Language Learning 45:3, September 1995, pp. 511–537

Review Article

Error Gravity: A Critical Review of Research Design

Benjamin Rifkin and Felicia D. Roberts
University of Wisconsin-Madison

This paper examines error gravity research design and its theoretical assumptions. Based on an analysis of 28 error gravity investigations (1977–1995), we study several aspects of error gravity research design (including, e.g., the authenticity of language sample), and theoretical constructs (such as the definition of "error"). The study demonstrates that investigators have only skimmed the surface of the process of error evaluation, which is undoubtedly shaped by extralinguistic factors. We conclude that researchers should reconceptualize error gravity research and should reassess earlier studies to confirm or disaffirm their stated outcomes.

The past 20 years have seen the publication of more than 25

Benjamin Rifkin, Department of Slavic Languages and Literatures; Felicia D. Roberts, Department of English.

The authors thank the anonymous reviewers for *Language Learning* as well as Sally Magnan, Charles Scott, Richard Young, and Jane Zuengler for their insightful comments on earlier versions of this manuscript.

Correspondence concerning this article should be addressed to Benjamin Rifkin, Department of Slavic Languages and Literatures, University of Wisconsin, 1432 Van Hise Hall, 1220 Linden Drive, Madison, Wisconsin 53706. Internet: brifkin@facstaff.wisc.edu

studies investigating native speaker reaction to second language (L2) learner error in speech or writing. In continuing efforts to revise and improve foreign language curricula, instructors have sought to determine exactly which learner errors most impede communication with native speakers (NSs) and/or those errors that most irritate NSs. Researchers in many of the more commonly taught foreign and second languages—English, French, German, Spanish—as well as some less commonly taught languages (e.g., Japanese and Russian) have attempted to unravel the complexities of NS response to L2 error, most notably in a series of studies of "error gravity" (cf. reviews in Eisenstein, 1983; Ludwig, 1982). This research has assumed that some linguistic errors are more serious than others in terms of disrupting a NS's comprehension of a nonnative speaker's (NNS's) message and that these error types can be identified. The apparent goal has been to establish hierarchies of L2 error types so that L2 teachers might focus on areas of language production judged by native speakers to be most disruptive to communication. Such hierarchies of error are determined through error evaluation studies, in which NSs are asked to respond in various ways to L2 spoken and written errors.

Discerning the features of L2 production that render communication with NSs more difficult (or conversely, more comfortable) has been a valuable pursuit for L2 researchers concerned with the practical problems of classroom instruction. Error evaluation research and derivative error gravity studies have provided some direction for answering the question "what should be corrected?" in the NNSs' speech and writing in order to render these more acceptable to NSs. However, error investigations' inconsistent findings make it difficult to point confidently in any one direction and proclaim it the route for improving native/nonnative interaction. As Eisenstein (1983) noted, the studies of NS reaction to NNS speech are limited by the difficulty of teasing apart linguistic, social, and contextual variables. Despite her excellent synopsis of the evaluation and reaction research, Eisenstein did not critically review the methods or assumptions driving the investigations of NS evaluation of NNS error. To advance this type of research in order both to equip teachers with an understanding of NS response to L2 students and to better understand NS/NNS interactions, we must examine the assumptions and methods of error evaluation research. Our review focuses on qualitative issues of study conception and design, rather than on problems of quantitative analysis. (For an interesting discussion of relevant quantitative issues in L1 research, see Clark, 1973.)

Review of Previous Error Gravity Research

We demonstrate, by reviewing a wide range of error gravity studies, that investigators have only skimmed the surface of a *process* (evaluation of error) that is undoubtedly shaped by extralinguistic factors. It is clear from the first language (L1) literature that evaluation of speakers varies according to their perceived, as well as real, characteristics (Bradac, 1990; Ryan & Sebastian, 1980; Stewart, Ryan, & Giles, 1985) and the context in which a speech event takes place (Brown, Giles, & Thakerer, 1985). Yet, to a large extent, the study of error gravity has conceived of "error" as a fundamentally linguistic phenomenon.

In the L2 literature, as early as the mid-1970s, Landén and Trankell (1975) demonstrated that "the listener's impression of [an L2] speaker is based on the content of what is said rather than on the occurrence of grammatical and phonological errors" (cited in Johansson, 1978, p. 17). Johansson speculated that content rather than "features of expression" may be a crucial variable in evaluation of L2 speakers, and called for replication of the Landén and Trankell study. To date, no one has answered that call.

By continuing to ignore the variability introduced by phenomena such as stereotyping or content, and by focusing only on linguistic detail, error gravity research has privileged a conception of error as a purely linguistic phenomenon. This focus on linguistic detail has entailed reliance on over-restricted experimental stimuli, leading to studies that tend to isolate language phenomena from situated production.

Table 1
Characteristics of Selected Error Gravity Studies

Study, Date	Language	Modality	*N*	Authenticity	Discourse Universe	Objective or Subjective
Albrechtsen, Henriksen, & Færch, 1980	English	Oral	150	Yes	Paragraph	Subjective
Chastain, 1980	Spanish	Written	27	No	Paragraph	Both
Delisle, 1982	German	Written	193	No	Sentence	Subjective
Earline Schairer, 1992	Spanish	Oral	28	No	Paragraph	Subjective
Ensz, 1982	French	Oral	250	No	Paragraph	Subjective
Ervin, 1979	Russian	Oral	12	Yes	Paragraph	Subjective
Fayer & Krasinski, 1987	English	Oral	128	Yes	Paragraph	Both
Galloway, 1980	Spanish	Oral	32	Yes	Paragraph	Subjective
Guntermann, 1978	Spanish	Oral	30	Yes	Sentence	Objective
Gynan, 1985	Spanish	Oral	186	Yes	Paragraph	Subjective
Hultfors, 1986	English	Written	366	No	Sentence	Subjective
Johansson, 1978	English	Both	Varied	No	Both	Both
Khalil, 1985	English	Written	240	No	Sentence	Subjective
Magnan, 1983	French	Oral	352	No	Sentence	Subjective
Okamura, 1995	Japanese	Oral	80	Yes	Paragraph	Subjective
Olsson, 1977	English	Both	Varied	No	Both	Subjective
Piazza, 1980	French	Both	264	No	Sentence	Subjective
Politzer, 1978	German	Oral	146	No	Sentence	Subjective
Rifkin, 1995	Russian	Oral	75	Yes	Sentence	Subjective

Roberts, 1993	English	Written	65	Yes	Paragraph	Objective
Santos , 1987	English	Written	40	Yes	Paragraph	Subjective
Santos , 1988	English	Written	158	Yes	Paragraph	Subjective
Sheorey, 1986	English	Written	96	No	Sentence	Subjective
Tardif & d'Anglejan, 1981	French	Oral	200	No	Sentence	Subjective
Tomiyana, 1980	English	Written	120	No	Paragraph	Objective
Vann et al., 1984	English	Both	164	Yes	Sentence	Subjective
Varonis & Gass, 1982	English	Oral	Varied	Varied	Both	Subjective

Table 1 summarizes some key characteristics of published error evaluation studies. In reviewing these studies, we have identified critical aspects of research design that may compromise the validity of their conclusions. The aspects of study design highlighted in Table 1 are primarily related to authenticity in terms of both "text" and "task". We focus on those issues in the following three sections. After that discussion, we proceed in the next four with an examination of the more complex problems of error gravity research design: reliance on a loosely defined notion of "irritation", lack of careful item control, focus on overt as opposed to covert errors in learner interlanguage (as discussed by Johansson, 1978, p. 2), and issues related to the recruitment of respondents.

Authenticity

Typically, error gravity researchers have tried to answer questions related to NS response to L2 error by collecting samples of learner language (e.g., student homework assignments or learner response to a picture stimulus) and then selecting some portion or, very rarely, all of the language sample for the experiment. Researchers may take a few sentences from each student's output and compile them for presentation to NS respondents. Alternatively, they might use a student sample to create "simulated learner discourse," based on parts of the learner sample but transformed so that, for example, there is an error in every verb phrase. This simulated learner discourse or any actual learner language performance is then ordered or sequenced, compiled and copied into what we will call an "R-text," a text that NSs *react* to and/or evaluate. This common approach to designing stimuli for error evaluation studies can be improved as a first step in refining error gravity research.

More than half of the error gravity studies reviewed here have used simulated rather than authentic learner discourse, seeking thus to control the number or range of errors that respondents evaluate. For example, Sheorey (1986), who sought to determine whether NS and NNS ESL teachers reacted differently to learner errors in English, collected a learner sample of 97 randomly chosen compositions, from which he created a list of eight major categories of error. He then "constructed twenty sentences, each containing an error representing one of the error types" (p. 307). Chastain (1980), working in Spanish, did not collect a learner sample, but surveyed a number of instructors of Spanish at the intermediate level at a US university and asked them to identify the most serious errors committed by their learners. Chastain then "generated" sentences containing one to three of these errors. Thus, the respondents were actually reacting to language produced by the researcher, in the sense that "instructor-identified" errors were compiled into an R-text. Although Chastain acknowledged that the responses might have been different had the utterances been actual learner language, he provided no evidence that respondents' reactions to the R-texts he created were in any way typical of respondents' reactions to actual learner error.

Guntermann (1978) used a set of 30 oral interviews with Peace Corps volunteers living in Latin America as a learner sample, then created 43 different sentences illustrating some of their errors, but gave no explanation of how the R-text was derived from the interview data. Finally, Piazza (1980) investigated whether response to written error differed from response to spoken error; some of the utterances used as stimuli were actual learner language, some were researcher-modified learner utterances, and some Piazza created. Piazza did not report whether there were any statistical differences in respondent reaction to each of the subtypes of texts.

The understandable need to restrict the number of variables in an experiment has compelled some investigators to create their own R-texts, rather than selecting material from an authentic sample of learner language. Yet, at this essential level of research design, namely selection of stimuli, we are faced with a basic challenge: If we are concerned with NS reaction to NNS error, then we must design studies that come as close as possible to using authentic language. As Table 1 indicates, a number of studies meet that challenge. Varonis and Gass' (1982) study of how American-born native speakers of English reacted to foreign students asking for directions to a train station in a midwestern American city is particularly interesting in this respect. Respondents in this multifaceted study reacted not merely to experimental R-texts, but to learners who stopped them on the street. There is no doubt as to the very real communicative context of these interactions.

Other studies have grappled with the problem of capturing authentic speech using naturalistic methods. Ervin (1979) collected learner language by asking students to tell stories based on pictures; Guntermann (1978) derived spoken language samples from oral interviews. These studies use approaches that, although perhaps somewhat contrived, come closer to representing actual learner language than do studies that enlist readers to record pre-scripted utterances (e.g., Magnan, 1983; Piazza, 1980; Politzer, 1978; Rifkin, 1995[1]). Although script recitation can, for example, capture authenticity of accent (though it is difficult to determine what is a "typical" accent of some group of learners and even more difficult to check the consistency of the accent across a recitation), the recitation of isolated utterances or paragraph-length texts is clearly not natural.

In sum, if the demands of the research question do not require authenticity but there is some theoretical issue at stake, then investigators should clearly indicate and explain the choices that have been made. Two examples of this type of theoretically grounded research are by Tomiyana (1980) and Santos (1987), both of whom attempted to identify the relative effects of global versus local errors (as defined by Burt & Kiparsky, 1974) in paragraph-long written R-texts.

Communicative Context

Many error gravity studies reviewed here (including Chastain, 1980; Delisle, 1982; Hultfors, 1986; Khalil, 1985; Magnan, 1983; Piazza, 1980; Politzer, 1978; Rifkin, 1995; Vann, Meyer, & Lorenz, 1984), did not provide a communicative context in which the R-text could be evaluated. These projects were based on the presentation of an R-text consisting of a series of unconnected single sentences. They required the respondents to assess language for which the discourse universe consisted of a single utterance, an event that simply does not occur in most types of interactions. Probably respondents cannot assess R-texts extracted from communicative contexts in the same way they would assess identical R-texts embedded in such contexts. In many languages, a sentence with a single error may be interpreted in several different ways depending on the linguistic context in which it occurs. For example, without the support of a larger communicative context, the English sentences below may be interpreted differently, as indicated by the alternative interpretations in italics:

> She bought the car tomorrow.
> *She will buy the car tomorrow.*
> *She bought the car yesterday.*

Without the presence of some other temporal indicator or a larger communicative context in which the utterance is set, the semantic contradiction between the adverb *tomorrow* and the tense of the verb *to buy* cannot be unambiguously resolved. This semantic contradiction suggests the importance of setting the R-text in a larger discourse universe. Error gravity studies based on R-texts for which the discourse universe is a single sentence cannot be used to prove the existence of a hierarchy of errors more or less likely to impede communication between NS and NNS. Unfortunately, none of the studies of this type has been linked with a follow-up investigation measuring error gravity in a larger and more authentic communicative context. Without clear contexts within which to evaluate language output, we cannot claim with certainty the source of trouble in any given utterance, and therefore cannot make serious claims as to the relative gravity of one type of error over another.

Some studies, however, have featured R-texts in a communicative context larger than a single sentence (Ensz, 1982; Ervin, 1979; Fayer & Krasinski, 1987; Galloway, 1977, 1980; Gynan, 1985; Roberts, 1993; Santos, 1987, 1988; Tomiyana, 1980; Varonis & Gass, 1982). All of them used actual learner compositions or speech, presented to respondents intact or nearly intact (i.e., in some studies some errors were filtered out by pretesting). For example, Santos (1988) asked learners (native speakers of Chinese or Korean) to write a 350-word composition explaining three aspects of their native cultures most likely to baffle an American encountering them for the first time. Although it may be argued that the content chosen for each R-text (such as home culture) was inappropriate to the audience (of academic professors) and thus affected the nature of their reactions, this is still a good example of naturalistic stimuli, well situated in a larger communicative context. Roberts (1993) did not filter any linguistic errors, but did correct punctuation before presenting the R-text to university faculty.

Objective vs. Subjective Assessment

Many error gravity studies have asked respondents either to perform some operation on the R-text or evaluate it, assessing whether it is "comprehensible" or "irritating" or "native-like", and so on. Some investigators combined these two approaches. The first, performing an operation on the R-text, has been called an "objective" assessment or "operation task"; the second has been called a "subjective" assessment or a "judgment" task (Hultfors, 1986, p. 7; Johansson, 1975, 34ff; Quirk & Svartvik, 1966, p. 23).

Some investigators have designed objective studies that directly ascertain comprehensibility by requiring respondents to perform a task, such as paraphrasing, that demonstrates evidence of their comprehension (Galloway, 1977; Guntermann, 1978). Khalil (1985) attempted an objective comprehension task by having respondents choose the intended meaning of an utterance from a four-option multiple-choice list following the item. He found, however, that the claim of comprehension on a subjective measure was not associated with ability to choose the correct intended meaning. To reduce the risk of multiple meanings for a single utterance, Magnan (1983) and Rifkin (1995) administered a pretest to NS informants asking them to "correct" each of the learner utterances to make sure that each had only one possible interpretation.

Other types of objective assessment include error identification and correction tasks. Both Tomiyana (1980) and Roberts (1993) asked respondents to correct errors encountered in the R-texts, restricting the experimental task to error identification. Chastain (1980) asked respondents to both locate errors in an R-text and then provide a subjective assessment of comprehensibility. He found that the respondents "understood" 90% of the utterances in his study solely on the basis of the respondents' notation that the utterances were comprehensible, rather than from a restatement or paraphrase that could be independently verified.

Clearly, objective approaches provide stronger direct evidence for respondent assessment of what constitutes an error and are also more likely to provide a better measure of comprehensibility. Asking respondents to display understanding of the R-text rather than simply rate it as comprehensible may provide greater insight into the actual intelligibility of learner language. Most error gravity studies, however, have utilized subjective assessments, in which respondents do not participate in the making of meaning as they would in a normal communicative context. Rather, they are asked to sit in judgment of the R-text. In the next section, we further evaluate the use of subjective assessments by scrutinizing the most common subjective measures in error gravity research: irritation and comprehensibility.

Irritation and Comprehension

Published error gravity studies of the early 1980s assumed that classroom teaching should address primarily those errors that caused a failure in communication (e.g., Chastain, 1980;

Delisle, 1982). Piazza (1980) explicitly stated that the goal of error gravity research is to determine those errors which "interfere with comprehensibility and [which] may irritate native speakers" (p. 422). The underlying assumption in many studies (e.g., Johansson, 1978; Magnan, 1983; Politzer, 1978; Tardif & d'Anglejan, 1981) has been that NS irritation with an L2 form is associated with a lack of comprehensibility. In effect, it was assumed, though not tested, that incomprehensible language production would be irritating, though the notion of irritation was never clearly defined. Early researchers in the field (Johansson, 1978; Ludwig, 1982; Piazza, 1980) theorized that this association could be described as an inverse relationship, where low ratings of comprehensibility are equated with high degrees of irritation and vice versa.

On the other hand, a few researchers have considered irritation to be at least partly socially determined—that is, influenced by the expectations and characteristics of the interlocutors. These investigators argued, for example, that a message can be both understandable *and* irritating, or judged to be "foreign" yet still highly comprehensible (Hultfors, 1986; Vann et al., 1984). In other words, comprehensibility is not necessarily linked to linguistic features. This recent work has undermined the earlier assumption that if a message is not understood, it is also irritating somehow.

In the L1 literature, Stewart, Ryan, and Giles (1985) provided an example of a population (American college students) in which "negative affect arousal does not always lead to unfavorable reactions to speakers" (p. 103). In their study, Standard British English, rated as more difficult to understand by Americans, was not downgraded in terms of status attributes. What is difficult to understand may not necessarily be irritating when the evaluators' social attitudes are taken into account.

As Santos (1988) pointed out, it is more likely that irritation includes notions of "acceptability", or the degree to which a NS judges a language sample as meeting implicit or explicit target norms. These norms could be either competence-based or performance-based. Competence-based errors would violate some innate knowledge of the target language that is invariable across speakers, such as SVO word order in English; performance-based errors do not violate the target language's core grammar but violate some standard view of its grammar for a given set of evaluators. For example, "he *don't* eat meat" is not an error in English except in the sense of divergence from standard form. In her study on reactions to NNS writing in English, Santos (1988) found that professors considered double negatives the most irritating of learner errors, but that sentences with such errors were still completely comprehensible. Thus, there is no clear connection between comprehensibility and irritation. Researchers who continue to use subjective assessment methods must take care to ground their concepts and to avoid spuriously assuming a connection between intelligibility of a language sample and the discomfort it may cause a native speaker.

Item Control

The issue of "item control" in error gravity studies ranges from problems as easy to correct as "order effect" to more complex problems of consistency in the classification of error types, the role of markedness, and the selection of global versus local errors.

The simplest example of the lack of item control is the question of item order. Briefly, the order in which items are presented to respondents may well influence response. This phenomenon ("order effect") is well documented in sociological research (cf. Schuman & Presser, 1981). In L2 error gravity research, Fayer and Krasinski's (1987) study on the intelligibility of ESL learner speech showed that the order of presentation of the items indeed had affected the results of that assessment: "a listener's judgment is influenced by the intelligibility of the previous speaker" (p. 318). Ensz (1982) also noted what may be called an order effect in her study. Of the other studies reviewed here, only Hultfors (1986), Magnan (1983), Rifkin (1995), and Varonis and Gass (1982) controlled for order effects. Other studies using isolated utterances or sentences presented items in only one sequence (Delisle, 1982; Guntermann, 1978; Khalil, 1985; Piazza, 1980; Polizer, 1978; Sheorey, 1986; Tardif & d'Anglejan, 1981; Vann et al., 1984). Whether randomized or not, failure to control for an order effect may weaken the conclusions of the majority of error gravity research projects. Controlling for order effect is easily accomplished and future investigations should attempt to do so.

There are other less easily corrected problems of item control. First is the question of error classification. Investigators have tended to establish categories of error types in an effort to generate hierarchies of relative gravity, hypothesizing, for example, that grammatical errors are more "serious" than lexical ones or vice versa. The difficulty, of course, is in the fuzziness of classifying errors. For example, Sheorey (1986) asked respondents to evaluate the R-text "She denied to help me" and classified this utterance as containing a lexical error (confusion of the verb *to deny* with the verb *to refuse*), but it could also be classified as a syntactical error (in the use of the verb *to deny* [someone something].) Delisle (1982) acknowledged the danger of misclassifying errors when she dismissed the use of broad categories "like word order or verb morphology" (p. 43) in error gravity studies. She argued that "errors that fall into [such] categories . . . are in fact, not judged uniformly by the native speaker" (p. 43). In other words, even though linguists may argue for the classification of an error as lexical or grammatical, we cannot be certain of the source of the disruption for the respondent judging the error. One way to ensure that investigators are tapping the source of disruption caused by an error is to ask respondents to perform objective tasks, such as corrections, as mentioned above in our discussion of tasks.

Related to the problem of error classification are the complex issues of markedness and global versus local distinctions. Santos (1987) investigated native speaker respondents' reactions to errors in learner compositions. She hypothesized that:

> there is a directionality of error gravity involving marked and unmarked pairs of forms and structures such that errors reflecting the unmarked-to-marked direction will arouse a greater degree of irritation in native speakers than errors reflecting the marked-to-unmarked direction. (p. 208)

For example, according to Santos, "with *an* great effort" should be more irritating than "such *a* event" because *an* is a marked form in English and is more disruptive when it appears out of its prescribed context. Santos concluded that respondents are indeed more irritated by errors in the unmarked-to-marked direction (1987, p. 215). None of the other studies reviewed here provided for consistent comparison of errors of a particular directionality. Virtually all asked respondents to compare errors of the unmarked-to-marked direction with errors of the marked-to-unmarked direction.

Tomiyana (1980) demonstrated that NS respondents found global errors (as described by Burt & Kiparsky, 1974) much more likely to disturb communication than local errors. The study's findings imply that global errors might also be more irritating than local errors. Accordingly, error gravity studies should consistently compare global errors with global errors, and local errors with local errors, but not mix the two in a single comparison. Only Magnan (1983) tried to provide this kind of item control by limiting item types to words in a particular word list "to minimize effects of word-use frequency, and placed in different positions in the sentences to minimize effects of global versus local errors" (p. 196).

Magnan's attention to "word use frequency" brings out an important point raised by Clark (1973) for L1 psycholinguistic research. Clark argued that those studies of reading aloud L1 word lists that do not randomly select words for the corpus cannot generalize the effect of those words (i.e., latency of response) to other similarly classed, though not identical words (i.e., other nouns and verbs). Clark convincingly described the "language-as-fixed-effect" fallacy underlying most L1 psycholinguistic research

(p. 336). Any response to a corpus of words (or in our case, corpus of errors) may be valid for that corpus and may be replicated with new respondents, but we cannot assume that similar words (or errors) drawn from elsewhere will evoke the same response. Clark argued that researchers must study the individual mean latencies for each item in order to determine any statistical significances, rather than studying the mean latencies of the responses from study participants. Few of the studies reviewed here have provided this kind of statistical analysis (e.g., Rifkin, 1995, for one). In other words, in Clark's view, unless we make a random selection of errors, we simply cannot generalize NS response to those *and* other errors of similar types. Researchers have, thus far, not been able to create R-texts with comparably egregious errors, nor have they been able to make a "random selection of errors".

In a similar vein, in the L2 literature, Bley-Vroman (1983) discussed the "comparative fallacy" in the study of systematicity in L2 production. He argued that, among other pitfalls, researchers must avoid treating language production as a matter of binary choices produced in obligatory contexts. In essence, Bley-Vroman explained that researchers cannot know in advance how many choices or factors learners consider before making a target-language utterance; therefore, it is impossible to know how "to count" learner errors. Many of the error gravity studies suffer thus. Further, Bley-Vroman warned that any study that "preselect[s] data for investigation (such as a study which begins with a corpus of errors) is even more liable to obscure the phenomenon under investigation" (p. 15). Thus, R-texts that use preselected (therefore inauthentic) data for investigation are less likely to provide a basis for valid comparisons.

Overt vs. Covert Errors

Virtually all the studies reviewed here focus on what Johansson (1978) called overt errors—those errors that are apparent at the level of grammatical, lexical, or phonological form. Covert errors, on the other hand, are errors of omission, arising from what Weinreich (1953) called a "poverty of expression" (p. 53) and which "as a rule are not recorded as lack of proficiency" (Johansson, 1978, p. 2). Johansson explained that:

> covert errors can be identified if the total performance of the learner is compared with the performance of native speakers in similar situations . . . or [compared with] the learner's performance in his native [language] and the foreign language in identical situations. (p. 2)

Only Johansson's own study of English and Ervin's (1979) study of Russian made any attempt to identify and assess NS reaction to both covert and overt errors by comparing learner language production in L2 with analogous L1 production. Ervin, for example, asked American learners of Russian to tell a story in English based on picture stimuli and then to tell the same story in Russian. Other studies reviewed here investigated only overt errors. Thus, even if a particular study finds that NSs of French are more sensitive to learner speech errors in grammar than in pronunciation (Piazza, 1980), or that NSs of German are more sensitive to certain types of learner errors in speech (Politzer, 1978) but to other types of errors in learner writing (Delisle, 1982), none of these errors may, in fact, be more significant than the learners' failure to use a broader range of lexical, syntactical, and grammatical items. Thus, the studies may not be identifying the most important problems in NNSs' efforts to communicate in the L2, which relate to a "poverty of expression" rather than to some overtly committed formal error. Future studies which examine and compare the richness of expression across the L1 and L2 may be instructive in this regard.

Recruiting of Respondents

Any survey project must have a plan for selecting respondents reflective of the larger population relevant for the study. The selection of respondents is of critical importance for the generalizability of results. Few of the studies reviewed here provide any information on how they recruited respondents: Was there an advertisement in a newspaper? Were people called randomly on the telephone? Were the respondents all acquaintances of the researchers?

Of all the studies reviewed here, four of the French projects (Ensz, 1982; Magnan, 1983; Piazza, 1980; Tardif & d'Anglejan, 1981) are distinguished by the number and range of the respondents used. Piazza used more than 250 lycée students, whereas Ensz used the same number of respondents selected from a broad range of regions in France, distributed across age groups and professions. Both Ensz and Magnan found that age and gender were critically significant for the assessment of R-texts, suggesting that those studies that did not provide for a range of respondents of different ages may not be generalizable to the larger population with which learners have contact. In addition, Roberts (1993), Santos (1987), and Vann et al. (1984) found that the response to L2 errors by professors was associated with the evaluator's academic field. So, for example, researchers concerned with response to L2 students in university settings should seriously consider randomizing their sample to include a broad range of academic disciplines.

Studies with small samples risk diverse reactions in a subgroup skewing results for the larger sample. Galloway (1980) noted this problem among a group of 8 nonteaching Spaniards living in Spain (a subgroup of the larger sample of 32 respondents). These Spaniards listened to a three-minute recording of an interview with an American student of Spanish. Although 3 of the respondents found the R-text virtually or very incomprehensible, 3 found the same text perfectly or nearly perfectly comprehensible.

Although a sample of convenience is sometimes a necessity, researchers must acknowledge the limitations of such a procedure. For example, Guntermann's (1978) respondents were the members of the families with whom the learners—Peace Corps volunteers—had lived; the respondents recognized the learners' voices on the tape recordings and this familiarity may have affected their reactions to the R-text speech.

Unless we can adhere to the most basic tenets of survey research design, such as adequate sample size to muster statistical power and a clear rationale for the sampling technique, we cannot purport to be engaging in quantitative analyses. Future experimenters must work more concertedly with statisticians in developing research designs that meet the basic criteria for recruitment of participants. Hultfors (1986) was perhaps most comprehensive in terms of using stratified random sampling to obtain a diverse and representative respondent population.

Problematic Conclusions

The research problems described above are symptoms of deeper problems and have led researchers to articulate conclusions that are not merely suspect but sometimes contradictory and occasionally meaningless.

First, much contradiction in the research stems from the different performance contexts of the investigations (written vs. oral presentation of stimuli) and the different tasks required of the judges. Furthermore, studies have lacked consistent methodology in terms of the tasks given to respondents to elicit evaluations. Some researchers have asked participants to respond on Likert-type scales to errors embedded in sentences, whereas others have asked them to choose one sentence as being preferable to another in pair-wise comparisons. It is nearly impossible to get comparable results from a set of studies that rely on such different methods. (For further discussion of this point, see Zuengler, 1980.)

To illustrate the widely varying outcomes of error gravity research, we will focus on a commonly used measure—assessments of comprehensibility of certain error types. Burt (1975) found that word order error was the most serious hindrance to NS/NNS communication in English. In later studies across languages, lexical errors were found to impair comprehensibility more seriously than word order errors (Chastain, 1980; Johansson, 1978; Olsson, 1977; Politzer, 1978). Still others found pronunciation to be more important than semantics or syntax in NS comprehension of spoken L2 production in English (Wigdorsky-Vogelsang, 1978, cited in Delamere, 1986). Varonis and Gass (1982) concluded, however, that there is no hierarchy of grammatical error over pronunciation error (or vice versa) in terms of comprehensibility in English. Other studies of this kind have been done in French, German, Russian and Spanish with mixed results. Finally, two investigations found that the number of errors in a sentence is more important than the type of error in terms of comprehensibility (Albrechtsen, Henriksen, & Færch, 1980; Guntermann, 1978).

Other findings, although not contradictory, do seem to lack practical value. Fayer and Krasinski (1987) reported that NSs' familiarity with a particular type of learner accent increases their comprehension of learners speaking with this accent. Guntermann (1978) concluded that the more errors learners make, the harder it is for NSs to understand them. Khalil (1985) observed that semantic errors were more likely than grammatical errors to reduce intelligibility. This might seem a practical observation, but grammar is that part of language which is essentially predictable. Yet, many researchers have come to the same conclusion as Khalil, including Politzer (1978) in German, Chastain (1981) and Galloway (1980) in Spanish, and Johansson (1978) in English.

Despite the far-reaching problems in design and the confusion of outcomes resulting from those problems, there is still opportunity to move error gravity studies toward producing research significant for L2 teachers and their learners. We now focus on summarizing our critique of error gravity research and suggesting ways for moving the research agenda forward.

Directions for Future Research

One can neither describe nor explain NS evaluations of L2 errors simply on the basis of response to manipulations of linguistic variables: the larger interactive framework within which these interactions take place must be accounted for as well. In turn, we may discover that *error,* conceived thus far as a linguistic event, is actually extralinguistically constrained. This is implied in a conclusion drawn by Albrechtsen et al. (1980): "Whether [or not] an error impairs the intelligibility of the IL [interlanguage] . . . is perhaps not primarily a function of its inherent qualities, but of the context in which it occurs" (p. 393). Although Albrechtsen et al. focused on the linguistic context, we use *context* to refer as well to the performance context (written, oral, formal, informal, etc.) and the social context (setting, participants, purpose, etc.). Error gravity studies have thus far failed to account fully for these variables.

Some researchers have demonstrated that response to error may stem from respondents' personal characteristics (e.g., Hultfors, 1986; Magnan, 1983; Santos, 1988; Sheorey, 1986; Vann et al., 1984; and others), but even these studies give only a glimpse of the true complexity of the social dimension of error evaluation. A richer examination of error response relative to social network, not simply social category, should be a concern. (See, e.g., Milroy, 1980; Milroy & Margrain, 1980, for a discussion of social network theory within a linguistic context.)

A handful of studies have focused on crucial intervening nonlinguistic variables and have shown that NS/NNS differences in belief affect interest in the content of the speech (Galloway, 1980) and that "the stereotype elicited by the non-native voice in some cases may have a more powerful effect on the listener than the effect of error in speech" (Delamere, 1986, p. 87). It has also been shown that increasing the accuracy of NNS speech does not necessarily evoke improved attitudes toward the L2 speaker (Albrechtsen et al., 1980).

To better understand the process of L2 error evaluation, we need to investigate further the nature of the target language norm(s) against which an error is being evaluated. Errors may not be absolute linguistic entities but rather flexible, norm-bounded constructs whose limits shift from judge to judge across speech communities. It may well be that linguistic errors take a "back seat" to larger rhetorical issues once language is assessed in the context of a discourse universe larger than a single utterance. The conclusions drawn thus far from experimental research must be tested in follow-up studies in fully natural, or at least naturalistic, language activities.

To contribute to an understanding of interactions between NS and NNSs, error gravity studies may well need to (a) conceptualize what it means to be a respondent and (b) incorporate into this concept what is known from variation studies, namely that interlocutors behave in relation to their social network and position (Eckert, 1989; Milroy, 1980; Milroy & Margrain, 1980). To be a respondent means to take on the task of judging some output as meeting an internal target norm—that is, some sense of correctness in purely linguistic terms—and/or meeting prescriptive target norms—what one thinks is appropriate given a particular context, interlocutor, and so on. Researchers must be careful to clearly identify contexts and expectations for respondents in order to understand the norms against which they assess a particular text.

Furthermore, researchers must coordinate their goals with their methodology. For example, if researchers are concerned about error gravity in the context of proper classroom discourse, then the respondents should be teachers; if researchers are concerned about the effect of errors in the workplace, then respondents should be coworkers, subordinates, or supervisors. We cannot expect undergraduate student respondents to assume the same perspective on foreign students' errors in English as we might expect L2 teachers or other adults to assume.

Finally, it is the responsibility of error gravity research to define what constitutes an error. So far, error has been explained in terms of L2 linguistic anomalies that disrupt NS comprehension and perhaps cause irritation. However, there are no psycholinguistic studies addressing whether or not any cognitive disruption actually occurs as the result of particular L2 errors; this, too, might prove to be an interesting avenue of inquiry.

Clearly, more attention must be focused on nonlinguistic variables that intervene in the error evaluation process, including not only the respondent's characteristics that affect judgments of L2 linguistic error but also the L2 producer's characteristics. Although we should not think in binary terms of sender and receiver characteristics, nevertheless a more holistic, naturalistic, qualitative approach to examining the full context of interaction is required. Contextual information is critical if one is to realistically gauge NS response to NNS language. However, virtually no attention has been given to setting—neither the social nor the performance context—and this must be incorporated into future studies. Departing radically from previous error gravity research, we could assume that error is socially determined to some extent, and then note whether or not we see evidence of negotiated outcomes as NS evaluators go about deciding what is an error, and indeed a grave error, in a given NNS text. It may be time to incorporate qualitative, ethnographic approaches into the study of error evaluation by NSs. Although qualitative methods are more time-consuming and have their own limitations, they provide the advantage of detailed mapping in a field of inquiry mined with questions of both linguistic and social import: questions that cannot be fully answered from a purely linguistic point of view.

In addition, previously published findings, problematic at best, should be reexamined. Methodological improvements would greatly enhance the value of findings. All of the studies reviewed here feature interesting solutions to the methodological problems discussed, though virtually all of them require replication in a loose sense. To be certain that a particular error hierarchy is valid, we must first check its soundness by conducting a second or third investigation of similar error variables using different design parameters, such as authentic learner discourse to check the findings of a study that used simulated learner discourse or a larger discourse universe to check the findings of a study whose discourse universe consisted of a single sentence, and so forth. Rather than pursuing research into entirely new sets of problems, it may be wise to design studies that could confirm or refute previous findings. Pursuing a better understanding of NS reaction to error in L2 production is clearly important for unraveling the intricacies of NS/NNS interactions. To date, research efforts have been too narrowly focused on the linguistic aspects of error evaluation. We must find new approaches by carefully considering what error and evaluation both mean. If not, we will continue to generate interesting but isolated, and even flawed, conclusions about the complex social interaction between NSs and NNSs of any given language.

Revised version accepted 3 May 1995

Note

[1]The R-text in Rifkin (1995) consisted of utterances selected from oral proficiency interviews and, in this sense, come closer to representing actual learner language than utterances generated by investigators themselves.

References

Albrechtsen, D., Henriksen, B., & Færch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning, 30*, 365–396.

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

Bradac, J. J. (1990). Language attitudes and impression formation. In H. Giles & W. P. Robinson (Eds.), *Handbook of language and social psychology* (pp. 387–412). New York: John Wiley & Sons.

Brown, B. L., Giles, H., and Thakerer, J. N. (1985). Speaker evaluations as a function of speech rate, accent and context. *Language & Communication, 5*, 207–220.

Burt, M. K. (1975). Error analysis in the adult ESL classroom. *TESOL Quarterly, 9*, 53–63.

Burt, M. K., & Kiparsky, C. (1974). Global and local mistakes. In J. H. Schumann & N. Stenson (Eds.), *New frontiers in second language learning* (pp.71–80). Rowley, MA: Newbury House.

Chastain, K. (1980). Native speaker reaction to instructor-identified student second language errors. *Modern Language Journal, 64*, 210–215.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior, 12*, 335–359.

Delamere, T. (1986). *The role of stereotyping in native speaker judgements of*

English as a second language learners' performance. Unpublished doctoral dissertation, Florida State University, Gainesville.

Delisle, H. H. (1982). Native speaker judgment and the evaluation of errors in German. *Modern Language Journal, 66,* 39–48.

Earline Schairer, K. (1992). Native speaker reaction to non-native speech. *Modern Language Journal, 76,* 309–319.

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change, 1,* 245–267.

Eisenstein, M. R. (1983). Native reactions to non-native speech: A review of empirical research. *Studies in Second Language Acquisition, 5,* 160–176.

Ensz, K. Y. (1982). French attitudes toward typical speech errors of American speakers of French. *Modern Language Journal, 66,* 133–139.

Ervin, G. L. (1979). Communicative strategies employed by American students of Russian. *Modern Language Journal, 63,* 329–334.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37,* 313–326.

Galloway, V. B. (1977). *Evaluations of oral communicative competence in Spanish of University of South Carolina students.* Unpublished doctoral dissertation, University of South Carolina, Columbia.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64,* 428–433.

Guntermann, G. (1978). A study of the frequency and communicative effects of errors in Spanish. *Modern Language Journal, 62,* 249–253.

Gynan, N. S. (1985). Comprehension, irritation and error hierarchies. *Hispania, 68,* 160–165.

Hultfors, P. (1986). *Reactions to non-native English: Native English-speakers' assessments of errors in the use of English made by non-native users of the language.* Stockholm: Almquist & Wiksell.

Johansson, S. (1975). *Papers in contrastive linguistics and language testing.* Lund, Sweden: Gleerup.

Johansson, S. (1978). *Studies of error gravity: Native reactions to errors produced by Swedish learners of English.* Göteborg: Acta Universitatis.

Khalil, A. M. (1985). Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of Arab EFL learners. *TESOL Quarterly, 19,* 335–351.

Landén, R., & Trankell, A. (1975). Människor som talar svenska med brytnin [People who speak Swedish with an accent]. In *Invandrarproblem: fem uppsatser om invandrar-och minoritetsproblem fran IMFO-gruppen vid Stockholms Universitet* [The immigration problem: Five essays concerning immigration and minority problems: IMFP group at Stockholm University] (pp. 23–71). Stockholm: Pan/Norstedts.

Ludwig, J. (1982). Native-speaker judgments of second-language learners' efforts at communication: A review. *Modern Language Journal, 66,* 274–283.

Magnan, S. S. (1983). Age and sensitivity to gender in French. *Studies in Second Language Acquisition, 5,* 194–212.

Milroy, L. (1980). *Language and social networks.* Oxford: Basil Blackwell.

Milroy, L., & Margrain, S. (1980). Vernacular language loyalty and social network. *Language and Society, 9,* 43–70.

Okamura, A. (1995). Teachers' and nonteachers' perception of elementary learners' spoken Japanese. *Modern Language Journal, 79,* 29–40.

Olsson, M. (1977). *Intelligibility: An evaluation of some features of English produced by Swedish 14-year-olds.* Göteburg, Sweden: Acta Universitatis.

Piazza, L. G. (1980). French tolerance for grammatical errors made by Americans. *Modern Language Journal, 64,* 422–427.

Politzer, R. L. (1978). Errors of English speakers of German as perceived and evaluated by German natives. *Modern Language Journal, 62,* 253–261.

Quirk, R., & Svartvik, J. (1966). *Investigating linguistic acceptability.* The Hague, The Netherlnds: Mouton & Co.

Rifkin, B. (1995). Error gravity in learners' spoken Russian: A preliminary study. *Modern Language Journal.*

Roberts, F. D. (1993, October). *Effect of perceived language background on evaluations of ESL writing.* Paper presented at the Milwaukee Linguistics Symposium, University of Wisconsin, Milwaukee.

Ryan, E. B., & Sebastian, R. J. (1980). The effects of speech style and social class background on social judgements of speakers. *British Journal of Social & Clinical Psychology,* 19, 229–233.

Santos, T. (1987). Markedness theory and error evaluation: an experimental study. *Applied Linguistics, 8,* 207–218.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22,* 69–90.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* New York: Academic Press.

Sheorey, R. (1986). Error perceptions of native-speaking and non-native-speaking teachers of ESL. *ELT Journal, 40,* 306–312.

Stewart, M. A., Ryan, E. B., & Giles, H. (1985). Accent and social class effects on status and solidarity evaluations. *Personality and Social Psychology Bulletin, 11,* 98–105.

Tardif, C., & d'Anglejan, A. (1981). Les erreurs en francais langue seconde et leurs effets sur la communication orale [Errors in French as a second language and their effects on oral communication]. *Canadian Modern Language Review, 37,* 706–723.

Tomiyana, M. (1980). Grammatical errors communication breakdown. *TESOL Quarterly, 14*, 71–79.

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*, 427–440.

Varonis, E. M., & Gass, S. M. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition, 4*, 114–136.

Weinreich, U. (1953). *Languages in contact: Findings and problems*. New York: Linguistic Circle of New York.

Zuengler, J. (1980). Review of the book *Studies of error gravity: Native reactions to errors produced by Swedish learners of English. Language Learning, 30*, 509–513.