical processing in bilinguals. *Second Language Research*, 11, 1–19.

Brown, C., Sagers, S. L., and LaPorte, C. 1999: Incidental vocabulary acquisition from oral and written dialogue journals. *Studies in Second Language Acquisition*, 21 (2), 259–83.

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. 2000: Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14 (1), 71–91.

Carrell, P. L. and Konneker, B. H. 1981: Politeness: comparing native and nonnative judgments. *Language Learning*, 31 (1), 17–30.

Carroll, M., Murcia-Serra, J., Watorek, M., and Bendiscioli, A. 2000: The relevance of information organization to second language acquisition studies. *Studies in Second Language Acquisition*, 22 (3), 441–66.

Carroll, S. and Swain, M. 1993: Explicit and implicit negative feedback: an empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15 (3), 357–86.

Cebrian, J. 2000: Transferability and productivity of L1 rules in Catalan–English interlanguage. *Studies in Second Language Acquisition*, 22 (1), 1–26.

Chafe, W. (ed.) 1980: *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Chaudron, C. 1983: Research on metalinguistic judgments: a review of theory, methods, and results. *Language Learning*, 33 (3), 343–77.

Chaudron, C. 1985a: The role of simplified input in classroom language. In G. Kasper (ed.), *Learning, Teaching and Communication in the Foreign Language Classroom*. Århus: Århus University Press, 99–110.

Chaudron, C. 1985b: Intake: on models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition*, 7 (1), 1–14.

Chaudron, C. 1988: *Second Language Classrooms: Research on Teaching and Learning*. Cambridge: Cambridge University Press.

Chaudron, C. 2001: Progress in language classroom research: evidence from *The Modern Language Journal*, 1916–2000. *Modern Language Journal*, 85 (1), 57–76.

Chaudron, C. forthcoming: *Elicited Imitation: Second Language and First Language Applications*.

Chaudron, C., Crookes, G., and Long, M. H. 1988: *Reliability and Validity in Second Language Classroom Research*. Technical Report No. 8, Center for Second Language Research. Honolulu: University of Hawai'i.

Chomsky, C. 1969: *The Acquisition of Syntax in Children from 5 to 10*. Cambridge, MA: MIT Press.

Clark, J. L. D. and Clifford, R. T. 1988: The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research. *Studies in Second Language Acquisition*, 10 (2), 121–47.

Cohen, A. D. 1991: Feedback on writing: the use of verbal report. *Studies in Second Language Acquisition*, 13 (2), 133–59.

Cohen, A. D. and Hosenfeld, C. 1981: Some uses of mentalistic data in second language research. *Language Learning*, 31 (2), 285–313.

Cohen, A. D. and Olshtain, E. 1994: Researching the production of second-language speech acts. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates,143–56.

Cook, V. 1986: The basis for an experimental approach to second language learning. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 3–21.

Coppetiers, R. 1987: Competence differences between native and near-native speakers. *Language*, 63 (3), 545–73.

Cowan, R. and Hatasa, Y. A. 1994: Investigating the validity and reliability of native speaker and second-language learner judgments about sentences. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 287–302.

Crookes, G. 1989: Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11 (4), 367–83.

Crookes, G. 1991: Second language speech production research. *Studies in Second Language Acquisition*, 13 (2) 113–32.

Damico, J. S., Oller, J. W., Jr, and Tetnowski, J. A. 1999: Investigating the interobserver reliability of a direct observational language assessment technique. *Advances in Speech-Language Pathology*, 1 (2), 77–94.

d'Anglejan, A. and Tucker, G. R. 1975: The acquisition of complex English structures by adult learners. *Language Learning*, 25 (2), 281–96.

Davies, W. D. and Kaplan, T. I. 1998: Native speaker vs. L2 learner grammaticality judgements. *Applied Linguistics*, 19 (2), 183–203.

Dechert, H. W. and Sandrock, U. 1986: Thinking-aloud protocols: the decomposition of language processing. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 111–26.

DeKeyser, R. M. 1997: Beyond explicit rule learning: automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19 (2), 190–221.

DeKeyser, R. M. and Sokalski, K. J. 1996: The differential role of comprehension and production practice. *Language Learning*, 46 (4), 613–42.

Demuth, K. 1996: Collecting spontaneous production data. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 3–22.

Dietrich, R., Klein, W., and Noyau, C. 1995: *The Acquisition of Temporality in a Second Language*. Philadelphia: John Benjamins.

Dittmar, N. (ed.) 1992: Grammaticalization in second language acquisition. Special issue.

Studies in Second Language Acquisition, 14 (3).

Doughty, C. 1991: Second language instruction does make a difference: evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition*, 13 (4), 431–69.

Doughty, C. and Long, M. H. 2000: Eliciting second language speech data. In L. Menn and N. Bernstein Ratner (eds), *Methods for Studying Language Production*. Mahwah, NJ: Lawrence Erlbaum Associates, 149–77.

Dube, B. 2000: Where are the minimal trees? Evidence from early Zulu L2 subordination. *Second Language Research*, 16 (3), 233–65.

Duff, P. A. 1993: Syntax, semantics, and SLA: the convergence of possessive and existential constructions. *Studies in Second Language Acquisition*, 15 (1), 1–34.

Duffield, N. and White, L. 1999: Assessing L2 knowledge of Spanish clitic placement: converging methodologies. *Second Language Research*, 15 (2), 133–60.

Eckman, F. R. 1994: Local and long-distance anaphora in second-language acquisition. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 207–25.

Eckman, F. R., Bell, L. H., and Nelson, D. 1988: On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics*, 9 (1), 1–20.

Edwards, A. L. 1957: *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts.

Ellis, R. 1989: Are classroom and naturalistic acquisition the same? A study of the classroom acquisition of German word order rules. *Studies in Second Language Acquisition*, 11, 305–28.

Ellis, R. 1991: Grammaticality judgments and second language acquisition.

Studies in Second Language Acquisition, 13 (2), 161–86.

Ellis, R. 1992: Learning to communicate in the classroom: a study of two language learners' requests. *Studies in Second Language Acquisition*, 14 (1), 1–23.

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

Ellis, R. and He, X. 1999: The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, 21 (2), 285–301.

Ericsson, K. A. and Simon, H. A. 1980: Verbal reports as data. *Psychological Review*, 87, 215–51.

Ericsson, K. A. and Simon, H. A. 1984: *Protocol Analysis*. Cambridge, MA: MIT Press.

Eubank, L. 1993: Sentence matching and processing in L2 development. *Second Language Research*, 9 (3), 253–80.

Færch, C. and Kasper, G. (eds) 1983: *Strategies in Interlanguage Communication*. London: Longman.

Færch, C. and Kasper, G. (eds) 1987a: *Introspection in Second Language Research*. Clevedon: Multilingual Matters.

Færch, C. and Kasper, G. 1987b: From product to process: introspective methods in second language research. In C. Færch and G. Kasper (eds), *Introspection in Second Language Research*. Clevedon: Multilingual Matters, 5–23.

Fathman, A. 1975: The relationship between age and second language productive ability. *Language Learning*, 25 (2), 245–53.

Felix, S. W. and Weigl, W. 1991: Universal Grammar in the classroom: the effects of formal instruction on second language acquisition. *Second Language Research*, 7 (2), 162–81.

Felix, S. and Zobl, H. 1994: Asymmetries in second language acquisition data.

Studies in Second Language Acquisition, 16 (4), 475–84.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., and Thal, D. 2000: Measuring variability in early child language: don't shoot the messenger. *Child Development*, 71 (2), 323–8.

Finney, M. A. 1997: Markedness, operator movement and discourse effects in the acquisition of purpose clause constructions in a second language. *Second Language Research*, 13 (1), 10–33.

Flanigan, B. O. 1995: Anaphora and relativization in child second language acquisition. *Studies in Second Language Acquisition*, 17 (3), 331–51.

Flege, J. E., Frieda, E. M., Walley, A. C., and Randazza, L. A. 1998: Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, 20 (3), 155–87.

Flynn, S. 1987: *A Parameter-Setting Model of L2 Acquisition: Experimental Studies in Anaphora*. Dordrecht: D. Reidel.

Foster, P. and Skehan, P. 1996: The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18 (3), 299–323.

Fraser, C. A. 1999: Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21 (2), 225–41.

Fraser, C., Bellugi, U., and Brown, R. 1963: Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior*, 2 (2), 121–35.

Gass, S. M. 1979: Language transfer and universal grammatical relations. *Language Learning*, 29 (2), 273–91.

Gass, S. M. 1980: An investigation of syntactic transfer in adult second language learners. In R. C. Scarcella and S. D. Krashen (eds), *Research in Second Language Acquisition*. Rowley, MA: Newbury House, 132–41.

Gass, S. M. 1983: The development of L2 intuitions. *TESOL Quarterly*, 17 (2), 273–91.

Gass, S. M. 1987: The resolution of conflicts among competing systems: a bidirectional perspective. *Applied Psycholinguistics*, 8 (3), 329–50.

Gass, S. M. 1989: How do learners resolve linguistic conflicts? In S. M. Gass and J. Schachter (eds), *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press, 183–99.

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

Gass, S. M. and Mackey, A. (eds) 1998: The role of input and interaction in second language acquisition. Special issue: *Modern Language Journal*, 82 (3).

Gass, S. M. and Mackey, A. 2000: *Stimulated Recall Methodology in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S., Mackey, A., Alvarez-Torres, M. J., and Fernández-García, M. 1999: The effects of task repetition on linguistic output. *Language Learning*, 49 (4), 549–81.

Gathercole, S. E. and Baddeley, A. D. 1993: *Working Memory and Language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Giacalone Ramat, A. 1992: Grammaticalization processes in the area of temporal and modal relations. *Studies in Second Language Acquisition*, 14 (3), 297–322.

Glesne, C. and Peshkin, A. 1992: *Becoming Qualitative Researchers: An Introduction*. London: Longman.

Goodluck, H. 1996: The act-out task. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 147–62.

Goss, N., Zhang, Y.-H., and Lantolf, J. P. 1994: Two heads may be better than one: mental activity in second-language grammaticality judgments. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 263–86.

de Graaff, R. 1997: The eXperanto experiment: effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19 (2), 249–76.

Green, B. 1954: Paired comparison scaling procedures. In G. Lindzey (ed.), *Handbook of Social Psychology. Vol. 1*. Reading, MA: Addison-Wesley, 344–7.

Green, P. S. and Hecht, K. 1992: Implicit and explicit grammar: an empirical study. *Applied Linguistics*, 13 (2), 168–84.

Greenbaum, S. (ed.) 1977: *Acceptability in Language*. The Hague: Mouton.

Grigg, T. J. 1986: The effects of task, time, and rule knowledge on grammar performance for three English structures. *University of Hawai'i Working Papers in ESL*, 5 (1), 37–60.

De Groot, A. M. B. and Keijzer, R. 2000: What is hard to learn is easy to forget: the roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50 (1), 1–56.

De Groot, A. M. B. and Kroll, J. F. (eds) 1997: *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Grotjahn, R. 1987: On the methodological basis of introspective methods. In C. Færch and G. Kasper (eds), *Introspection in Second Language Research*. Clevedon: Mulitilingual Matters, 54–81.

Hagen, L. K. 1994: Constructs and measurement in parameter models of second-language acquisition.

In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 61–87.

Hakuta, K. 1974: A preliminary report on the development of grammatical morphemes in a Japanese girl learning English as a second language. *Working Papers on Bilingualism*, 3, 18–43.

Hakuta, K. 1976: A case study of a Japanese child learning English as a second language. *Language Learning*, 26 (2), 321–51.

Hamilton, R. L. 1994: Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning*, 44 (1), 123–57.

Handorf, S., Watson-Gegeo, K., and Sato, C. J. 1993: *Research on Narrative: A Bibliography and Selected Annotations*. Technical Report No. 10, Center for Second Language Research. Honolulu: University of Hawai'i.

Hardison, D. M. 1996: Bimodal speech perception by native and nonnative speakers of English: factors influencing the McGurk effect. *Language Learning*, 46 (1), 3–73.

Harley, B. 1989: Functional grammar in French immersion: a classroom experiment. *Applied Linguistics*, 19 (3), 331–59.

Harley, B., Howard, J., and Hart, D. 1995: Second language processing at different ages: do younger learners pay more attention to prosodic cues to sentence structure? *Language Learning*, 45 (1), 43–71.

Harrington, M. 1987: Processing transfer: language-specific processing strategies as a source of interlanguage variation. *Applied Psycholinguistics*, 8 (3), 351–77.

Hatch, E. (ed.) 1978a: *Second Language Acquisition: A Book of Readings*. Rowley, MA: Newbury House.

Hatch, E. 1978b: Introduction. In E. Hatch (ed.), *Second Language*

Acquisition: A Book of Readings. Rowley, MA: Newbury House, 1–18.

Hatch, E. and Lazaraton, A. 1991: *The Research Manual.* Rowley, MA: Newbury House.

Haznedar, B. 2001: The acquisition of the IP system in child L2 English. *Studies in Second Language Acquisition*, 23 (1), 1–39.

Helms-Park, R. 2001: Evidence of lexical transfer in learner syntax: the acquisition of English causatives by speakers of Hindi-Urdu and Vietnamese. *Studies in Second Language Acquisition*, 23 (1), 71–102.

Hendriks, H. 2000: The acquisition of topic marking in L1 Chinese and L1 and L2 French. *Studies in Second Language Acquisition*, 22 (3), 369–97.

Holmen, A. 1993: Syntactic development in Danish L2. In K. Hyltenstam and Å. Viberg (eds), *Progression and Regression in Language: Sociocultural, Neuropsychological, and Linguistic Perspectives.* Cambridge: Cambridge University Press, 267–88.

Hoover, M. L. and Dwivedi, V. D. 1998: Syntactic processing by skilled bilinguals. *Language Learning*, 48 (1), 1–29.

House, J. 1996: Developing pragmatic fluency in English as a foreign language: routines and metapragmatic awareness. *Studies in Second Language Acquisition*, 18 (2), 225–52.

Huang, J. and Hatch, E. 1978: A Chinese child's acquisition of English. In E. Hatch (ed.), *Second Language Acquisition: A Book of Readings.* Rowley, MA: Newbury House, 118–31.

Hudson, T., Detmer, E., and Brown, J. D. 1995: *Developing Prototypic Measures of Cross-Cultural Pragmatics.* Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center.

Hulstijn, J. 1997: Second language acquisition research in the laboratory: possibilities and limitations. *Studies in Second Language Acquisition*, 21 (2), 131–43.

Hulstijn, J. and DeKeyser, R. (eds) 1997: Testing SLA theory in the research laboratory. Special issue. *Studies in Second Language Acquisition*, 19 (2).

Hyltenstam, K. 1988: Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development*, 9 (1 and 2), 67–84.

Hyltenstam, K. 1992: Non-native features of near-native speakers: on the ultimate attainment of childhood L2 learners. In R. J. Harris (ed.), *Cognitive Processing in Bilinguals.* Amsterdam: Elsevier Science, 351–68.

Inagaki, S. 1997: Japanese and Chinese learners' acquisition of the narrow-range rules for the dative alternation in English. *Language Learning*, 47 (4), 637–69.

Ioup, G., Boustagui, E., El Tigi, M., and Moselle, M. 1994: Reexamining the critical period hypothesis: a case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, 16 (1), 73–98.

Izumi, S., Bigelow, M., Fujiwara, M., and Fearnow, S. 1999: Testing the output hypothesis: effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition,* 21 (3), 421–52.

Jin, H. G. 1994: Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer. *Language Learning*, 44 (1), 101–22.

Johnson, D. M. 1992: *Approaches to Research in Second Language Learning.* New York: Longman.

Johnson, J. S. 1992: Critical period effects in second language acquisition: the effect of written versus auditory materials on the assessment of grammatical competence. *Language Learning*, 42 (2), 217–48.

Johnson, M. 2000: Interaction in the Oral Proficiency Interview: problems of validity. *Pragmatics*, 10 (2), 215–31.

Jones, F. R. 1995: Learning an alien lexicon: a teach-yourself case study. *Second Language Research*, 11 (2), 95–111.

Jourdenais, R. 1998: The effects of textual enhancement on the acquisition of Spanish preterit and imperfect. Ph.D. dissertation. Georgetown University.

Jourdenais, R. 2001: Cognition, instruction, and protocol analysis. In P. Robinson (ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, 354–75.

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*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center, 183–216.

Ju, M. K. 2000: Overpassivization errors by second language learners: the effect of conceptualizable agents in discourse. *Studies in Second Language Acquisition*, 22 (1), 85–111.

Juffs, A. 1998: Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning*, 48 (1), 107–47.

Juffs, A. and Harrington, M. 1995: Parsing effects in second language sentence processing. *Studies in Second Language Acquisition*, 17 (4), 483–516.

Juffs, A. and Harrington, M. 1996: Garden Path sentences and error data in second language sentence processing. *Language Learning*, 46 (2), 283–326.

Just, M. A. and Carpenter, P. A. 1987: *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon.

Kasper, G. (ed.) 1996: The development of pragmatic competence. Special issue. *Studies in Second Language Acquisition*, 18 (2).

Kasper, G. 1998: Datenerhebungsverfahren in der Lernersprachenpragmatik. *Zeitschrift für Fremdsprachenforschung*, 9 (1), 85–118.

Kasper, G. and Dahl, M. 1991: Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, 13 (2), 215–47.

Kasper, G. and Kellerman, E. (eds) 1997: *Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives*. London: Longman.

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

Kempe, V. and MacWhinney, B. 1998: The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition*, 20 (4), 543–87.

Klein, E. C. 1995: Second versus third language acquisition: is there a difference? *Language Learning*, 45 (3), 419–65.

Klein, W. and Perdue, C. 1992: *Utterance Structure: Developing Grammars Again*. Philadelphia: John Benjamins.

Klein, W., Coenen, J., van Helvert, K., and Hendriks, H. 1995: The acquisition of Dutch. In R. Dietrich, W. Klein, and C. Noyau (eds), *The Acquisition of Temporality in a Second Language*. Philadelphia: John Benjamins, 117–43.

Koike, D. A. 1998: What happens when there's no one to talk to? Spanish foreign language discourse in simulated oral proficiency interviews. In R. Young and A. W. He (eds), *Taking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Philadelphia: John Benjamins, 69–98.

Kormos, J. 1999: Simulating conversations in oral-proficiency assessment: a conversation analysis of

role plays and non-scripted interview in language exams. *Language Testing*, 16 (2), 163–88.

Kroll, J. F. and De Groot, A. M. B. 1997: Lexical and conceptual memory in the bilingual: mapping form to meaning in two languages. In A. M. B. De Groot and J. F. Kroll (eds), *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 169–99.

Labov, W. 1969: The study of language in its social context. *Studium General*, 23, 30–87. Reprinted in 1972, *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lakshmanan, U. and Teranishi, K. 1994: Preferences versus grammaticality judgments: some methodological issues concerning the governing category parameter in second-language acquisition. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 185–206.

Larsen-Freeman, D. and Long, M. H. 1991: *An Introduction to Second Language Acquisition Research*. London: Longman.

Laufer, B. and Eliasson, S. 1993: What causes avoidance in L2 learning: L1–L2 difference, L1–L2 similarity, or L2 complexity? *Studies in Second Language Acquisition*, 15 (1), 35–48.

Leather, J. (ed.) 1999: *Phonological Issues in Language Learning: A Supplement to Language Learning*, 49, Supplement 1.

Leopold, W. 1939: *Speech Development of a Bilingual Child. Vol. I: Vocabulary Growth in the First Two Years*. Evanston, IL: Northwestern University Press.

Leopold, W. 1948: The study of child language and infant bilingualism. *Word*, 4, 1–17.

Leow, R. P. 1993: To simplify or not to simplify: a look at intake. *Studies in Second Language Acquisition*, 15 (3), 333–55.

Leow, R. P. 1997: Attention, awareness, and foreign language behavior. *Language Learning*, 47 (3), 467–505.

Liceras, J. M. 1985: The role of intake in the determination of learners' competence. In S. M. Gass and C. G. Madden (eds), *Input in Second Language Acquisition*. Rowley, MA: Newbury House, 354–73.

Liceras, J. M. 1989: On some properties of the "pro-drop" parameter: looking for missing subjects in non-native Spanish. In S. M. Gass and J. Schachter (eds), *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press, 109–33.

Liceras, J. M. and Díaz, L. 1999: Topic-drop versus pro-drop: null subjects and pronominal subjects in the Spanish L2 of Chinese, English, French, German and Japanese speakers. *Second Language Research*, 15 (1), 1–40.

Lightbown, P., Spada, N., and Wallace, R. 1980: Some effects of instruction on child and adolescent ESL learners. In R. Scarcella and S. D. Krashen (eds), *Research in Second Language Acquisition*, Rowley, MA: Newbury House, 162–72.

Liskin-Gasparro, J. 1998: Linguistic development in an immersion context: how advanced learners of Spanish perceive SLA. *Modern Language Journal*, 82, (2), 159–75.

Long, M. H. 1983: Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, 4 (2), 126–41.

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*. Clevedon: Multilingual Matters, 123–67.

Lust, B., Flynn, S., and Foley, C. 1996: What children know about what they say: elicited imitation as a research

method for assessing children's syntax. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 55–76.

Lyster, R. 1998a: Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, 20 (1), 51–81.

Lyster, R. 1998b: Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, 48 (2), 183–218.

Lyster, R. and Ranta, L. 1997: Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, 19 (1), 37–66.

Mackey, A. 1999: Input, interaction, and second language development: an empirical study of question formation in ESL. *Studies in Second Language Acquisition*, 21 (4), 557–587.

MacWhinney, B. 1987: The Competition Model. In B. MacWhinney (ed.), *Mechanisms of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 249–308.

MacWhinney, B. 1997: Second language acquisition and the Competition Model. In A. M. B. De Groot and J. F. Kroll (eds), *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 113–42.

MacWhinney, B. and Bates, E. (eds) 1989: *The Crosslinguistic Study of Sentence Processing*. Cambridge: Cambridge University Press.

Major, R. C. (ed.) 1998: Interlanguage phonetics and phonology. Special issue. *Studies in Second Language Acquisition*, 20 (2).

Major, R. C. and Faudree, M. C. 1996: Markedness universals and the acquisition of voicing contrasts by Korean speakers of English. *Studies in Second Language Acquisition*, 18 (1), 69–90.

Malakoff, M. and Hakuta, K. 1991: Translation skill and metalinguistic awareness in bilinguals. In E. Bialystok (ed.), *Communication Strategies: A Psychological Analysis of Second-Language Use*. Oxford: Blackwell, 141–66.

Mandell, P. B. 1999: On the reliability of grammaticality judgement tests in second language acquisition research. *Second Language Research*, 15 (1), 73–99.

Markee, N. P. 1994: Toward an ethnomethodological respecification of second-language acquisition studies. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 89–116.

Masterson, D. 1992: Instruments for measuring grammaticality and preference: the status of modified NPs in English and Korean. *University of Hawai'i Working Papers in ESL*, 11 (1), 125–56.

Matusmura, M. 1994: Japanese learners' acquisition of the locality requirement of English reflexives. *Studies in Second Language Acquisition*, 16 (1), 19–42.

McDade, H. L., Simpson, M. S., and Lamb, D. E. 1982: The use of elicited imitation as a measure of expressive grammar. *Journal of Speech and Hearing Disorders*, 47 (1), 19–24.

McDaniel, D. and Cairns, H. Smith. 1996: Eliciting judgments of grammaticality and reference. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 233–54.

McDaniel, D., McKee, C., and Cairns, H. Smith (eds) 1996: *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press.

McKee, C. 1996: On-line methods. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 189–208.

Meara, P. 1986: The *Digame* project. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 101–10.

Mehnert, U. 1998: The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20 (1), 83–108.

Meisel, J. M. 1994a: Code-switching in young bilingual children. *Studies in Second Language Acquisition*, 16 (4), 413–39.

Meisel, J. M. (ed.) 1994b: *Bilingual First Language Acquisition: French and German Grammatical Development*. Philadelphia: John Benjamins.

Mellow, J. 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 (3), 325–50.

Menn, L. and Bernstein Ratner, N. (eds) 2000: *Methods for Studying Language Production*. Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, G. A. 1969: A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 6, 169–91.

Milroy, L. 1987: *Observing and Analysing Natural Language*. Oxford: Blackwell.

Mohan, B. A. 1977: Acceptability testing and fuzzy grammar. In S. Greenbaum (ed.), *Acceptability in Language*. The Hague: Mouton, 133–48.

Moyer, A. 1999: Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition*, 21 (1), 81–108.

Munnich, E., Flynn, S., and Martohardjono, G. 1994: Elicited imitation and grammaticality judgment tasks: what they measure and how they relate to each other. In E. Tarone, S. M. Gass, and A. D. Cohen (eds), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 227–43.

Murdock, B. B., Jr 1982: Recognition memory. In C. R. Puff (ed.), *Handbook of Research Methods in Human Memory and Cognition*. New York: Academic Press, 1–26.

Myles, F., Hooper, J., and Mitchell, R. 1998: Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, 48 (3), 323–63.

Myles, F., Mitchell, R., and Hooper, J. 1999: Interrogative chunks in French L2. *Studies in Second Language Acquisition*, 21 (1), 49–80.

Nation, P. 2001: *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, R. and McLaughlin, B. 1986: Language learning in multilingual subjects: an information-processing point of view. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 41–53.

Nicolaidis, E. and Genesee, F. 1996: A longitudinal study of pragmatic differentiation in young bilingual children. *Language Learning*, 46 (3), 439–64.

Norris, J. M. and Ortega, L. 2000: Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, 50 (3), 417–528.

Norris, J. M., Brown, J. D., Hudson, T., and Yoshioka, J. 1998: *Designing Second Language Performance Assessments*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center.

Nunan, D. 1989: *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Nunan, D. 1992: *Research Methods in Language Learning*. Cambridge: Cambridge University Press.

Nunan, D. 1996: Issues in second language acquisition research: examining substance and procedure.

In W. C. Ritchie and T. K. Bhatia (eds), *Handbook of Second Language Acquisition*. New York: Ritchie and Bhatia, 349–74.

Olson, G. M. and Clark, H. H. 1976: Research methods in psycholinguistics. In E. C. Carterette and M. P. Friedman, (eds), *Handbook of Perception. Vol. VII: Language and Speech*. New York: Academic Press, 25–74.

Ortega, L. 1999: Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21 (1), 109–48.

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

Papp, S. 2000: Stable and developmental optionality in native and non-native Hungarian grammars. *Second Language Research*, 16 (2), 173–200.

Paradis, J. and Genesee, F. 1996: Syntactic acquisition in bilingual children: autonomous or interdependent? *Studies in Second Language Acquisition*, 18 (1), 1–25.

Paribakht, T. S. and Wesche, M. 1999: Reading and "incidental" L2 vocabulary acquisition: an introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21 (2), 195–224.

Park, H. 2000: *When*-questions in second language acquisition. *Second Language Research*, 16 (1), 44–76.

Perdue, C. (ed.) 1982: *Second Language Acquisition by Adult Immigrants: A Field Manual*. Strasbourg: European Science Foundation.

Perdue, C. (ed.) 2000: The Structure of Learner Varieties. Special issue. *Studies in Second Language Acquisition*, 22 (3).

Perdue, C. and Klein, W. 1992: Why does the production of some learners not grammaticalize? *Studies in Second Language Acquisition*, 14 (3), 259–72.

Pérez-Leroux, A. T. and Glass, W. 1999: Null anaphora in Spanish second language acquisition: probabilistic versus generative approaches. *Second Language Research*, 15 (2), 220–49.

Perkins, K. and Larsen-Freeman, D. 1975: The effect of formal language instruction on the order of morpheme acquisition. *Language Learning*, 25 (2), 237–43.

Pfaff, C. W. 1992: Grammaticalization processes in the area of temporal and modal relations. *Studies in Second Language Acquisition*, 14 (3), 273–96.

Pica, T. 1994: Research on negotiation: what does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, 44 (3), 493–527.

Pica, T., Kanagy, R., and Falodun, J. 1993: Choosing and using communication tasks for second language instruction. In G. Crookes and S. Gass (eds), *Tasks and Language Learning: Integrating Theory and Practice*. Clevedon: Multilingual Matters, 9–34.

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

Pienemann, M. and Håkansson, G. 1999: A unified approach toward the development of Swedish as L2: a processability account. *Studies in Second Language Acquisition*, 21 (3), 393–420.

Polio, C. 1995: Acquiring nothing? The use of zero pronouns by nonnative speakers of Chinese and the implications for the acquisition of nominal reference. *Studies in Second Language Acquisition*, 17, 353–77.

Poplack, S., Wheeler, S., and Westwood, A. 1989: Distinguishing language contact phenomena: evidence from Finnish–English bilingualism. In K. Hyltenstam and L. Obler (eds), *Bilingualism Across the Lifespan*.

Cambridge: Cambridge University Press, 132–54.

Ravem, R. 1968: Language acquisition in a second language environment. *International Review of Applied Linguistics*, 6, 165–85. Reprinted 1974 in J. Richards (ed.), *Error Analysis*. London: Longman, 124–33.

Ravem, R. 1970: The development of Wh-questions in first and second language learners. *Occasional Papers*, Language Centre, University of Essex. Reprinted 1974 in J. Richards (ed.), *Error Analysis*. London: Longman, 134–55.

Rayner, K. (ed.) 1983: *Eye Movements in Reading: Perceptual and Language Processes*. New York: Academic Press.

Riney, T. J. and Flege, J. E. 1998: Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition*, 20 (3), 213–43.

Rintell, E. M. and Mitchell, C. J. 1989: Studying requests and apologies: an inquiry into method. In S. Blum-Kulka, J. House, and G. Kasper (eds), *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex, 248–72.

Robertson, D. 2000: Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16 (2), 135–72.

Robinson, P. 1994: Universals of word formation processes: noun incorporation in the acquisition of Samoan as a second language. *Language Learning*, 44 (4), 569–615.

Robinson, P. 1997: Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, 19 (2), 223–47.

Robinson, P. J. 2001: Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22 (1), 27–57.

Robinson, P. J. and Ha, M. A. 1993: Instance theory and second language rule learning under explicit conditions. *Studies in Second Language Acquisition*, 15 (4), 413–38.

Roebuck, R. F., Martínez-Arbelaiz, M. A., and Pérez-Silva, J. I. 1999: Null subjects, filled CPs and L2 acquisition. *Second Language Research*, 15 (3), 251–82.

Rosa, E. and O'Neill, M. D. 1999: Explicitness, intake, and the issue of awareness: another piece to the puzzle. *Studies in Second Language Acquisition*, 21 (4), 511–56.

Rose, K. R. 1994: On the validity of DCTs in non-western contexts. *Applied Linguistics*, 15 (1), 1–14.

Rose, K. R. 2000: An exploratory cross-sectional study of interlanguage pragmatic development. *Studies in Second Language Acquisition*, 22 (1), 27–67.

Rose, K. R. and Ono, R. 1995: Eliciting speech act data in Japanese: the effect of questionnaire type. *Language Learning*, 45 (2), 191–223.

Ross, J. R. 1979: Where's English? In C. J. Fillmore, D. Kemper, and W. S.-Y. Wang (eds), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press, 172–83.

Ross, S. and Berwick, R. 1992: The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14 (2), 159–76.

Rott, S. 1999: The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 21 (4), 589–619.

Rounds, P. L. and Kanagy, R. 1998: Acquiring linguistic cues to identify AGENT: evidence from children learning Japanese as a second language. *Studies in Second Language Acquisition*, 20 (4), 509–42.

Salaberry, M. R. 1998: The development of aspectual distinctions in L2 French classroom learning. *Canadian Modern Language Review/La revue canadienne des langues vivantes*, 54 (4), 508–42.

Salaberry, M. R. 1999: Procedural vocabulary: lexical signalling of conceptual relations in the development of past tense verbal morphology in classroom L2 Spanish. *Applied Linguistics*, 20 (2), 151–78.

Salaberry, M. R. and López-Ortega, N. 1998: Accurate L2 production across language tasks: focus on form, focus on meaning, and communicative control. *Modern Language Journal*, 82 (4), 514–32.

Sasaki, M. 1998: Investigating EFL students' production of speech acts: a comparison of production questionnaires and role plays. *Journal of Pragmatics*, 30 (4), 457–84.

Sasaki, Y. 1994: Paths of processing strategy transfers in learning Japanese and English as foreign languages. *Studies in Second Language Acquisition*, 16 (1), 43–72.

Sato, C. J. 1985: Task variation in interlanguage phonology. In S. M. Gass and C. G. Madden (eds), *Input in Second Language Acquisition*. Rowley, MA: Newbury House, 181–96.

Sato, C. J. 1990: *The Syntax of Conversation in Interlanguage Development*. Tübingen: Gunter Narr.

Scarcella, R. and Zimmerman, C. 1998: Academic words and gender: ESL student performance on a test of academic lexicon. *Studies in Second Language Acquisition*, 20 (1), 27–49.

Schachter, J. 1989: Testing a proposed universal. In S. M. Gass and J. Schachter (eds), *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press, 73–88.

Schachter, J. and Yip, V. 1990: Why does anyone object to subject extraction? *Studies in Second Language Acquisition*, 12 (4), 379–92.

Schlyter, S. 1993: The weaker language in bilingual Swedish–French children. In K. Hyltenstam and Å. Viberg (eds), *Progression and Regression in Language: Sociocultural, Neuropsychological, and Linguistic Perspectives*. Cambridge: Cambridge University Press, 289–308.

Schmidt, R. (ed.) 1995a: *Attention and Awareness in Foreign Language Learning*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center.

Schmidt, R. 1995b: Consciousness and foreign language learning: a tutorial on the role of attention and awareness in learning. In R. Schmidt (ed.), *Attention and Awareness in Foreign Language Learning*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center, 1–63.

Schmidt, R. and Frota, S. 1986: Developing basic conversational ability in a second language: a case study of an adult learner of Portuguese. In R. Day (ed.), *Talking to Learn: Conversation in Second Language Acquisition.* Rowley, MA: Newbury House, 237–326.

Schmitt, N. and Meara, P. 1997: Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, 19 (1), 17–36.

Schreuder, R. and Weltens, B. (eds) 1993: *The Bilingual Lexicon*. Philadelphia: John Benjamins.

Schumann, F. M. 1980: Diary of a language learner: a further analysis. In R. C. Scarcella and S. D. Krashen (eds), *Research in Second Language Acquisition*. Rowley, MA: Newbury House, 51–7.

Schumann, J. 1978: *The Pidginization Process: A Model for Second Language Acquisition*. Rowley, MA: Newbury House.

Schütze, C. T. 1996: *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Scott, M. L. 1994: Auditory memory and perception in younger and older adult second language learners. *Studies in Second Language Acquisition*, 16 (3), 263–281.

Seliger, H. 1989: Deterioration and creativity in childhood bilingualism. In K. Hyltenstam and L. Obler (eds), *Bilingualism Across the Lifespan*. Cambridge: Cambridge University Press, 173–84.

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

Shehadeh, A. 1999: Non-native speakers' production of modified comprehensible output and second language learning. *Language Learning*, 49 (4), 627–75.

Shirai, Y. and Kurono, A. 1998: The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning*, 48 (2), 245–79.

Simard, D. and Wong, W. 2001: Alertness, orientation, and detection: the conceptualization of attentional functions in SLA. *Studies in Second Language Acquisition*, 23 (1), 103–24.

Skehan, P. 1986: Cluster analysis and the identification of learner types. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 81–94.

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

Skehan, P. and Foster, P. 1997: Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1, 185–211.

Skiba, R. and Dittmar, N. 1992: Pragmatic, semantic, and syntactic constraints and grammaticalization. *Studies in Second Language Acquisition*, 14 (3), 323–49.

Slavoff, G. R. and Johnson, J. S. 1995: The effects of age on the rate of learning a second language. *Studies in Second Language Acquisition*, 17 (1), 1–16.

Slobin, D. I. and Welsh, C. A. 1973: Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson and D. Slobin (eds), *Studies of Child Language Development*. New York: Holt, Rinehart, and Winston, 485–97. Reprint of D. I. Slobin and C. A. Welsh (1968) Elicited imitation as a research tool in developmental psycholinguistics. Working Paper 10. Language Behavior Research Laboratory, University of California, Berkeley.

Smith, M. C. 1997: How do bilinguals access lexical information? In A. M. B. De Groot and J. F. Kroll (eds), *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 145–68.

Snodgrass, J. G. 1993: Translating *versus* picture naming: similarities and differences. In R. Schreuder and B. Weltens (eds), *The Bilingual Lexicon*. Philadelphia: John Benjamins, 83–114.

Snow, C. E., Cancino, H., de Temple, J., and Schley, S. 1991: Giving formal definitions: a linguistic or metalinguistic skill? In E. Bialystok (ed.), *Communication Strategies: A Psychological Analysis of Second-Language Use*. Oxford: Blackwell, 90–112.

Sorace, A. 1990: Indeterminacy in first and second languages: theoretical and methodological issues. In J. de Jong and D. Stevenson (eds), *Individualizing the Assessment of Language Abilities*. Clevedon: Multilingual Matters, 127–53.

Sorace, A. 1996: The use of acceptability judgments in second language acquisition research. In W. C. Ritchie and T. K. Bhatia, (eds), *Handbook of Second Language Acquisition*. San Diego: Academic Press, 375–409.

Su, I.-R. 2001: Transfer of sentence processing strategies: a comparison of

L2 learners of Chinese and English. *Applied Psycholinguistics*, 22 (1), 83–112.

Swain, M. 1972: Bilingualism as a first language. Ph.D. dissertation. University of California at Irvine.

Swain, M. and Cumming, A. 1989: Beyond methodology: behind research. In J. H. Esling (ed.), *Multicultural Education and Policy: ESL in the 1990s*. Toronto: OISE Press, 88–106.

Swain, M. and Lapkin, S. 1998: Interaction and second language learning: two adolescent French immersion students working together. *Modern Language Journal*, 82 (3), 320–37.

Swain, M. and Wesche, M. 1973: Linguistic interaction: case study of a bilingual child. *Working Papers on Bilingualism*, 1, 10–34.

Swain, M., Dumas, G., and Naiman, N. 1974: Alternatives to spontaneous speech: elicited translation and imitation as indicators of second language competence. *Working Papers on Bilingualism*, 4, 68–79.

Tanaka, S. and Kawade, S. 1982: Politeness strategies and second language acquisition. *Studies in Second Language Acquisition*, 5 (1), 18–33.

Tarone, E. 1979: Interlanguage as chameleon. *Language Learning*, 29 (1), 181–91.

Tarone, E. 1983: On the variability of interlanguage systems. *Applied Linguistics*, 4, 143–63.

Thomas, M. 1992: What do elicited imitation data reveal about comprehension? In D. Staub and C. Delk (eds), *The Proceedings of the Twelfth Second Language Research Forum*. East Lansing, MI: Michigan State University, 187–99.

Thomas, M. 1994: Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44 (2), 307–36.

Thornton, R. 1996: Elicited production. In D. McDaniel, C. McKee, and H. Smith Cairns (eds), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press, 77–102.

Tomiyama, M. 2000: Child second language attrition: a longitudinal case study. *Applied Linguistics*, 21 (3), 304–32.

Tomlin, R. S. 1984: The treatment of foreground–background in the on-line descriptive discourse of second language learners. *Studies in Second Language Acquisition*, 6 (2), 115–42.

Towell, R., Hawkins, R., and Bazergui, N. 1993: Systematic and nonsystematic variability in advanced language learning. *Studies in Second Language Acquisition*, 15 (4), 439–60.

Trahey, M. 1996: Positive evidence in second language acquisition: some long-term effects. *Second Language Research*, 12 (2), 111–39.

Trahey, M., and White, L. 1993: Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition*, 15 (2), 181–204.

UCLES (University of Cambridge Local Examinations Syndicate) 1990: *First Certificate in English/Certificate of Proficiency in English: Instructions to Oral Examiners*. Cambridge: UCLES.

Valdman, A. (ed.) 1988: The assessment of foreign language oral proficiency. Special issue. *Studies in Second Language Acquisition*, 10 (2).

Van den Branden, K. 1997: Effects of negotiation on language learners' output. *Language Learning*, 47 (4), 589–636.

van Lier, L. 1988: *The Classroom and the Language Learner*. London: Longman.

VanPatten, B. and Cadierno, T. 1993: Explicit instruction and input processing. *Studies in Second Language Acquisition*, 15 (2), 225–43.

VanPatten, B. and Oikkenon, S. 1996: Explanation versus structured input in processing instruction. *Studies in*

Second Language Acquisition, 18 (4), 495–510.

Verhoeven, L. T. 1994: Transfer in bilingual development: the linguistic interdependence hypothesis revisited. *Language Learning*, 44 (3), 381–415.

Viberg, Å. 1993: Crosslinguistic perspectives on lexical organization and lexical progression. In K. Hyltenstam and Å. Viberg (eds), *Progression and Regression in Language: Sociocultural, Neuropsychological, and Linguistic Perspectives.* Cambridge: Cambridge University Press, 340–85.

Walters, J. 1979: The perception of politeness in English and Spanish. In C. A. Yorio, K. Perkins, and J. Schachter (eds), *On TESOL '79: The Learner in Focus*. Washington, DC: TESOL, 288–96.

Wang, C. and Lee, T. H. 1999: L2 acquisition of conflation classes of prenominal adjectival participles. *Language Learning*, 49 (1), 1–36.

Watson, I. 1991: Phonological processing in two languages. In E. Bialystok (ed.), *Communication Strategies: A Psychological Analysis of Second-Language Use*. Oxford: Blackwell, 25–48.

Weber-Fox, C. and Neville, H. J. 1999: Functional neural subsystems are differentially affected by delays in second language immersion: ERP and behavioral evidence in bilinguals. In D. Birdsong (ed.), *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum Associates, 23–38.

Wesche, M. and Paribakht, T. S. 1996: Assessing vocabulary knowledge: depth vs. breadth. *Canadian Modern Language Review*, 53 (1), 13–40.

White, L. 1986: Implications of parametric variation for adult second language acquisition: an investigation of the pro-drop parameter. In V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon Press, 55–72.

White, L. 1989: The adjacency condition on case assignment: do L2 learners observe the Subset Principle? In S. M. Gass and J. Schachter (eds), *Linguistic Perspectives on Second Language Acquisition*. Cambridge: Cambridge University Press, 134–58.

White, L. 1991: Adverb placement in second language acquisition: some effects of positive and negative evidence in the classroom. *Second Language Research*, 7 (2), 133–61.

White, L., Spada, N., Lightbown, P. M., and Ranta, L. 1991: Input enhancement and L2 question formation. *Applied Linguistics*, 12 (4), 416–32.

White, L., Bruhn-Garavito, J., Kawasaki, T., Pater, J., and Prévost, P. 1997: The researcher gave the subject a test about himself: problems of ambiguity and preference in the investigation of reflexive binding. *Language Learning*, 47 (1), 145–72.

Williams, J. 1999: Learner-generated attention to form. *Language Learning*, 49 (4), 583–625.

Wode, H. 1999: Incidental vocabulary acquisition in the foreign language classroom. *Studies in Second Language Acquisition*, 21 (2), 243–58.

Wolfe-Quintero, K. 1992: Learnability and the acquisition of extraction in relative clauses and *Wh*-questions. *Studies in Second Language Acquisition*, 14 (1), 39–70.

Wolfe-Quintero, K., Inagaki, S., and Kim, H. 1998: *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center.

Wolter, B. 2001: Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition*, 23 (1), 41–69.

Yang, L. R. and Givón, T. 1997: Benefits and drawbacks of controlled laboratory studies of second language

acquisition. *Studies in Second Language Acquisition*, 19 (2), 173–93.

Ying, H. G. 1996: Multiple constraints on processing ambiguous sentences: evidence from adult L2 learners. *Language Learning*, 6 (4), 681–711.

Young, R. 1995: Conversational styles in language proficiency interviews. *Language Learning*, 45 (1), 3–42.

Young, R. and He, A. W. (eds) 1998: *Taking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Philadelphia: John Benjamins.

Young, R. and Milanovic, M. 1992. Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14 (4), 403–24.

Yuan, B. 1995: Acquisition of base-generated topics by English-speaking learners of Chinese. *Language Learning*, 45 (4), 567–603.

Yuan, B. 1997: Asymmetry of null subjects and null objects in Chinese speakers' L2 English. *Studies in Second Language Acquisition*, 19 (4), 467–97.

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

Zobl, H. 1992: Sources of linguistic knowledge and uniformity of nonnative performance. *Studies in Second Language Acquisition*, 14 (4), 387–402.