0670-1

Teachers' and Students' Evaluations of Foreign Language Errors: A Meeting of Minds?

DAVID BIRDSONG and MARGARET ANN KASSEN
University of Florida *University of Texas*

RESEARCH IN THE AREA OF FOREIGN-LANGUAGE error evaluations has typically focused on three groups of evaluators: 1) teachers who are native speakers of the target language; 2) nonnative teachers of the target language; and 3) native speakers not involved in language teaching. The objectives of such studies have been quite diverse. Ervin and Galloway, for example, investigate the issue of uniformity of error judgments among the three groups of evaluators, while Chastain, Ensz, Gynan (17, 18), and Piazza *inter alia* are concerned with these raters' reactions to various types of errors. Other research has examined the array of experiential and attitudinal variables that influence judgments of foreign language.[1]

While error judgments among instructors and native speakers have received ample attention, a significant group of error evaluators has been systematically overlooked: students. Foreign language learners regularly commit errors; they often hear and read other learners' mistakes; they dutifully submit to correction by their instructors. Given this error-intensive environment, we should not be surprised if learners themselves developed a sense of which errors are most serious.

Students' error evaluations take on particular significance when compared to those of teachers. For example, if students and instructors agreed in principle on the seriousness of a given error pattern, remediation might become a more cooperative enterprise. Moreover, initial remedial efforts could be constructively directed at what both parties identified as the biggest blunders. This type of accord could also help students understand their teachers' criteria for assigning grades and their teachers' rationales for correction of errors in oral classroom performance (see Hendrickson; Cathcart & Olsen). Quite possibly, this understanding could contribute to a lowering of individual and group anxiety levels in the foreign-language classroom (Horwitz et al.). Finally, it has been suggested (Bewell & Straw; Gass) that development of metalinguistic skills—among them assessments of interlanguage and target language utterances—may coincide with or even promote advances in overall language ability.

Though agreement among teachers and students in error evaluations may be desirable in the foreign-language classroom, whether such a "meeting of minds" is attainable remains to be seen. Indeed, this idealized consensus would appear somewhat unrealistic, given the idiosyncrasies of individual teachers and students, along with the well-documented response variability inherent in metalinguistic tasks in general, and in error judgments in particular (Birdsong, 6; Chaudron; Galloway). Add to this the disagreements in error evaluation among native and nonnative teachers found by Ervin, and it would appear unreasonable to expect uniformity of judgment, whether within groups or between groups of students and teachers.

Nevertheless, to suppose that certain factors may influence the degree of congruence in error judgments is not unreasonable. Let us take the common situation of English natives who teach French. These teachers have themselves studied (or, in some cases, are presently studying) the target language. One might conjecture that, if teachers and students shared linguistic background and learning experiences, they might also share rationales for error judgment. For

The Modern Language Journal, 72, i (1988)
0026-7902/88/0001/001 $1.50/0

example, both groups could relate target-language errors to comparable errors in their common mother tongue; or they might, by having all committed much the same errors as learners, somehow "empathize" in error evaluations.

Another factor which could influence comparability of students' and teachers' judgments is the length of students' exposure to the target language. As teachers, we serve in many ways as models for our students. By our reactions to errors we may be cultivating in students standards for error evaluation similar to our own. Over time, our students' reactions to errors may begin to coincide with ours.

From the preceding speculations emerge two principal hypotheses tested in the present study. Hypothesis I predicts that student and teacher error judgments will better conform if teachers have the same native language as their students. Hypothesis II predicts that, relative to those of beginners, advanced students' error judgments will better conform to teachers' judgments.

A third hypothesis is a logical derivative of the first two. Hypothesis III predicts that, of all comparisons of student and teacher error evaluations, the greatest conformity will be observed among advanced students and teachers with the same native language as the students. These hypotheses are tested with materials and procedures designed to reduce the number of confounding and/or latent variables that complicate interpretation of results.

In addition to exploring these hypotheses, our investigation offers new perspectives on the congruence of error judgments among native and nonnative teachers of the target language. We also examine the extent to which learners at different levels of instruction agree with one another in error evaluations.

Our discussion of the data centers on differences in results obtained by various statistical procedures. These differences suggest the need to identify two types of agreement in error evaluation: agreement in terms of rating severity and agreement about the seriousness of one item relative to another. We conclude by situating the present study within current research in metalinguistic performance and by discussing the implications of our study for foreign language instruction and curriculum planning.

METHOD

Subjects. Participants in this study included ten French native-speaking (FNS) and ten English native-speaking (ENS) instructors of French at the University of Texas. The FNS instructors, aged 24–49, had from fourteen to twenty-five years' total schooling and three to twenty-four years of teaching experience. The ENS instructors ranged in age from twenty-nine to forty-two years, had had twenty to twenty-six years of schooling, and from one to fourteen years of teaching experience.[2] All ENS instructors had studied French as a foreign language and had attained near-native or native fluency. (Note: the ENS and FNS instructors were recruited as participants in the Birdsong & Thoren study; a subset of their judgment data is reported in the present study and is compared to that of students described below.)

The twenty-one second-language learners (L2L's) in this study were enrolled at the University of Texas in French language courses; twelve of these subjects (hereafter L2L-A) were near the end of their second semester of French instruction; nine of these subjects (hereafter L2L-B) were near the end of their fifth semester of French instruction. Subjects ranged in age from eighteen to thirty; roughly half of the subjects in each group were male, half were females. All students in the sample had received a grade of "B" or higher in previous French courses.

Instrument. The French-language portion of the Birdsong & Thoren questionnaire was used in the present study. The instrument consisted of a set of thirty-two deviant French sentences, each containing one error. All errors had been validated by three experienced French teachers (not subjects in the instructor sample) as being representative of those typically made by beginning students. Errors were equally distributed across four broad categories or types: morphology, syntax, lexicon, and phonology. A complete list of test items is found in the Appendix. Before presentation to subjects, the deviant sentences had been recorded in computer-randomized order by a female American undergraduate student who spoke French with a slight but detectable accent.

Criterion. A precondition for any error judg-

ment task is the establishment of a cr
evaluation. Some traditional criter
comprehensibility (the degree to whic
is an impediment to understanding)
tion (the degree to which an error
negative reaction) are geared to refl
guistic experiences and attitudes of t
natives. Learners, however, and es
ginners, are not likely to understar
these criteria in the manner of inst
native speakers.

The Birdsong & Thoren study
neutral label, *seriousness of error* (cf
son) to its criterion, which was als
in the present study. The term m
tionally interpreted as the likeli
given error will provoke interrupt
rection by a listener in a specified
tion. The criterion of seriousness
for subjects in terms of a realistic c
context in which they were called c
correct the speaker when errors
Significantly, the speech situatio
one that was plausible and under
students and teachers alike. L2L
instructed as follows:

> Imagine that the speaker on the tape
> mate or close friend. She is planning
> and has asked you, knowing that you
> to help prepare her for this experienc
> her errors when the two of you casua
> French. Assuming that the more seri
> the more important it is to correct i
> not possible to correct all her errors
> formal situation, how important is it
> following errors be corrected?

An essentially identical situ
presented to instructors by
Thoren:

> Imagine that you are having an in
> tion in French with a student of th
> has asked you to correct her error
> is planning a trip to a country in wh
> is spoken. Assuming that the more
> is, the more important it is to cor
> is not possible to correct all error
> is it that each of the following er

These instructions were n
(vain) hope of eliminating c
and irritation as factors in sub
there can be little doubt that

ment task is the establishment of a criterion for evaluation. Some traditional criteria such as *comprehensibility* (the degree to which an error is an impediment to understanding) and *irritation* (the degree to which an error provokes a negative reaction) are geared to reflect the linguistic experiences and attitudes of teachers and natives. Learners, however, and especially beginners, are not likely to understand or apply these criteria in the manner of instructors and native speakers.

The Birdsong & Thoren study assigned a neutral label, *seriousness of error* (cf. Hendrickson) to its criterion, which was also employed in the present study. The term may be functionally interpreted as the likelihood that a given error will provoke interruption and correction by a listener in a specified speech situation. The criterion of seriousness was defined for subjects in terms of a realistic conversational context in which they were called on to stop and correct the speaker when errors were made. Significantly, the speech situation chosen was one that was plausible and understandable for students and teachers alike. L2L subjects were instructed as follows:

> Imagine that the speaker on the tape is your roommate or close friend. She is planning a trip to France and has asked you, knowing that you study French, to help prepare her for this experience by correcting her errors when the two of you casually converse in French. Assuming that the more serious an error is, the more important it is to correct it, and that it is not possible to correct all her errors in such an informal situation, how important is it that each of the following errors be corrected?

An essentially identical situation had been presented to instructors by Birdsong and Thoren:

> Imagine that you are having an informal conversation in French with a student of that language who has asked you to correct her errors, as the student is planning a trip to a country in which that language is spoken. Assuming that the more serious the error is, the more important it is to correct it, and that it is not possible to correct all errors, how important is it that each of the following errors be corrected?

These instructions were not devised in the (vain) hope of eliminating comprehensibility and irritation as factors in subjects' judgments: there can be little doubt that one or both factors entered into the evaluation routines of at least some subjects. (For discussion of overlapping criteria, see Gynan, 17; Ludwig.) However, unlike comprehensibility and irritation, a neutral criterion such as "seriousness" can be embraced by all respondent groups. In addition, contextualization of the criterion injects a desirable element of realism into the experimental setting. We also felt that, by giving explicit instructions and contextualization of the criterion, we might reduce the likelihood of subjects' idiosyncratically interpreting ("reading into") the experimental task.

Procedure. Once given the specification of the criterion, raters were told that they would be evaluating errors of a five-point scale, 1 being "'not at all' important to correct," 5 being "'extremely' important to correct," and 2, 3, and 4 having intermediate values.

Before hearing each taped deviant sentence, L2L subjects consulted a written transcript of that item, along with a gloss (in French) of the corresponding non-deviant utterance. To ensure that L2L raters would recognize and focus on the deviant portion of each item, errors of morphology, syntax, and lexicon were italicized on the transcript, while pronunciation errors were underlined. This procedure was intended to attenuate a possible confounding factor of salience or detectability (see discussion, below, and Gynan, 18).

ENS and FNS instructors in the Birdsong & Thoren study were also provided with copies of the correct French utterance corresponding to each deviant test item. When deviant pronunciation was involved, the mispronounced portions were underlined in the non-deviant gloss.

Instructor raters were given approximately ten seconds between items. L2L subjects were given fifteen to twenty seconds between items, as they were obliged to consult two written versions of each stimulus item.

ANALYSIS OF THE DATA

The data analyzed in this study include raw data from the study by Birdsong & Thoren on ENS and FNS instructors' error evaluations, along with our own judgment data elicited from the two groups of learners (L2L-A and L2L-

0670-4

B). For each of the thirty-two deviant items, mean ratings were calculated; these figures are included in the Appendix. Grand means calculated across all items were as follows: FNS = 4.17, std dev = .60; ENS = 3.69, std dev = .70; L2L-B = 3.44, std dev = .87; L2L-A = 2.99, std dev = .87.[3]

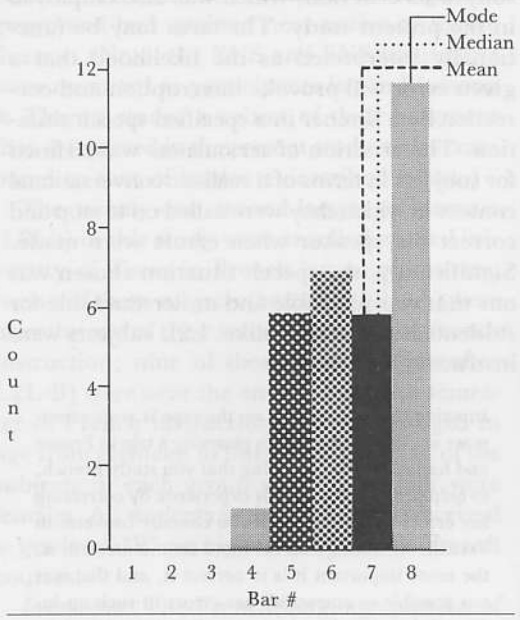
Recalling that a rating of five indicates the judgment "'extremely' important to correct," these grand mean figures point to a tendency among teachers, and especially FNS, to rate items more harshly than students.[4] The severity of native-speaking French teachers' responses is consistent with Ervin's finding that native-speaking teachers of Russian are more critical of errors than nonnative teachers and than natives who are not teachers. Also, among FNS, response means across all items were less dispersed than among other groups; for FNS, most of the means for individual items clustered near the upper end of the 1 to 5 scale. The frequency distribution of item means for FNS and all other groups is detailed in Table I, below. For the four respondent groups, tabular and bar chart representations of the frequency distribution of means for individual test items (test item n = 32). NOTE: Bar charts recapitulate graphically the numerical data in corresponding tables. Each bar corresponds to a range: Bar #1 depicts numbers of items with response means greater than 1.0 and less than or equal to 1.5; Bar #2 has a range of $>1.5 \leq 2.0$; Bar #3: $>2.0 \leq 2.5$, etc. The vertical axis (labelled "Count") in each bar graph indicates how many response means (out of a total of 32) fall within the range of each bar. The vertical axes are *scaled;* thus the heights of the bars are properly proportioned within each chart but should not be used for comparison across charts.

The depiction of response frequency in Table I reveals divergent patterns of evaluations among the four groups of raters. As noted above, one of the more striking features of the tables and bar charts is the high frequency of FNS responses clustering toward numerically high ("'extremely' important to correct") ratings. Among FNS respondents, only one item received a mean rating at or below neutrality (3.0). ENS subjects are not quite so biased toward assigning severe (high numeric) ratings; five items (fewer than sixteen percent) were evaluated at or below neutrality. The L2L-A responses are much less harsh, as fewer than

TABLE I

Bar	From (>)	To (≤)	Count	Percent	
		FNS			
1	1.0	1.5	0	0	
2	1.5	2.0	0	0	
3	2.0	2.5	0	0	
4	2.5	3.0	1	3.125	
5	3.0	3.5	6	18.75	
6	3.5	4.0	7	21.875	
7	4.0	4.5	6	18.75	
8	4.0	5.0	12	37.5	Mode

Bar Chart of FNS



ten percent of item means were above four. In contrast, more than thirty-one percent of L2L-B means fell above four, a clue to the difference in severity of evaluation between the two groups of students.

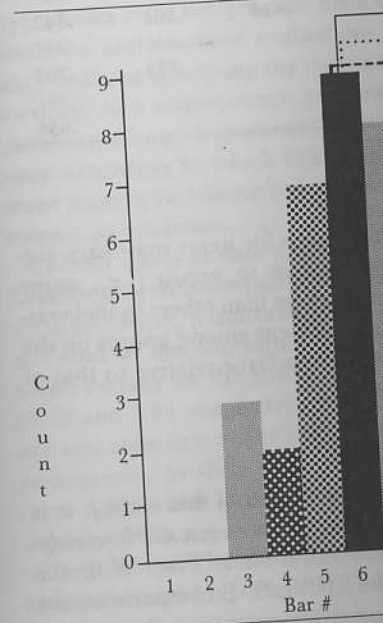
Mean data were then subjected to analyses of variance (ANOVA). A one-way ANOVA performed on all four groups revealed significant differences among the groups: $F(3, 124) = 13.625$, $p < .0001$. To isolate these differences, pairwise F-tests were performed on individual groups. These results are displayed in Table II. F-ratios in pairwise F-tests ($p < .05$ for all comparisons except ENS vs L2L-B).[5]

Data from the paired comparisons above confirm the intergroup disparities suggested in the raw mean and distributional data. Signifi-

TABLE I (continued)

Bar	From (>)	To (≤)	Count	Per
		ENS		
1	1.0	1.5	0	0
2	1.5	2.0	0	0
3	2.0	2.5	3	9
4	2.5	3.0	2	6
5	3.0	3.5	7	21
6	3.5	4.0	9	28
7	4.0	4.5	8	25
8	4.5	5.0	3	

Bar Chart of ENS



cant differences are found for all comparisons except one, ENS greatest magnitude of differen comparison of FNS to L2L-A: speaking teachers stand out as severe in their error judgments learners are. Recalling the ra all four groups, it would appea in error judgments correlates sure to the target language, as severe, followed by nonnative mediate students, and beginn Ludwig 279ff; Ervin; discus

While the data displayed i suggest considerable differenc groups of respondents, an im son remains to be made. Sp

TABLE I (continued)

Bar	From (>)	To (≤)	Count	Percent	
		ENS			
1	1.0	1.5	0	0	
2	1.5	2.0	0	0	
3	2.0	2.5	3	9.375	
4	2.5	3.0	2	6.25	
5	3.0	3.5	7	21.875	
6	3.5	4.0	9	28.125	Mode
7	4.0	4.5	8	25	
8	4.5	5.0	3	9.375	

Bar Chart of ENS

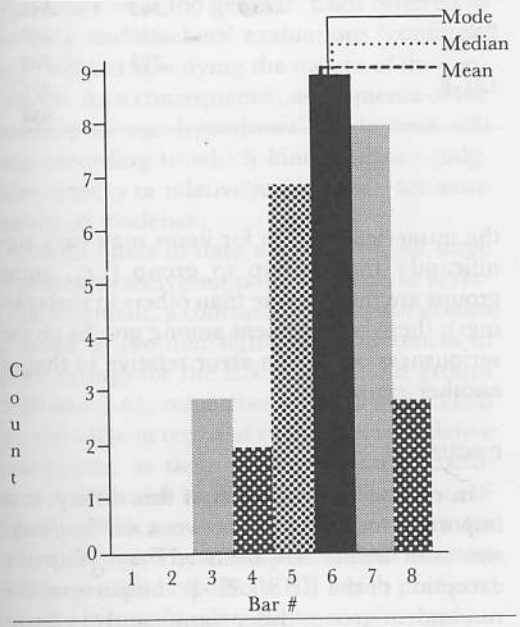
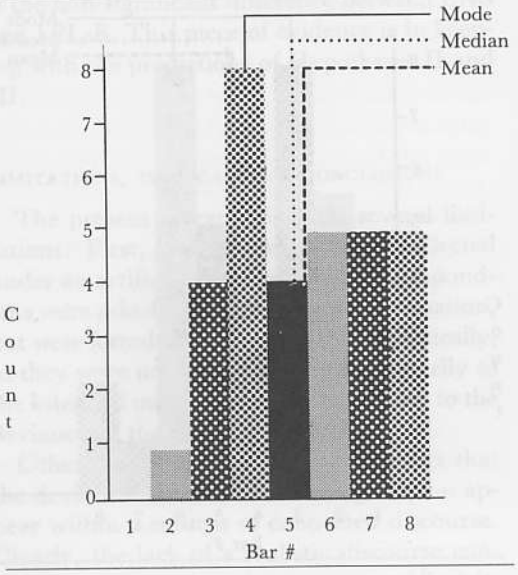


TABLE I (continued)

Bar	From (>)	To (≤)	Count	Percent	
		L2L-B			
1	1.0	1.5	0	0	
2	1.5	2.0	1	3.125	
3	2.0	2.5	4	12.5	
4	2.5	3.0	8	25	Mode
5	3.0	3.5	4	12.5	
6	3.5	4.0	5	15.625	
7	4.0	4.5	5	15.625	
8	4.5	5.0	5	15.625	

Bar Chart of L2L-B



cant differences are found for all between-group comparisons except one, ENS vs L2L-B. The greatest magnitude of difference is seen in the comparison of FNS to L2L-A: French native-speaking teachers stand out as being far more severe in their error judgments than beginning learners are. Recalling the raw mean data for all four groups, it would appear that harshness in error judgments correlates well with exposure to the target language, as natives are most severe, followed by nonnative teachers, intermediate students, and beginning students (see Ludwig 279ff; Ervin; discussion below).

While the data displayed in Tables I and II suggest considerable differences among the four groups of respondents, an important comparison remains to be made. Specifically, we are interested in knowing whether respondents tend to judge a given error as more serious or less serious than another error. Let us imagine, for example, that FNS respondents assign a mean rating of 4.7 to error *x*, while the L2L-A mean rating for *x* is 3.4; for error *y* the ratings are 4.3 and 3.1, respectively; for error *z*, 3.9 and 2.5, and so on. Such a pattern would indicate that beginning students (despite their low numeric mean ratings) agree with French native teachers (despite their high numeric ratings) on the seriousness of some errors relative to others.

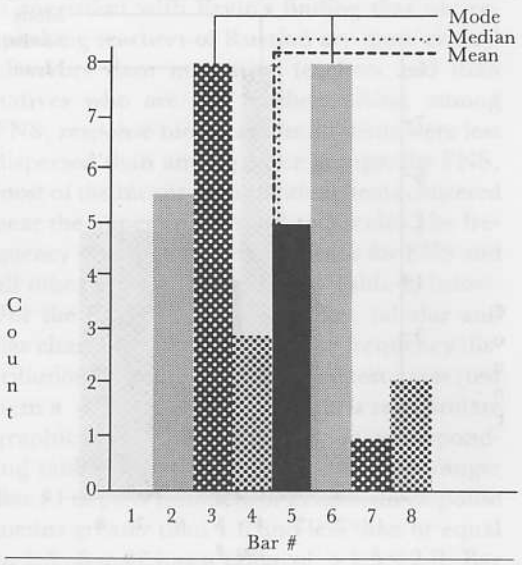
To determine if such an agreement exists, two correlations of group means for individual items were performed. The first was a simple correlation, yielding Pearson product-moment

0670-6

TABLE I (continued)

Bar	From (>)	To (≤)	Count	Percent	
		L2L-A			
1	1.0	1.5	0	0	
2	1.5	2.0	5	15.625	
3	2.0	2.5	8	25	Mode
4	2.5	3.0	3	9.375	
5	3.0	3.5	5	15.625	
6	3.5	4.0	8	25	Mode
7	4.0	4.5	1	3.125	
8	4.5	5.0	2	6.25	

Bar Chart of L2L-A



coefficients of linear correlation (r). The second comparison was a rank-order correlation, using a conservative measure sensitive to ties (Kendall's tau). In the latter procedure, interval scale data (item means ranging from one to five) are converted to rankings (items ordered from most to least serious along a range of one to thirty-two); correlations are then made by comparing the respondent groups' rank-orderings of the thirty-two errors. Table III displays correlation coefficients and Kendall's tau rank-order coefficients, both based on mean ratings for individual items. Correlation coefficients (left figure in each cell) and Kendall's tau corrected for ties (right figure, bold type) for item means. All correlations are significant at $p < .01$.

The results of both the simple correlation and the rank-order correlation indicate that, though

TABLE II

	FNS	ENS	L2L-B	L2L-A
FNS		8.51	15.35	39.89
ENS			1.69 [NS]	12.88
L2L-B				4.33

TABLE III

	FNS	ENS	L2L-B	L2L-A
FNS		.824	.734	.663
		.644	**.565**	**.542**
ENS			.846	.844
			.723	**.702**
L2L-B				.791
				.636

the numerical ratings for items may vary significantly from group to group (i.e., some groups are more severe than others in their ratings), there is agreement among groups on the seriousness of a given error relative to that of another error.

DISCUSSION

In reviewing the results of this survey, it is important to distinguish between *severity of judgment* and *relative seriousness of error*. With the exception of the ENS/L2L-B comparison, the respondent groups differ significantly in terms of severity. Consistent with the findings of Ervin, FNS are harsher than ENS teachers; teachers, with the exception just noted, are harsher than students. A progression of severity of ratings is observed: as subjects approach native speaker status, their responses become more severe.

While all indices of severity (mean ratings, frequency distribution of item means, and analyses of variance) point to differences among error evaluators, rank-order and simple correlational data suggest a great deal of similarity in the groups' assignments of relative seriousness. The latter procedures suggest that the numerical ratings across groups are merely logarithmic variants of a single underlying effect; a focus on the former type of error evaluation data, therefore, may obscure between-group agreements about the relative seriousness of errors.

A glance at the ratings compiled
pendix is revealing. All groups ag
ample, that *Je suis bois un café* (Item 1
serious error than *J'ai aimé ce livre bea*
30), and that the latter is more s
diphthongizing the final vowel in
1). The strength of the concorda
groups in this respect is suggested
robust Pearson and Kendall coef
played in Table III.

Having distinguished between
judgment and relative seriousness
now see that the original statement
hypotheses were too general. Eac
students' and teachers' evaluatio
ing," without specifying the natu
formity. As a consequence, asses
accuracy of our hypotheses' pr
vary according to which kind o
ment severity or relative seriousn
moned as evidence.

If both kinds of data are allo
hypotheses receive support. In t
ity of judgment, a concordance b
is found in the non-significant
mean ratings for the ENS and
(3.69 and 3.44, respectively). E
are also alike in terms of evalua
seriousness, as their product-
cient (.846) and rank-order c
ficient (.723) are marginally
other pairings. These two piece
consistent with Hypothesis III,
that the strongest agreement
tion would be found among ac
and teachers with the same na
the students.[6] Hypothesis II
that advanced students' ratin
conform to teachers' than wo
ginners) receives support from
ity data. The L2L-B mean
cantly higher than the L2L
and, as we have already obse
nificantly different from the
ever, the L2L-B mean is s
than the FNS mean (4.17).
favoring the first hypothesis
cordance of judgment will o
dents and teachers who sh
guage—is found in the re
data. Correlations are mo
ENS instructors and learne
.70) than between FNS ins
ers (r's < .74; taus < .57).

A glance at the ratings compiled in the Appendix is revealing. All groups agree, for example, that *Je suis bois un café* (Item 19) is a more serious error than *J'ai aimé ce livre beaucoup* (Item 30), and that the latter is more serious than diphthongizing the final vowel in *Entrez* (Item 1). The strength of the concordance among groups in this respect is suggested by the very robust Pearson and Kendall coefficients displayed in Table III.

Having distinguished between severity of judgment and relative seriousness of error, we now see that the original statements of the three hypotheses were too general. Each referred to students' and teachers' evaluations "conforming," without specifying the nature of the conformity. As a consequence, assessments of the accuracy of our hypotheses' predictions will vary according to which kind of data—judgment severity or relative seriousness—are summoned as evidence.

If both kinds of data are allowed, all three hypotheses receive support. In terms of severity of judgment, a concordance between groups is found in the non-significant differences in mean ratings for the ENS and L2L-B groups (3.69 and 3.44, respectively). ENS and L2L-B are also alike in terms of evaluations of relative seriousness, as their product-moment coefficient (.846) and rank-order correlation coefficient (.723) are marginally larger than for other pairings. These two pieces of evidence are consistent with Hypothesis III, which predicted that the strongest agreement in error evaluation would be found among advanced learners and teachers with the same native language as the students.[6] Hypothesis II (which predicted that advanced students' ratings would better conform to teachers' than would those of beginners) receives support from judgment severity data. The L2L-B mean (3.44) is significantly higher than the L2L-A mean (2.99); and, as we have already observed, it is not significantly different from the ENS mean. However, the L2L-B mean is significantly lower than the FNS mean (4.17). Finally, evidence favoring the first hypothesis—that better concordance of judgment will obtain between students and teachers who share a native language—is found in the relative seriousness data. Correlations are more robust between ENS instructors and learners (r's > .84; taus > .70) than between FNS instructors and learners (r's < .74; taus < .57).

An altogether different assessment of the hypotheses results if only relative seriousness data are considered. In terms of these data, all groups are essentially the same, with minor differences in robustness of the correlations lending modest support to the spirit of Hypotheses I and III. Given the overall strength of between-group agreements, the three hypotheses —at least as far as evaluations of relative seriousness are concerned—are vacuous.

Discrimination among respondent groups is achieved, on the other hand, with judgment severity data. As we have noted, the key datum is the non-significant difference between ENS and L2L-B. This piece of evidence is in keeping with the predictions of Hypotheses II and III.

LIMITATIONS, IMPLICATIONS, CONCLUSIONS

The present investigation has several limitations. First, our responses were collected under an artificial set of conditions. Respondents were asked to imagine a realistic situation, but were forced to act somewhat unrealistically, as they were not to pay attention primarily to the intended message of the speaker, but to the deviance of the utterance.

Other limitations derive from the fact that the deviant sentences in our study did not appear within a context of connected discourse. Clearly, the lack of a realistic discourse context places respondents in an unnatural linguistic situation and puts constraints on the interpretation of our results. The "discourse context" for a given error in our investigation consisted of the items that preceded it: when rendering error judgments for that item, subjects doubtless reflected on their judgments of earlier items. The implication for interpretation of results is that the severity ratings for a given item in our study should not be viewed as absolute measures or as hard-and-fast evaluations of an isolated linguistic deviance. Instead they should be understood in a relative sense, to some degree a function of ratings of other errors.

Another drawback resulting from lack of discourse context involves the notion of "comprehensibility." Though subjects had at their disposal target-language glosses of the intended message for each item, it is still possible that judgment routines took into account whether an error affected the understanding of its sen-

tence. If all items were couched in a rich discourse context rather than in isolated sentences, some errors might emerge as more comprehensible, other errors less so. The question of ecological validity—whether results similar to those of our study would obtain in real-life situations—is still open.[7]

Despite its limitations, the study invites reflection on the notion of a meeting of minds: that is, the extent to which instructors and students agree—or should agree—in their evaluations of foreign language errors. Our results suggest that in terms of judging the relative seriousness of errors, students and teachers are in agreement. In fact, for our four respondent groups, such a meeting of minds is found to obtain in all six intergroup comparisons, as each group agrees with all the others. This finding might surprise—and encourage—those teachers who perceive students as generally oblivious to errors. It may be argued, however, that our results are not representative of the entire population of foreign-language learners, since the learners in our study had received grades of "B" or better. Does congruence of teachers' and students' assessments of relative seriousness correlate with student achievement?[8] This question merits further research. If such a correlation were found, then the issue of causality would have to be explored: is the agreement of students and teachers on error evaluations attributable to achievement, or is achievement in some way enhanced by sensitivity to errors?

In terms of the harshness or severity of ratings, our four groups largely disagree. Apparently this is a principled, not random, disagreement, as the severity of rating correlates positively with language proficiency. In this finding our study is consonant with those of Odlin (23) and Birdsong (5), who report that with increased language exposure comes a tendency toward rendering metalinguistic judgments that depart significantly from neutrality. As was the case with assessments of relative seriousness, causality questions emerge from the attested correlation of severity and proficiency. Future researchers may consider the possibility that severity reflects confidence in judgments, and that confidence accrues with proficiency (see Yule et al.; cf. Ross). Relative lack of confidence in judgments could be the source of Ervin's finding that, compared to native-speaking teachers, even nonnative teachers are often "cautious" in their evaluations (p. 58).

At the outset of this paper, we discussed the desirability of a concordance in teachers' and students' error assessments. The type of meeting of minds we described then—an agreement about the relative seriousness of errors—appears to exist. The question now becomes whether the other meeting of minds is also desirable. That is, should students become as severe as teachers in their error judgments? Our findings suggest that, desirable or not, such a meeting of minds develops over time, and that this convergence is most obvious among students and teachers with the same native language. The concordance of L2L-B and ENS ratings of severity might be traceable to inherently similar evaluative dispositions deriving from shared native-language background and foreign language learning experience. Alternatively, this phenomenon could result from ENS teachers' success—deliberate or unintentional—in communicating their attitudes toward errors in their students.

On the basis of discussions in curriculum planning meetings and informal surveys of FNS and ENS colleagues, we are persuaded that an accord among students and teachers on error gravity is a goal of language instruction, especially at advanced levels. A common lament goes, "If only we could get students to see how bad some of their errors are!" Our upper-division French students and majors seem to be aware that their instructors frown on certain errors more than others, and that often their grades suffer if such errors are not quickly eradicated. Students express their eagerness to achieve a meeting of minds with instructors—and especially with FNS instructors—in terms of being "on the same wavelength as the teacher," and in terms of knowing "what is expected" of them, "what the pitfalls of the French language are," and "where the teacher is coming from" in error correction and classroom exercises. Unclear, however, is whether students and teachers wish to achieve agreement in rating severity or agreement about the relative seriousness of errors. Before a meeting of minds can be achieved, we must know what kind of agreement we want.

Whatever form the agreement takes, an important precondition must be met: students must be able to *detect* errors before t judge them.[9] Error detection, like oth linguistic abilities (recognition of ambi synonymy, appreciation of puns, m and rhymes, etc.), is not possessed by speakers or even by all bilinguals (E Scholes & Willis; Scribner & Cole). I explicit, analyzed knowledge of langua with a well-developed mental appara trieving and manipulating such k (Bialystok & Ryan; Odlin, 24). The which these requirements are pres individual can be enhanced by cer of linguistic training and experience & Ryan; Olsen et al.; Van Kleeck). not all language-teaching methodol development of error-recognition ski ciple, deductive and intentional should be more successful in this r inductive, incidental approaches.

In this paper we have attempt apart and describe several compo concordance in error evaluation a ers and students. In so doing we

NOTES

[1]Since to chronicle the findings of the would take us beyond the scope of the pres are referred to the review article of Ludw in Omaggio 284ff. We wish to express ou students and teachers at the University who participated in this study. We also t (OSU) and Dan Moors (Univ. of Flori ments on a draft of this paper.

[2]Magnan, Politzer, and Delisle inves effects in error judgments. Their findi ences between children, adolescents, a *among* adults. The variable of sex was exa Delisle, and Ensz, and was found not t significant differences in ratings. In neither years of schooling nor years of correlated systematically with the seve error judgments.

[3]These standard deviation figures a with standard deviations in ratings of i FNS, standard deviations for each ranged from 0 to 1.4, with a mean s .82; for ENS, .52 to 1.65, mean = 1 1.32, mean = .52; for L2L-A, .13 to responding to *a given item*, then, as a g to behave most homogeneously, whi homogeneously. As Table II reveals, persion of means across all items is sponses.

must be able to *detect* errors before they can judge them.[9] Error detection, like other metalinguistic abilities (recognition of ambiguity and synonymy, appreciation of puns, metaphors, and rhymes, etc.), is not possessed by all native speakers or even by all bilinguals (Bertelson; Scholes & Willis; Scribner & Cole). It requires explicit, analyzed knowledge of language, along with a well-developed mental apparatus for retrieving and manipulating such knowledge (Bialystok & Ryan; Odlin, 24). The degree to which these requirements are present in an individual can be enhanced by certain forms of linguistic training and experience (Bialystok & Ryan; Olsen et al.; Van Kleeck). However, not all language-teaching methodologies foster development of error-recognition skills. In principle, deductive and intentional approaches should be more successful in this respect than inductive, incidental approaches.[10]

In this paper we have attempted to tease apart and describe several components of the concordance in error evaluation among teachers and students. In so doing we have had to distinguish between severity and relative seriousness, and among broad respondent categories such as native/nonnative teachers and advanced/beginning students. Fundamental issues, such as the role of error detection in error judgments and the role of metalinguistic abilities in language development, have also been raised. The depth and breadth of the topic of error evaluations should be remembered when instructors and curriculum directors wish aloud that students could "see how bad their errors are." What in fact are we wishing for? Do we want students to be able to detect errors? Do we want them to be as severe as we are in their evaluations of errors, or do we just want them to agree with us about which errors are most serious? Do we really care about students as error detectors and evaluators, so long as they aren't error makers? These multiple goals should be kept conceptually distinct as we consider the appropriateness and feasibility of programmatic attention to learners' foreign-language errors.[11]

NOTES

[1]Since to chronicle the findings of these investigations would take us beyond the scope of the present paper, readers are referred to the review article of Ludwig and discussion in Omaggio 284ff. We wish to express our gratitude to the students and teachers at the University of Texas at Austin who participated in this study. We also thank Terry Odlin (OSU) and Dan Moors (Univ. of Florida) for their comments on a draft of this paper.

[2]Magnan, Politzer, and Delisle investigated age-related effects in error judgments. Their findings suggest differences between children, adolescents, and adults, but not *among* adults. The variable of sex was examined by Magnan, Delisle, and Ensz, and was found not to be responsible for significant differences in ratings. In the present study, neither years of schooling nor years of teaching experience correlated systematically with the severity of respondents' error judgments.

[3]These standard deviation figures are not to be confused with standard deviations in ratings of individual items. For FNS, standard deviations for each of the 32 test items ranged from 0 to 1.4 with a mean standard deviation of .82; for ENS, .52 to 1.65, mean = 1.08; for L2L-B, 0 to 1.32, mean = .52; for L2L-A, .13 to 1.42, mean = 86. In responding to *a given item,* then, as a group L2L-B appeared to behave most homogeneously, while ENS behaved least homogeneously. As Table II reveals, however, the least dispersion of means across all items is observed in FNS responses.

[4]As we see below, the difference between ENS and L2L-B severity ratings is not statistically significant.

[5]Since multiple pairwise comparisons were performed, the significance levels given in Table II may be misleading. More accurate figures are achieved by applying the Bonferroni procedure which (details aside) would meaningfully alter only one p value, that of the L2L-A/L2L-B comparison, changing it from $p<.05$ to $p<.26$.

[6]Differences in robustness, when all correlations are significant, are admittedly not persuasive evidence. This point is discussed further below.

[7]Other criticisms might focus on our small sample size and on the fact that, in the correlational analyses, simple means (without standard deviations) were used, thus obscuring variance figures.

[8]If so, this would parallel the observed tendency for judgments to become more severe as higher levels of proficiency are achieved.

[9]In the discussion of our experimental methodology, we mentioned that we had attempted to eliminate error detection as a possible confounding variable. In fact, error detection presents at least two variables to control for: some errors are more salient or "glaring" than others, and some subjects are more adept than others at spotting errors. In performance on an error detection task, we would intuitively expect these two variables to interact, such that the most salient errors would be detected by the majority of subjects and that some less salient errors would go undetected by the less adept subjects.

An anonymous reviewer raised the possibility that, in spite of our having explicitly indicated errors by underlining, italics, and non-deviant glosses, some errors may not

have been perceived by some subjects. According to the reviewer, this perceptual asymmetry could account for differences in error evaluation among groups, and specifically for the divergence of L2L-B/L2L-A, the assumption being that some or all L2L-A subjects were unable to detect certain errors, while L2L-B's linguistic experience made them better error detectors. (This comment pertained, presumably, to intergroup differences in terms of *severity* of ratings, since strong correlations in judgments of relative seriousness were found for all intergroup comparisons.)

While some errors in our study may have gone undetected despite our procedural safeguards (e.g., *pas des [de] frères* is phonetically similar to *pas de [də] frères*), the majority of our items incorporated errors that were phonetically and orthographically quite distinct from their non-deviant counterparts. Nevertheless, to test whether detection variables for putatively less salient errors may have been a factor in intergroup rating severity differences overall, we analyzed separately those items whose deviant and non-deviant forms were phonetically and/or orthographically similar. These items included all pronunciation errors, one lexical error (*la vacance/les vacances*), one morphology error (*pas des frères/pas de frères*), and one syntax error (*Qu'est-ce que fait ce bruit/Qu'est-ce qui fait ce bruit*). If a variable of error detection were at play, we would expect the intergroup differences in rating severity for these eleven items to be larger than for the thirty-two items overall. This was not the case, as F-ratios for three of the six intergroup comparisons involving the isolated items (FNS/ENS, ENS/L2L-B, and L2L-B/L2L-A) did not achieve statistical significance at $p < .05$ (cf. Table II). Moreover, judgments of relative seriousness for these items did not depart markedly from the overall figures: all intergroup r's were above .7 ($p < .05$) except for the FNS/L2L-B comparison ($r = .59$, $p < .1$; cf. Table III).

While intergroup severity differences cannot be attributed to detection factors, it is of interest to note that the mean severity ratings *within groups* for the eleven putatively less-detectable items (FNS = 3.71; ENS = 3.2; L2L-B = 2.88; L2L-A = 2.33) were significantly lower than the means for the other twenty-one items (FNS = 4.41; ENS = 3.96; L2L-B = 3.74; L2L-A = 3.34): F's (1, 30) all exceeded 8.9, $p < .01$. This difference suggests a possible influence of error salience. It is conceivable that subjects recognized that such errors might be less noticeable within a conversational situation, and lowered their severity ratings accordingly.

Error detection is obviously a crucial element in error evaluation in real-life settings. In some experimental studies of error judgments, however, it may have been overlooked as a factor. For further discussion, see Gynan (18).

[10]Naturally, discussion of the mechanics of various methodologies—whether errors are permitted, expected, encouraged, corrected, etc.—would be an unnecessary digression here. However, we will briefly touch on an even more fundamental pedagogical consideration lurking beneath the issues of error detection and error evaluation. We refer to a question of causality raised earlier in our discussion of the correlation between student-teacher congruence in error judgments and levels of foreign language achievement: what is the relationship between metalinguistic skills and proficient linguistic performance? Though this question has been studied from the perspectives of both first- and second-language development, there is to date no consensus on the answer to this question. Some researchers argue for a reciprocal relationship (as linguistic performance becomes more sophisticated, metalinguistic skills are enhanced, and vice versa); some feel the two are orthogonal, that is, related conceptually but not causally; others argue that metalinguistic skills are merely an artifact of linguistic and cognitive development; still others feel that there are metalinguistic prerequisites to certain linguistic advances. Finally, certain researchers feel that relationships described above are overgeneralizations, and argue that the nature of the relationship between metalinguistic performance and L1/L2 competence is a matter of individual differences in education, linguistic experience, and level of literacy. For elaborations of these positions, see Bewell & Straw, Bialystok & Ryan, Gass, Smith & Tager-Flusberg, Schachter, Scribner & Cole, and Van Kleeck.

[11]Because of the functional overlap of morphology, syntax, and lexicon in grammar, the categorization of some of the errors may at times seem arbitrary. Thus, examples of errors of verb inflection such as *je va* appear properly classified under MORPHOLOGY, while errors in morpho-syntax (e.g., *Qu'est-ce que fait ce bruit?*) may seem misplaced under SYNTAX. In the present study, the four categories merely serve the pragmatic purpose of assuring roughly even distribution of items across broad types of errors. Issues in error typology are discussed by Adiv and Delisle; their findings suggest that patterns of error judgment are not related to broad categories such as those used in this study.

BIBLIOGRAPHY

1. Adiv, Ellen. "Native Speaker Reactions to Errors Made by French Immersion Students" [EDRS, 1982; FL 014 654].
2. Bertelson, Paul. "The Onset of Literacy: Liminal Remarks." *Cognition* 24 (1986, special issue): 1–30.
3. Bewell, Diane V. & Stanley B. Straw. "Metalinguistic Awareness, Cognitive Development, and Language Learning." *Research in the Language Arts: Language and Schooling.* Ed. Victor Froese & Stanley B. Straw. Baltimore: Univ. Park Press, 1983.
4. Bialystok, Ellen & Ellen Bouchard Ryan. "A Metacognitive Framework for the Development of First and Second Language Skills." *Metacognition, Cognition, and Human Performance.* Ed. D. L. Forrest-Pressley, G. E. MacKinnon, & T. Gary Waller. New York: Academic, 1985.
5. Birdsong, David. "Psycholinguistic Perspectives on the Phonology of Frozen Word Order." Diss., Harvard Univ., 1979.
6. ———. "Empirical Impediments to Theories of Second Language Acquisition." Kentucky Foreign Language Conference. Lexington, 26
7. ——— & Krista Thoren. "Variables Affecting Evaluations of Errors in Second Language Performance" [in preparation].
8. Cathcart, Ruth L. & Judy E. W. B. Cohen. "Native and Students' Preferences for Correction in Classroom Conversation Errors." *On TESOL '77.* Ed. Fanselow & Ruth H. Gryme TESOL, 1976: 41–53.
9. Chastain, Kenneth. "Native Speaker and Instructor-Identified Student Errors." *Modern Language Journal* 6
10. Chaudron, Craig. "Research on Methods: A Review of Theory, Meth *Language Learning* 33 (1983): 343
11. Delisle, Helga. "Native Speaker Judgment and Evaluation of Errors in German *Journal* 66 (1982): 39–48.
12. Ensz, Kathleen. "French Attitudes Speech Errors of American Spe *Modern Language Journal* 66 (198
13. Ervin, Gerard. "A Study of the Use of Target Language Communica ployed by American Students o (1978): 6579A–80A (Ohio Stat
14. Galloway, Vicki. "Perceptions of t Efforts of American Students *Language Journal* 64 (1980): 42
15. Gass, Susan. "The Development *TESOL Quarterly* 17 (1983): 2
16. Guntermann, Gail. "A Study of Communicative Effects of Erro *ern Language Journal* 62 (1978)
17. Gynan, Shaw N. "Attitudes to What is the Object of Study?" *nal* 68 (1984): 315–21.
18. ———. "Comprehension, Irrita archies." *Hispania* 68 (1985):
19. Hendrickson, James. "Error Corr guage Teaching: Recent Th Practice." *Modern Language Jou*
20. Horwitz, Elaine K., Michael B. H "Foreign Language Classro *Language Journal* 70 (1986):
21. Ludwig, Jeannette. "Native-S Second-Language Learners'

APPENDIX

Test Sentences by Error Type, w

Item No.	FNS	ENS	Mean
1	3.2	2.2	
2	4.3	4.3	
3	3.8	3.2	

ond Language Acquisition." Kentucky Foreign Language Conference. Lexington, 26 Apr. 1986.
7. ——— & Krista Thoren. "Variables Affecting Teachers' Evaluations of Errors in Second Language Performance" [in preparation].
8. Cathcart, Ruth L. & Judy E. W. B. Olsen. "Teachers' and Students' Preferences for Correction of Classroom Conversation Errors." *On TESOL 76.* Ed. F. Fanselow & Ruth H. Grymes. Washington: TESOL, 1976: 41–53.
9. Chastain, Kenneth. "Native Speaker Reaction to Instructor-Identified Student Second Language Errors." *Modern Language Journal* 64 (1980): 210–15.
10. Chaudron, Craig. "Research on Metalinguistic Judgments: A Review of Theory, Methods, and Results." *Language Learning* 33 (1983): 343–77.
11. Delisle, Helga. "Native Speaker Judgment and the Evaluation of Errors in German." *Modern Language Journal* 66 (1982): 39–48.
12. Ensz, Kathleen. "French Attitudes Toward Typical Speech Errors of American Speakers of French." *Modern Language Journal* 66 (1982): 133–39.
13. Ervin, Gerard. "A Study of the Use and Acceptability of Target Language Communication Strategies Employed by American Students of Russian." *DAI* 38 (1978): 6579A–80A (Ohio State Univ.).
14. Galloway, Vicki. "Perceptions of the Communicative Efforts of American Students of Spanish." *Modern Language Journal* 64 (1980): 429–33.
15. Gass, Susan. "The Development of L2 Intuitions." *TESOL Quarterly* 17 (1983): 273–91.
16. Guntermann, Gail. "A Study of the Frequency and Communicative Effects of Errors in Spanish." *Modern Language Journal* 62 (1978): 249–53.
17. Gynan, Shaw N. "Attitudes toward Interlanguage: What is the Object of Study?" *Modern Language Journal* 68 (1984): 315–21.
18. ———. "Comprehension, Irritation and Error Hierarchies." *Hispania* 68 (1985): 160–65.
19. Hendrickson, James. "Error Correction in Foreign Language Teaching: Recent Theory, Research, and Practice." *Modern Language Journal* 62 (1978): 387–98.
20. Horwitz, Elaine K., Michael B. Horwitz & Joann Cope. "Foreign Language Classroom Anxiety." *Modern Language Journal* 70 (1986): 125–32.
21. Ludwig, Jeannette. "Native-Speaker Judgments of Second-Language Learners' Efforts at Communication: A Review." *Modern Language Journal* 66 (1982): 274–83.
22. Magnan, Sally. "Native Speaker Reaction as a Criterion for Error Correction." *ESL and the Foreign Language Teacher.* Ed. Alan Garfinkel. Skokie, IL: National Textbook, 1982: 30–46.
23. Odlin, Terence. "Part-of-Speech Anomalies in a Second Language." Diss., Univ. of Texas, 1983.
24. ———. "On the Nature and Use of Explicit Knowledge." *International Review of Applied Linguistics* 24 (1986): 123–44.
25. Olsen, Janet L., Bernice Y. L. Wong & Ronald W. Marx. "Linguistic and Metacognitive Aspects of Normally Achieving and Learning Disabled Children's Communication Process." *Learning Disability Quarterly* 6 (1983): 289–304.
26. Omaggio, Alice C. *Teaching Language in Context.* Boston: Heinle, 1986.
27. Piazza, Linda G. "French Tolerance for Grammatical Errors Made by Americans." *Modern Language Journal* 64 (1980): 422–27.
28. Politzer, Robert. "Errors of English Speakers of German as Perceived and Evaluated by German Natives." *Modern Language Journal* 62 (1978): 253–58.
29. Ross, John R. "Where's English?" *Individual Differences in Language Ability and Language Behavior.* Ed. Charles J. Fillmore, Daniel Kempler & William S.-Y. Wang. New York: Academic, 1979: 172–83.
30. Schachter, Jacqueline. "Three Approaches to the Study of Input." *Language Learning* 36 (1986): 211–25.
31. Scholes, Robert & Brenda J. Willis. "Age and Education Factors in Illiteracy." *Developmental Neuropsychology,* in press.
32. Scribner, Sylvia & Michael Cole. *The Psychology of Literacy.* Cambridge: Harvard Univ. Press, 1981.
33. Smith, Carol L. & Helen Tager-Flusberg. "Metalinguistic Awareness and Language Development." *Journal of Experimental Child Psychology* 34 (1982): 449–68.
34. Van Kleeck, Anne. "The Emergence of Linguistic Awareness: A Cognitive Framework." *Merrill-Palmer Quarterly* 28 (1982): 237–65.
35. Yule, George, Jerry L. Yanz, & Atsuko Tsuda. "Investigating Aspects of the Language Learner's Confidence: An Application of the Theory of Signal Detection." *Language Learning* 35 (1985): 473–88.

APPENDIX

Test Sentences by Error Type, with Mean Ratings by Respondent Group[11]

Item No.	Mean Ratings FNS	ENS	L2L-B	L2L-A	
					PHONOLOGY
1	3.2	2.2	2.22	1.58	Entrez [atre$\underline{j}$]
2	4.3	4.3	4.44	4.17	C'est jaune [d$\underline{ʒ}$ on]
3	3.8	3.2	2.33	2.75	Elle prend [pran$\underline{d}$] l'autobus

0670-12

APPENDIX (continued)

Item No.	Mean Ratings FNS	ENS	L2L-B	L2L-A	
4	4.5	3.7	3.56	2.33	C'est la meilleure revue [rəvju]
5	3.3	3.1	3.67	2.17	Il va le vendre [vādrə] (retroflex /r/)
6	3.0	2.3	1.56	1.67	Nous étudions à la bibliothèque [bIbliɔtɛk]
7	3.4	2.5	2.89	2.00	La maison [mɛsõ] est très jolie
8	3.4	3.0	3.00	1.83	J'adore le café filtre [fIltrə]
			LEXICON		
9	4.3	3.9	3.67	3.08	Je suis étudiante à l'universitaire '. . . à l'université'
10	4.4	3.9	3.33	2.58	J'attends pour l'autobus 'J'attends l'autobus'
11	4.9	4.7	4.56	3.92	Nous sommes faim '. . . avons faim'
12	4.6	4.5	4.22	3.58	J'attends toutes les réunions 'J'assiste à . . .'
13	4.0	3.8	4.22	3.83	Il y a beaucoup de traffic '. . . circulation'
14	3.3	3.8	2.89	3.42	Je dois laver mes cheveux tous les jours 'Je dois me laver les cheveux . . .'
15	4.7	4.6	4.67	4.58	J'étudie à la librairie '. . . à la bibliothèque'
16	3.9	3.3	2.44	2.08	Pendant la vacance de Noël, elle a voyagé 'Pendant les vacances . . .'
			MORPHOLOGY		
17	4.5	4.3	4.33	3.58	Je va à l'église le dimanche 'Je vais . . .'
18	3.6	3.5	2.67	2.33	Je suis un professeur 'Je suis professeur'
19	5.0	4.8	5.00	4.58	Je suis bois un café 'Je bois un café'
20	4.9	4.5	4.56	4.00	Nous avons prendu le soleil 'Nous avons pris . . .'
21	4.8	4.1	4.33	3.75	Vous avez sorti chaque soir 'Vous êtes sorti . . .'
22	3.8	3.2	2.22	1.92	Je n'ai pas des frères '. . . pas de frères'
23	4.7	3.8	3.11	3.25	Ma mère exige que je fais mon lit '. . . que je fasse mon lit'
24	4.7	3.6	3.00	2.50	Nous avons une nouveau maison '. . . une nouvelle maison'
			SYNTAX		
25	3.8	3.3	2.89	3.00	C'est le plus intelligent étudiant de la classe 'C'est l'étudiant le plus intelligent de . . .'
26	4.9	4.0	3.67	3.75	Parle-Jean français? 'Jean parle-t-il . . .'
27	4.3	4.4	3.33	3.08	Qu'est-ce que fait ce bruit? 'Qu-est-ce qui fait ce bruit?'
28	4.7	4.2	4.56	2.42	Qu'est-ce qu'il travaille avec? 'Avec quoi est-ce qu'il travaille?'
29	4.6	3.3	3.00	2.17	Il continue parler 'Il continue à parler'
30	3.5	2.6	2.78	2.42	J'ai aimé ce livre beaucoup 'J'ai beaucoup aimé ce livre'
31	3.9	3.8	3.22	3.50	Je ne sais pas qu'est-ce qu'il veut 'Je ne sais pas ce qu'il veut'
32	4.8	4.1	3.78	3.83	Je ne vois pas rien 'Je ne vois rien'

History, Literat and Conversati

LESLIE A. ADELSON
Ohio State University

DURING SPRING QUARTER OF 1986 I visiting assistant professor in the D of German at the University of Irvine. Among the courses I had to teach was an undergraduate c vanced composition and conversat spring quarter approached, I learn "skills" course had as its topical sub the Weimar Republic and the rise fascism. This, I thought, was a bold eign language pedagogy, an assess on the following assumptions: 1) st best in a composition and conver when they are encouraged to spea freely, unencumbered by any oblig grasped any subject other than "t itself" (vocabulary, syntax, gram etc.); 2) most American college st lack extensive knowledge of natio they thus cannot be expected to d substantial prior knowledge of t German history that was to com stantive content of the course.[1] I anticipated was that I could e improving the students' speakin proficiency, or I could teach th Weimar Republic and the ris fascism, but I was not at all certa bine the two effectively.[2]

Composition and conversati by definition classes which pr passivity. How then to eliminat that my students would become sive recipients of historical inf Germany from 1918 to 1945? any, could literature play in t

The course had been concei

The Modern Language Journal, 72, i (19
0026-7902/88/0001/013 $1.50/0