

THE BILINGUAL RESEARCH JOURNAL  
Winter 1996, Vol. 20, No. 1, pp. 69-92

## **NATURALISTIC LANGUAGE ASSESSMENT OF LEP STUDENTS IN CLASSROOM INTERACTIONS**

Leo Gomez  
University of Texas, Pan American

Richard Parker, Rafael Lara-Alecio, & Salvador Hector Ochoa  
Texas A&M University, College Station

Richard Gomez, Jr.  
Texas Tech University

### **Abstract**

The purpose of this study was to assess language competence of second language learners through observing natural interactions in transitional bilingual classrooms. We first developed an instrument in accordance with current second language acquisition theory, and obtained inter-rater reliability on observation categories. We then piloted the instrument in six Grade 5 bilingual classrooms, targeting the social language of 24 individual students. Each student was observed for a total of 12 hours, during twenty-four 30-minute sessions, over a four-week period. Resulting data were examined for measurement stability over time, criterion-related validity, and construct validity evidence (through item clustering). We also discuss the efficiency and utility of the overall naturalistic language observation procedure.

### **Introduction**

The assessment of Limited-English-Proficient (LEP) students' oral language proficiency has traditionally been conducted through formal testing. Despite their popularity, formal language assessments have been challenged on several fronts (Damico, 1991). First, most tests segment

language into linguistic components, losing a more holistic picture of communicative effectiveness. Second, test response behaviors are often artificial; they would not be found in natural situations of language communication. Third, scoring of formal tests is prescriptive; only one answer is deemed correct. The contextual variability of language is not considered in assessment. Fourth, formal testing typically omits a receiver of communication, or an audience; the student is asked to communicate in a vacuum. Fifth, formal tests usually lack purpose or communicative intent for the student; purpose and intent are the sole province of the examiner. Finally, the psychometric qualities of many formal oral language proficiency measures are inadequate (Damico, 1991; Lieberman & Michael, 1986; McCauley & Swisher, 1984). Given these concerns, accurate assessment of oral language proficiency of LEP students is considered to be a challenge (Doughty & Pica, 1986). The use of formal instruments to assess oral language proficiency lags behind recent second language theory and theoretical research. Most influential second language acquisition (SLA) theories and theoretical research regard SLA as a social phenomenon, developed within and measured by growing competency within meaningful social exchanges (Cummins, 1981, 1986; Enright & McCloskey, 1985; Hatch, 1978; Hatch, Flashner, & Runt 1986; Krashen, 1982, 1985a, 1985b, 1985c; Sato, 1986). Language acquisition involves interaction with others in a variety of contexts, toward the goal of meaningful, functional communication (Krashen, 1982). This view of second language acquisition suggests that natural language assessment is essential. This social learning perspective on SLA helps identify those language attributes or components which should be assessed in measuring language competency. The social learning perspective recommends assessing oral language in meaningful, realistic situations over extended periods of time, and within a range of social situations involving two-way communication (Morrow, 1985; Nix, 1983). The emphasis is on "negotiation of meaning" between persons (Swain, 1985). Little evidence exists supporting the discrimination or utility of traditional components of oral language (e.g., vocabulary, fluency, pronunciation) (Hendricks, Scholz, Spurling, Johnson, & Vandenburg, 1980; Mullen, 1980). Instead, new components are suggested for SLA assessment, each of which should be measured in integrated, holistic social situations, e.g.

(a) effectiveness of meaning transmission, (b) fluency of meaning transmission, and (c) appropriateness of meaning transmission (Damico 1991; Mullen, 1980; Oiler 1991). One social situation providing language assessment opportunities is the public school classroom, where communicative tasks increasingly are performed by cooperative dyads and small groups (Alvarado, 1992). The implications for SLA assessment are clear and direct. Since the late seventies, SLA assessment has been urged to focus on functional and social aspects of the language acquisition process rather than form (Day, 1986; Doughty & Pica, 1986; Hatch, 1978; Long, 1981; Sato, 1986). SLA should be assessed through conversation rather than through acquisition and expansion of component structures (Hatch, 1978). Yet most formal (often mandated) tests of SLA depend on artificially structured, non-social tasks, with isolated (and arguably artificial) vocabularies and syntactic structures. Such tests omit or treat superficially (e.g. teacher perception checklists) language competence in negotiating meaning in real social tasks. Five exceptions have been identified, all prototypes without well-established measurement properties, and not generally known or used in the field: Social Interactive Coding System (SICS) (Rice, Sell, & Hadley, 1987); Environmental Communication Profile (ECP) (Calvert & Murray, 1985); Systematic Observation of Communicative Interaction (SOCI) (Damico & Oiler, 1985); Spotting Language Problems (SLP) (Damico & Oiler, 1985); Student Oral Language Observation Matrix (SOLOM) (Monte Bello USD, 1987). Of these five, none meets the following basic measurement-related and theory-related criteria: (a) adequate interrater reliability; (b) adequate stability over time; (c) adequate criterion-related validity; (d) inclusion of linguistic accuracy as well as social effectiveness of language. Only two present evidence for interrater reliability. Three present evidence for criterion-related validity. Two include content on linguistic accuracy plus social effectiveness. Only one presents evidence for stability over time. The gap between SLA theory and theoretical research on the one hand and formal SLA assessment on the other is understandable. All measurement in group social situations is challenging - especially so in classrooms, with the ambient noise level, multiple simultaneous interactions, rapid changes in activity structures, and potential reactivity to observers. Four other measurement obstacles to the naturalistic measurement of social

language are: (a) Variability across contexts, or inability to generalize: The social use of language may depend highly on the particular academic or social situation, with little interpretability or generalizability beyond that context (Cathcart, 1986; Cummins, 1981). (b) Complexity of context specification: As language meaning and effectiveness is contextually embedded, it must be described in light of a wide range of situational and verbal cues (Cummins, 1981). (c) Non-observable language competence: Much important non-observable interaction occurs which is important to understanding or judging social language (Faerch & Kasper, 1980; Van Lier, 1988). (d) Inability to separate using a language from knowing a language: Linguistic knowledge and social application are different, but inseparable in performance (Chomsky, 1980). These obstacles have led researchers to recommend a discourse analysis approach to measurement (Alvarado, 1992; Cathcart, 1986; 1989; Cazden, 1988; Fine, 1988; Hatch, 1978; Long, 1981), wherein verbatim transcriptions and copious contextual notes are relied upon to illuminate the social language use of a single student, dyad, or triad. However, the exhaustive data collection, sophisticated transcript analysis and complex interpretations from discourse analysis make that technique less useful when our goal is a practical language assessment tool. Verhoeven (1992), Alvarado (1992), and Higgs and Ray (1982) have emphasized the need when conducting SLA assessment to do more than describe intricate relationships. Relatively simple data analysis and interpretation are required, yielding judgments of language quality and accuracy, as well as language description.

### **Attributes of Social Language**

We attempted to measure language within a naturalistic social setting in which meaning was negotiated over a real task. To determine which facets of language to measure, we turned to both the theoretical literature and to other extant assessment instruments which followed modern SLA theory. Reviewing attempts by several writers to define the facets of effective social language (Cathcart, 1986; Hatch, 1992; Higgs & Ray; 1982), we were most influenced by Hatch (1992). Hatch (1992) contends that to fully describe and make judgments about social language requires inclusion of its cognitive, linguistic, and social facets.

We also learned and borrowed from the contents of five extant instruments: SICS (Rice, Sell, & Hadley, 1987); ECP (Calvert & Murray, 1985); SOCI (Damico & Oller, 1985); SLP (Damico & Oller, 1985); and SOLOM (Monte Bello USD, 1987). Anchored in the review of theoretical research and existing measurement scales of natural language, and tempered by our need for reliable qualitative judgments, we proposed to observe and judge the following fifteen social language attributes, categorized in Table I by Hatch's three facets.

Table 1  
*Initial Fifteen Observation Variables, by Three Language Facets*

Variable	Social*	Ling*	Cog*
B-2. Under: Understandability by others.	••	••	
F-6. Convers: Maintains conversation.	••	•	•
G-7. Delay: Absence of hesitations/delay interference.	••	•	
H-8. Self-corr: Absence of self-correction interference.	••	•	
K-11. Under: Apparent understanding of conversation.	••		
L-12. Partic: Willingness to participate in conversation.	••		
SIU: Self-Initiated Utterance: Student utterance not in	••		
M- 13. Attend: Attentiveness to important verbal information.	••		
N-14. Gesture: Appropriate Gestures, Body language, Humor, Expressions.	••		
A-I. Rel/Sens: Relevance and sensibility of utterance.	•	••	•
C-3. Prov.Info: Provides information needed by listener.	•		••
D-4. Top.Dev: Demonstrates topic development in conversation.	•	•	••
E-5. Spec.Voc: Uses appropriately specific vocabulary.		•	•
I-9. Accur: Accuracy in grammar, usage, & vocabulary.		••	
J-10. Simple: Uses unsimplified vocabulary and syntax.		••	

\*Social: Socially appropriate and functional or effective.

\*Linguistic: Language is intelligible, correct or accurate, and fluent.

\*Cognitive: Content of conversation is relevant and developed.

• = minor emphasis; •• = major emphasis.

### **Purpose**

The purpose of this study was to assess language competence of second language learners through observing natural interactions in bilingual classrooms. We wished to investigate the practicability of passive observation of language without structuring the social context. Besides concerns of practicability and efficiency, we were concerned with the measurement attributes of the resulting data. First, could we obtain interrater reliability? Second, how stable are natural language ratings from one observation episode to the next (to what extent are our observation results temporally bound)? Temporally or contextually dependent data are less useful for assessment, as they disallow general statements about generalizable habits or abilities. Finally, we examined evidence for construct validity of our new observation instrument by interpreting item intercorrelations in light of second language acquisition theory.

### **Research Questions**

The following research questions were posed for this study:

1. *Interrater reliability*: Does language assessment within natural classroom interactions exhibit acceptable interrater reliability?
2. *Score Stability*: Do data from naturalistic language assessment in classrooms demonstrate acceptable stability over time?
3. *Item Structure*: (3-a.) Internal Consistency: What overall internal consistency does the observation scale possess? (3-b.) Subscale Structure: Do intercorrelation patterns among language categories reflect facets of social language suggested by SLA theory?
4. *Criterion-related Validity*: What is the relationship between language assessment results from naturalistic observations on the one hand, and students' recent formal language assessment scores on the other?
5. *Utility*: Can naturalistic observation in classrooms be conducted with sufficient efficiency for applied, group research?

### Method Context

The context for this study was a summer transitional English program for Grade 5 at-risk LEP Hispanic students focusing on mathematics. The six-week program was funded through a grant from the federal Office of Bilingual and Minority Language Affairs (OBEMLA). The program was located in an urban school district of over 27,000 students. Of this total, 36% (9,700) are Hispanic, and 22% (6,000) are LEP. One hundred seven district students were bused to a single school to participate in this intensive summer program. Eight heterogeneous classrooms (each with 12-14 students) operated from 9:00 A.M. to 1:30 P.M., with breakfast served prior to class, and with a half hour lunch break. The teacher and instructional assistant for each class were bilingual/ESL certified, and easily adapted their mixture of English to Spanish to the needs of individual students, following a pedagogical transitional bilingual model (Lara-Alecio & Parker, 1994). Each school day included a 45 minute period for paired reciprocal learning, in which students were paired to work on math problems. Interactions with other classmates also were permitted, and students were permitted to move about the room. The dyads provided opportunities for intensive, task-related verbal interaction, liberally interspersed with social language. Students with similar English and Spanish language capabilities were paired, to facilitate communication. The teacher and instructional assistant rotated through the small classes, varying the language of instruction as needed for each individual and pair. Teachers participated in an initial language screening of individual students at program intake, so became familiar with individual needs.

*Participants.* The 5th grade LEP students were all from Bilingual or ESL classrooms, and were all classified "at-risk" (performing academically below the 23rd percentile by district norms on any of a variety of nationally standardized achievement tests). According to standardized language assessments (I.P.T., L.A.S.) on file, updated by an individual structured interview, students' English proficiency levels were: Level I-Non English Speaking: 14%; Level II-Limited English Speaking: 58%; Level III-Fluent English Speaking: 28%. For this study, students were selected who would likely exhibit both languages in unprompted interactions. These students were Limited English Speakers

(LES), and Limited or Fluent Spanish Speakers (LSS or FSS). We anticipated more use of English than Spanish, since the thrust of this transitional program was to use Spanish mainly to: (a) introduce new and difficult academic content, (b) explain or support English presentations, (c) socially interact in a relaxed atmosphere. Because of the intensive nature of this observation study, four students (two pairs) were randomly selected from each of the six heterogeneous program classrooms, a total of 24 respondents.

*Instrumentation.* We desired a practical, efficient instrument which also reflected current second language acquisition theory. Our criteria were that the instrument would: (a) rely on passive observations of naturalistic social situations in which students actually negotiate meaning, (b) yield descriptive results, as well as judgments of quality and accuracy, (c) include generally accepted social, linguistic, and cognitive attributes of language, (d) yield generalizable results through the desirable psychometric properties of interrater reliability and stability of scores over time, (e) permit efficient observations, scoring, and summary, and not require extensive staff training.

*Instrument development.* Establishing reliability occurred over a two-week period, and entailed several instrument revisions. Through trial and error, a five-minute interval time sample (ITS) technique provided the best window for capturing language attributes. A five-minute interval time sample was used; at the end of each five minutes, the observer makes a summative judgment of the content of the entire 5-minute interval. Independent, concurrent ratings were obtained from program classrooms by two bilingual ESL graduate students, both experienced bilingual teachers. Following short, 10-20 minute observational periods, the two researchers would operationalize and clarify language attributes, edit observation categories, re-structure judgment scales, and create descriptive anchors. After 20 hours of trial observations conducted over two weeks, the researchers had a completed instrument. Three-point or four-point judgment scales were constructed for each category. All scale values were "anchored" by written descriptors. Of the original 23 coding categories, only fifteen could be unambiguously operationalized. Further elimination of language attributes which were not observable, overlapped, or could not be consistently scored, resulted in only seven surviving categories. A 20-30

minute time observation period was selected as the minimum length required for somewhat stable scores. Each observation session entailed concurrently observing and coding two students.

*Final reliability sample.* The final reliability sample was obtained on five different students in three classrooms, carried out for 2.5 hours, over a two-day period. Observations were videotaped, for later review and double-checking disagreements. Reliability was calculated with Cohen's Kappa (Cohen, 1960), a conservative index for categorical agreement beyond chance. Calculations were based on a reliability corpus of 60 tallies per language variable per observer (12 codes per hour per student, over 2.5 hours). Kappa has been criticized for being overly conservative, as it eliminates all agreement attributable to chance. Kappa coefficients of .6 to .8 usually indicate very good agreement (Fleiss, 1981). To make the Kappa statistic less conservative, some statisticians recommend the ratio of Kappa to its maximum value: Kappa/KappaMax (Umesh, Peterson, & Sauber, 1989). We prefer a balanced consideration of all three indices: Percent Agreement, Kappa, and Kappa/KappaMax.

Table 2  
*Kappa Statistic for Interrater Reliability in 8 Language Variables*

Variable	Percent Agreement	Kappa Kappa	Kappa/Kappa Max
Under: Understandability by others.	.69	.53	.72
Convers: Maintains conversation.	deleted	deleted	deleted
Delay: Absence of hesitations/delay interference.	.90	.83	1.00
Self-corr: Absence of self-correction interference.	deleted	deleted	deleted
Under: Apparent understanding of conversation.	deleted	deleted	deleted
Partic: Willingness to participate in conversation.	.77	.64	.72
Attend: Attentiveness to important verbal info.	deleted	deleted	deleted
Gesture: Appropriate Gestures, Body language, Humor, Expressions.	.91	.81	1.00
Rel/Sens: Relevance and sensibility of utterance.	deleted	deleted	deleted
Prov.Info: Provides information needed by listener.	.73	.55	.83
Top.Dev: Demonstrates topic development in conversation.	.67	.43	1.00
Spec.Voc: Uses appropriate specific vocabulary.	deleted	deleted	deleted
Accur: Accuracy in grammar, usage, & vocabulary.	.76	.64	.67
Simple: Uses unsimplified vocabulary and syntax.	deleted	deleted	deleted

Table 2 shows simple Percent Agreement, Kappa, and the ratio Kappa/KappaMax for the final eight language variables, each rated on a four-point scale. The table also includes the prior set of 14 categories to show which were eliminated. High or high-moderate agreement was obtained for more than half of the categories, with moderate to low-moderate agreement for "Prov.Info." (K=.55), "Under" (K=.53) and "Top.Dev." (K=.43). The three lower Kappa scores were largely an artifact of a "skewed" reliability data sample, i.e. our data were unbalanced or unequally represented all points on the scale. This diagnosis is drawn from comparing Kappa with the Kappa/KappaMax statistic. Results of these calculations answered part of research question two, and encouraged continuance of the study.

*Procedure.* Following the reliability study, the main language sample was obtained over the remaining four weeks of the summer program. Six weekly language samples of 30-minutes each were obtained from 4 students per classroom, within 6 classrooms, a total of 24 observations for each of 24 students. Students were repeatedly observed in the same dyads on Tuesday, Wednesday, and Thursday mornings, over four weeks. Students engaged in paired, collaborative math problem-solving, with assistance from rotating bilingual teachers and instructional assistants. Researchers video-recorded selected observation sessions during this time to augment the observers' field notes. The video recorders were placed in classrooms before the study began, and continuously thereafter so students were accustomed to their presence. During the designated thirty minute period, observers coded every five minutes, a summary rating to represent the content of the preceding 5-minute period (interval recording). The codes were 1-4 ratings of eight language attributes, plus three additional features: (a) the language spoken, (b) whether code switching (English to Spanish or vice versa) occurred within the five minute period, and (c) the count of self-initiated utterances (not direct responses) per student. Coders later played back videotapes to check their ratings. Although videotaping did not prove to be essential, it was useful for resolving some ambiguities.

## Results

The research question of interrater reliability was answered in the previous section. The main language corpus provided answers to the remaining questions of: (a) Score stability over time, (b) Whole-Scale internal consistency and meaningful item patterns (c) criterion-related

The main language corpus was six weekly 30-minute language samples collected over four weeks on each of 24 students, within six classrooms. Thus, for each student we obtained a total of 24 (6 sessions x 4 weeks) average session scores over the four weeks. The complete dataset was composed of 576 (24 sessions x 24 students) average session ratings on each of seven language attributes - 4032 mean scores in all. These mean scores represented aggregates of 24,192 individual ratings.

Score stability was assayed through three analyses: (a) partitioning of variance, (b) internal consistency of weekly scores, and (c) visual inspection of scores over the four weeks.

**Partitioning of Variance.** The generalizability approach to measurement reliability entails partitioning logical sources of score variability through analysis of variance. For each language code, we conducted a mixed ANOVA with one repeated measure, Weeks (4 levels), and with two grouping variables: Student (24 levels), and Session (6 levels). In the desirable case of reliable measurement, we would anticipate little variability accounted for by individual sessions, separate from that accounted for by weeks. Furthermore, we would expect some variability due to the variable, Weeks, but most due to individual differences, the variable Students. Table 3 presents Eta-squared effect sizes for main effects and for interactions. Eta-squared is an index of effect-size, the proportion of total variance accounted for by each effect. The table also indicates with asterisks those effect sizes from statistically significant ( $p < .01$ ) F-ratios.

Table 3  
*Effect Sizes (Eta-squared) for Main Effects and Repeated Measures*

Language Attribute	Main Effects			Weeks	Repeated Measures Effects		
	Student	Session	Error		Weeks *	Weeks * Student	Weeks * Session
Under.	.75*	.02	.23	.03	.47*	.05	.45
Prov. Info.	.61*	.20	.20	.03	.22	.32	.42
Top.Dev.	.58*	.34*	.09	.00	.63*	.20	.16
Delay	.92*	.02	.06	.00	.63*	.03	.34
Accur.	.62*	.01	.37	.08*	.41*	.05	.45
Partic.	.34*	.04*	.11	.11*	.51*	.05	.33
SIUs	.80*	.03	*.17	.12*	*.33*	.06	.44

\*p < .01

Table 3 shows that main effects (first three columns), students' individual differences accounted for a large proportion (58%-92%) of score variance. For only two variables Prov.Info. and Top.Dev., Session also accounted for at least 20% of the variance, reflecting less stable measurement. For repeated measures effects, the pattern of results was similar, but a smaller percent (22%-63%) of the variance was accounted for by our predictors. Weeks accounted for a small proportion (12% or less) of the total variance. Again, for both Prov.Info. and Top.Dev., Session accounted for 32% and 20% of the variance, neither statistically significant. In summary, our partitioning of variance supports stable measurement for five of the seven variables. The variables, Prov.Info. and Top.Dev. are less stable.

Weekly consistency. Reliability of each language attribute also was assessed through intercorrelating scores from the six sessions for each week, for N=24 students. In the desirable case of reliable measurement, we would anticipate high item intercorrelations within a particular week, and similar coefficients across the four weeks. Table 4 below contains average intercorrelations among six sets of weekly scores, presented by week and by attribute.

Table 4  
*Average intercorrelations \*\* of Session Scores  
 by Language Attribute and Week (N=24 students)*

Language Attribute	Week 1	Week 2	Week 3	Week 4	Average
Under.	*49*	.33	.57*	.58*	*49*
Prov. Info.	.36	.35	.44	*95*	.52*
Top. Dev.	.56*	.56*	.51*	.17	.44
Delay	.87*	.84*	.76*	.73*	.80*
Accur.	.40	.28	.46*	.47*	.40
Partic.	.68*	.78*	.68	*57*	.68*
S.I.U.	.60*	.58*	.48	.70*	.59*

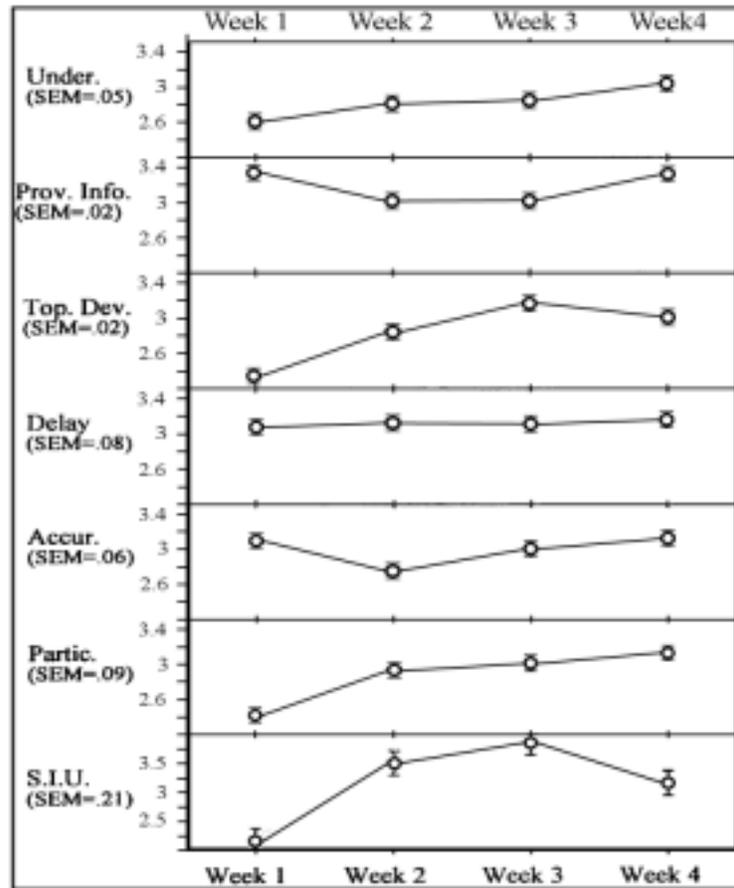
\*  $p < .01$

\*\* Tabled coefficients are averages among six sets of scores per week.

Tabled average weekly intercorrelations (first four data columns) ranged from .17 (Topic Dev.) to .87 (Delay). Grand averages over four weeks (last data column) ranged from .40 (Accur.) to .80 (Delay). Three attributes (Delay, Partic., S.I.U.) showed the desirable pattern of moderately high and stable intercorrelations. An additional three (Under., Accur., and Top.Dev.) showed moderate intercorrelations for all weeks but one. In summary, the data are only somewhat supportive of within-week data stability.

*Visual inspection.* Evidence for stable measurement also may include lack of "bounce" of scores over time in relationship to the standard error of measurement each week. Trend lines of stable measurements should either be relatively flat, or change in linearity over time periods sufficient for growth or development. We anticipated either no growth (flat trend lines) or slight growth over the short, four-week program period. In Figure 1, mean weekly scores are plotted, with 95% Confidence Interval (2-SEM) error bars. The average weekly SEM is printed under each variable name. Most error bars are so short they are barely visible. Note that six attributes are graphed on the same scale; the exception, S.I.U., was a frequency count rather than a 1-4 rating.

Figure 1  
*Mean Weekly SEM Scores*



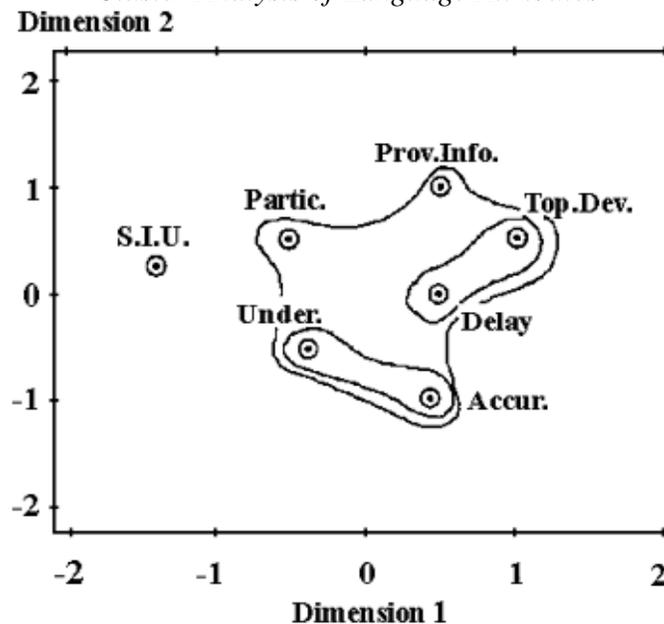
Only three of the seven language attributes (Under., Delay, Partic.) showed the desirable flat or linear trends. Variability of the other four attributes reflected language which changed by situation, rather than stable language skill. Four of the attributes (Under., Top.Dev., Partic., S.I.U.) showed general improvement over the four weeks, although two (Top.Dev., S.I.U.) evidenced a pattern of deterioration during the final week.

### Whole Scale Internal Consistency and Item Patterns

*Cronbach's Alpha.* Inasmuch as our whole observation scale represents a unitary construct of social language performance, one might expect high internal consistency of all items. Results of a scale reliability test, yielded a moderately high Cronbach's Alpha of .78, and a standardized Alpha of .90. The difference between Alpha and standardized Alpha indicates widely differing variances among items, especially by the differently scaled S.I.U. The item-to-total scale correlations ranged from .56 to .81: Under: .81, Delay: .69, Partic.: .62, Gesture: .54, Prov. Info.: .69, Top.Dev.: .65, Accur.: .72, and S.I.U.: .56.

*Cluster analysis.* We also were interested in relationship patterns among the language attributes. We expected that the cognitive, linguistic, and social facets of natural language interactions would cluster together in intercorrelations. In Figure 2, we submitted the seven ratings (Under., Prov.Inf., Top.Dev., Delay, Accur, Partic., S.I.U.) to multidimensional scaling (MDS).

Figure 2  
*Cluster Analysis of Language Attributes*



MDS is a non-parametric "poor man's" factor analysis, suitable when small N's, ordinal measurement, and non-normal samples disallow use of the stronger parametric alternative. MDS creates a two-dimensional map of item clusters, permitting two types of interpretation. First, we can identify clusterings of language attributes with similar score patterns across students and weeks. The second interpretation is to name the two map dimensions, given the positions of each attribute score. Prior to MDS scaling, S.I.U. was rescaled to prevent scale artifacts appearing as interpretable map configurations.

Kruskal's monotonic MDS procedure resulted in a two-dimensional map which accounted for 92% of the variance among the seven plotted scores. Clusters were identified from a supplemental scree plot. The map shows two internally cohesive clusters: Under. + Accur. ( $r=.81$ ), and Top.Dev. + Delay ( $r=.79$ ). S.I.U. was most isolated, correlating only .50 with the nearest cluster. Most closely related to S.I.U. was another relatively isolated attribute, Partic. Both S.I.U. and Partic. relate to initiative and risk-taking in a social situation. In the first cluster, Under. and Accur. had both been categorized as representing the "Linguistic" facet of language. In the second cluster, Top.Dev. and Delay were not categorized under the same facet. Therefore, the scaling map only somewhat bore out the language facets identified in the literature.

### **Criterion-Related Validity**

Finally, we investigated the relationship between the language attribute scores and an external criterion measure, the level of English proficiency of the 24 students. English proficiency was judged as "non-proficient" (Level I), "limited proficient" (Level II), or "proficient" (Level III), based on existing IPT test scores and an individual intake interview. We performed F-tests on the differences in attribute ratings. Table S shows a pattern of increasing mean scores from Level I to Level II to Level III for all seven language attributes. Six of the seven trends were highly significant; S.I.U. was the exception. Therefore, the criterion variable of English proficiency level (as indicated by standardized test score records) was highly supportive of the validity of this naturalistic language assessment.

Table 5  
*Tests for Differences among Mean Language Attribute Scores,  
 by English Proficiency Level*

<b>Language Attribute</b>	<b>Level I</b>	<b>Level II</b>	<b>Level III</b>	<b>F-Value</b>	<b>P-Value</b>
Under.	2.29	2.90	2.99	10.73	.0002
Prov.Inf	2.36	2.81	2.94	9.39	.0005
Top.Dev.	1.68	2.82	3.13	7.56	.0024
Delay	2.09	3.45	3.76	27.70	<.0001
Accur	2.48	3.02	3.21	9.74	.0004
Partic	2.29	3.11	3.22	10.23	.0003
S.I.U	2.44	3.66	3.74	1.67	.2015

*Efficiency of Use.* In considering "efficiency of use," we will comment separately on the three phases of (a) reliability training, (b) observation, and (c) scoring and reporting of results. In this pilot study, reliability training is difficult to isolate, as it was mixed with instrument development. However, we estimate that a two-hour training session, followed by 1-2 hours of supervised scoring in classrooms would suffice to bring bilingual teachers to the level of reliability we achieved. The actual coding is not burdensome, as it does not entail verbatim transcription, nor rapid judgments, but rather simple coding after each 5-minute interval.

Observation efficiency is another matter; larger time samples are required in unstructured language situations, as the lack of structure permits more response variability. This study did not satisfactorily indicate a sufficient time sample. However, we estimate that a minimum of six 30-minute samples, over two or three weeks would be needed for reasonable data stability. This time expenditure exceeds that of most standardized language tests.

Scoring and reporting of results was relatively efficient, requiring approximately 5 minutes per session. On the protocol for each 30-minute observation session, approximately six ratings were totaled and averaged for each of the seven language categories. These session averages were then averaged again over each week.

## Discussion

We observed 24 students in natural classroom interactions over a five-week period, using observational content from current second language acquisition theory. Combining this content with rigorous measurement methodology, we pursued a practical and technically adequate assessment tool. We posed questions about: (a) interrater reliability, (b) stability of results over time, (c) theoretically meaningful item cluster patterns, (d) criterion-related validity with formal language test scores, and (e) efficiency of use.

One notable outcome of this instrumentation study was the reduction of our first list of language attributes (coding categories) from 15 to 7. Item overlap (with mutual confusion) was one reason for dropping attributes, but more important were problems in operationalizing and consistently interpreting scale values. Reduced fullness and richness of description has usually accompanied the quest for improved judgment reliability. The surviving seven items still succeed in representing the three conceptual facets of language: Social (Under., Prov. Info., Delay, Partic., S.I.U.'s), Linguistic (Under., Accur.), and Cognitive (Prov.Info., Top.Dev.). The omitted eight language attributes may warrant inclusion in a more qualitative commentary, in which reliable judgment is not sought.

At least moderately strong interrater reliability was soon achieved with the surviving set of seven coding categories. However, this was accomplished by the same two persons earlier involved in several hours of instrument development. We cannot yet assay our efficiency in training professionals to reliability who have no prior history with the observation instrument. Our success in making reliable qualitative judgments on a 4-point scale for relatively global attributes (e.g. accuracy in grammar, usage and vocabulary) was rewarding. When informal checklists and ratings are used to make important decisions based on language status or improvement, we should consider improving the technical adequacy of these measures.

Stability of scores over time is a complex and too-often ignored criterion of sound assessment. In repeated observations of unstructured contexts we expect performance to vary due to mood and fatigue, varying opportunity, interest level, etc. Yet this variability should be less

than the similarity of scores obtained over a limited time period. Lack of score stability reduces the generalizability of our summaries of a student's language performance; such summaries of typical performance are commonly required by schools. One additional consideration in measurement stability is the phenomenon of linear change (typically growth, learning, or habit formation), a type of score variability which usually is not a sign of instability.

Our first approach to estimating score stability, partitioning of variance, consistently indicated stable scores. The second and more demanding investigative method, intercorrelating each set of weekly scores, was also supportive, but less so, for about 2/3 of the analyses only. Another demanding analysis, a line graph of mean scores with error bars, provided more equivocal evidence for stability. The graph depicted low within-week score variability, but high (non-linear) between-week variability for nearly half of the language attributes (Prov.Info., Top. Dev., S.I.U.). Thus, our observation procedures do not yet demonstrate a desirable level of score stability, although our data suggest that with additional development work, stability can be obtained with some indices, e.g. Delay, Under, Partic. We cannot compare our findings with other naturalistic language observation studies, as we found no others which thoroughly investigated score stability.

We investigated scale structure in two ways, whole scale internal consistency and multidimensional scaling, hypothesizing theoretically meaningful item clusters. Cronbach's standardized Alpha of .90 indicated only reasonably strong overall internal consistency for our scale, especially considering the small number of items. This result suggests that the scale tends to represent a unitary construct of social language competence. Our follow-up multidimensional scaling analysis of item clustering revealed two cohesive dyad clusters, Under.+Accur., and Top.Dev.+Delay, with a clear outlier-S.I.U., followed by Partic. Our first dyad (Under.+Accur.) appeared to well-represent the "linguistic" facet of language. However, the other dyad was not so readily interpretable. The two furthest outliers appeared to represent risk-taking in a social situation. Thus, our cluster analysis raised more questions than it did confirm our hypothesized groupings.

The criterion-related validity analysis was limited to a gross categorization of English language proficiency (NES, LES, FES).

However, six of the seven language attributes in the scale (all except S.I.U.) showed consistent and highly significant relationships with the English proficiency scores.

Regarding efficiency of use of the naturalistic observation protocol, reliability training time was difficult to isolate, but we estimate needing approximately four hours, including supervised coding of actual classrooms. The experience of coding and summarizing results also proved not to be too difficult. The major obstacle to efficiency encountered was the number of observation minutes and sessions required to obtain a stable estimate of student language performance level. We estimated needing a minimum of six 30-minute samples, over two or three weeks, which is beyond that required by any standardized language test.

Several criteria were investigated to help establish technical adequacy of a new naturalistic language observation procedure: (a) interrater reliability, (b) score stability over time, (c) whole-scale internal consistency, (d) meaningful item patterns (e) criterion-related validity, and (f) efficiency of use. Evidence was strongest for interrater reliability and criterion-related validity, and weakest for meaningful item patterns (evidence for construct validity), and score stability over time. Considered together, the evidence encourages development of a naturalistic observation tool, indicates the further development needs, and cautions against the current use of results for important educational decisions.

Especially in need of further development work are the two criteria of score stability and meaningful item patterns. For an observation tool to be useful to schools, observation time must be reduced to two-to-four sessions. The item patterns obtained in this study demand a re-examination of the theorized "facets" of social language. Further work on empirically defined factors is needed to help inform a field which is now dominated by theory with little supportive validation evidence.

### References

- Alvarado, G. S. (1992). Discourse styles and patterns of participation on ESL interactive tasks. *TESOL Quarterly*, 26, 589-593.
- Brown, G., & Yule, G. (1984). *Discourse analysis*. Cambridge: Cambridge University Press.
- Calvert, M. B., & Murray, S. L. (1985). Environmental Communication Profile: An assessment procedure. In C. S. Simon (Ed.), *Communication skills and classroom success: Assessment of language-learning disabled students* (pp. 135-165). Austin, TX: Pro-ed.
- Catheart, R. (1986). Situational differences and the sampling of young children's school language. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 118-139). Rowley, MA: Newbury House.
- Catheart, R. L. (1989). Authentic discourse and the survival English curriculum. *TESOL Quarterly*, 23, 105-126.
- Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.
- Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.
- Cohen, J. A. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review*, 56(1), 18-36.
- Cummins, J. (Ed.). (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles: California State University, Evaluation, Dissemination and Assessment Center.
- Damico, J. S. & Oller, J. W., Jr. (1985). *Spotting language problems*. San Diego: Los Amigos Research Associates.
- Damico, J. S. (1991). *Performance assessment of language minority students*. Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement. Washington, D.C.: U.S. Dept. of Education (OBEMLA), 137-171.

- Day, R. R. (1986). Introduction. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 3-4). Rowley, MA: Newbury House.
- Doughty, C., & Pica, T. (1986). "Information gap" tasks: Do they facilitate second language acquisition? *TESOL Quarterly*, 20, 305-325.
- Enright, S., & McCloskey, M. (1985). Yes, talking!: Organizing the classroom to promote second language acquisition. *TESOL Quarterly*, 19, 431-453.
- Faerch, C., & Kasper, G. (1980). Processes in foreign language learning and communication. *Interlanguage Studies Bulletin*, 5, 47-118.
- Fillmore, L. W. (1985). *Second language learning in children: A proposal model*. National Clearinghouse for Bilingual Education, 33-42.
- Fine, J. (1988). The place of discourse in second language study. In J. Fine (Ed.), *Second language discourse: A textbook of current research* (pp. 1-16). Norwood, NJ: Ablex Publishing Corporation.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley & Sons.
- Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401-435). Rowley, MA: Newbury House.
- Hatch, E. (1992). *Discourse and language education*. New York: Cambridge University Press.
- Hatch, E., Flashner, V., & Hunt, L. (1986). The experience model and language teaching. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 5-22). Rowley, MA: Newbury House.
- Hendricks, D., Scholz, G., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Oral proficiency testing in an intensive English language program. In J. Oiler & K. Perkins (Eds.), *Research in language testing* (pp. 77-90). Rowley, MA: Newbury House.
- Higgs, T., & Ray, C. (1982). The push toward communication. In T. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57-79). Lincolnwood, IL: National Textbook Company.

- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.
- Krashen, S. D. (1985a). *Inquiries and insights: Second language learning*. London: Longman.
- Krashen, S. D. (1985b). *The input hypothesis: Issues and implications*. London: Longman.
- Krashen, S. D. (1985c). *Input in second language acquisition*. Oxford: Pergamon Press.
- Lara-Alecio, R., & Parker, R. (1994). A pedagogical model for transitional English bilingual classrooms, *Bilingual Research Journal*, 18(3&4), 119-133.
- Lieberman, R. J., & Michael, A. (1986). Content relevance and content coverage in tests of grammatical ability. *Journal of Speech and Hearing Disorders*, 51, 71-81.
- Long, M. H. (1981). Input, interaction, and second language acquisition. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 259-278), Annals of the New York Academy of Sciences.
- McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, 49, 34-42.
- Monte Bello Unified School District (1987). Student Oral Language Observation Matrix (SOLOM). Instructional Division, Monte Bello Unified School District, CA.
- Morrow, D. 6. (1985). Prominent characters and events organize narrative understanding. *Journal of Memory and Language*, 24, 304-319.
- Mullen, K. A. (1980). Rater reliability and oral proficiency evaluations. In J. Oller & K. Perkins (Eds.), *Research in language testing* (91-101). Rowley, MA: Newbury House.
- Nix, D. H. (1983). Links: A teaching approach to developmental progress in children's reading comprehension and meta-comprehension. In J. Fine & R. O. Freedle (Eds.), *Developmental issues in discourse* (pp. 103-142). Norwood, NJ: Ablex Publishing Company.
- Oller, J. (1991). *Language testing research.' Lessons applied to LEP students and programs*. Proceedings of the Second National

- Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement. Washington, D.C.: U.S. Dept. Of Education (OBEMLA).
- Rice, M. L., Sell, M. A., & Hadley, P. A. (1992). *The social interactive coding system (SICS): An on-line, clinically relevant descriptive tool*. Language, Speech, and Hearing Services in Schools.
- Sato, J. C. (1986). Conversation and interlanguage development: Rethinking the connection. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 23-45). Rowley, MA: Newbury House.
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.
- Umesh, U. N., Peterson, R. A., & Sauber, M. H. (1989). Interjudge agreement and the maximum value of Kappa. *Educational and Psychological Measurement*, 49, 835-855.
- Van Lier, L. (1988). *The classroom and the language learner*. New York: Longman.
- Verhoeven, L. (1992). Assessment of bilingual proficiency. In L. Verhoeven & J. De Jong (Eds.), *The construct of language proficiency* (pp. 125-135). Philadelphia: John Benjamins Publishing Company.
- Vermeer, A. (1992). Exploring the second language learner lexicon. In L. Verhoeven & J. De Jong (Eds.), *The construct of language proficiency* (pp. 47-173). Philadelphia: John Benjamins Publishing Company.