Methodological Problems of Error Evaluation Research: A Review

高 田 智 子

Introduction

Error evaluation is a field of study in which communicative effects of linguistic errors committed by ESL (English as a second language)/EFL (English as a foreign language) learners are investigated. It has drawn attention as second/foreign language teaching has shifted its emphasis from grammatical accuracy to overall communicative competence. Research has been conducted to identify hierarchies of error gravity for the purpose of eliciting pedagogical implications.

Although previous research has presented empirical results and pedagogical implications based on them, it also poses a number of methodological problems (Ellis, 1994). For example, theoretical constructs of error evaluation have not been clearly defined. Methods of measuring the seriousness of errors involve weaknesses, resulting in some conflicting findings in empirical studies.

The present review highlights the background from which these problems arise. It focuses on two issues: theoretical constructs of error evaluation research and the measurements of these constructs. Methodological weaknesses involved in this field of study are also discussed.

Theoretical Constructs of Error Evaluation

Previous studies have investigated communicative effects of interlanguage (IL) errors with multiple criteria (Piazza, 1980; Santos, 1980; Fayer and Krasinski, 1987; Hadden, 1991; Conner-Linton, 1995; Okamura, 1995). These criteria include comprehensibility, acceptability, naturalness, irritation, tolerance, and seriousness.

This multidimensionality of error evaluation is attributed to the nature of practical goals of communication. Major goals of

communication are (1) transmission of information, and (2) establishment of personal/social relationships. For the first goal, foreign language (FL) learners' language product containing errors must be comprehensible. For the second goal, FL learners should not demand psychological burden on the part of interlocutors in interpreting the message senders' intention.

Based on the two goals of communication described above, the criteria for error evaluation may be grouped into two major categories: (1) comprehensibility of IL errors (comprehensibility), and (2) audience's attitudes toward IL errors (acceptability, naturalness, irritation, tolerance, or seriousness). In other words, these two criteria constitute theoretical constructs of error evaluation research.

Comprehensibility is of topmost importance for successful communication. However, even if non-native speakers' (NNSs') error-loaded messages are intelligible, they may irritate their audience (Fayer and Krasinski, 1978, p.315; Piazza, 1980, p.422; Hadden, 1991, p.3). Native speakers receiving the messages may judge the NNSs' intelligence and personality in a negative direction based on the non-standard English they use. This is why Chastain (1980) argues that "although comprehensibility is the most important goal for non-native speakers, in interpersonal communications they must also ultimately be concerned that their language does not lead to a negative reaction on the part of the native speakers with whom they are communicating (p.212)".

Chastain (1980) and Khalil (1985) provide empirical evidence that audience utilize at least two major criteria in error evaluation: comprehensibility of IL messages and attitudes towards IL errors. Chastain (1980) presents proof that there are linguistic deviances whose meanings are comprehensible but whose forms are unacceptable. He studied Spanish native speakers' (NSs') reactions to generated sentences containing errors which are typical of second language (L2) learners of Spanish. His subjects read and rated each sentence as

comprehensible and acceptable, comprehensible but unacceptable, or incomprehensible. He found that 23 of the 48 errors were considered comprehensible but unacceptable by 50 percent or more of the evaluators, concluding that many errors are unacceptable for reasons other than frustration resulting from trying to comprehend what NNSs are attempting to express.

Chastain's findings are supported by Khalil (1985). He investigated the extent to which judgments of comprehensibility and naturalness differed. He selected 20 grammatically and 10 semantically deviant utterances from the errors that occurred in 150 compositions written by EFL Arab freshman students. 240 American undergraduates rated them on a four-point scale of comprehensibility and on a four-point scale of naturalness. The results showed that evaluations of naturalness and that of comprehensibility differed, with utterances generally judged to be more comprehensible than they were natural.

Although comprehensibility has been clearly defined, attitudes towards IL errors pose theoretical confusion. Santos (1988) further classifies it into two constructs: acceptability and irritation. Fayer and Krasinski further classifies irritation into annoyance and distraction. Gynan, (1985) on the other hand, presents evidence that acceptability and irritation are empirically indistinguishable.

The complicated nature of this problem is intensified by obscure definition of related terms. An example is the term "tolerance", which is widely used in the literature. Vann et al. (1984) consider tolerance a criterion to measure acceptability whereas Piazza (1980) assumes that it results from comprehensibility. This terminological confusion betrays precarious nature of theoretical frameworks of error evaluation studies, and eventually threatens the validity of studies in this field.

This review first discusses comprehensibility, followed by acceptability and irritation. Then, a problem related to theoretical and terminological confusion is discussed.

Comprehensibility

Considering that the ultimate objective of communication is conveyance of messages, comprehensibility is a foremost criterion to measure communicative effects of IL errors. According to Johansson (1978a), researchers first examine whether deviant utterances are comprehensible or not. If they are not, they are considered serious errors. If they are fully comprehensible, the next step of research is to examine how they are perceived by addressees. In the following two subsections, the definition of comprehensibility is presented, and two methods of measurement of comprehensibility are described.

Definition

Comprehensibility is the degree to which the interlocutor understands what is said or written (Ludwig, 1982, p.275; Santos, 1988, p.70). The term "comprehensibility" is used interchangeably with "intelligibility" (Piazza, 1980, p.422), but the former is consistently used in the present review.

Measurement

Comprehensibility is measured either on a bipolar scale or on a continuous scale. Fayer and Krasinski (1987) named the former an objective design, and the latter a subjective design.

The objective design involves judges' restating or rewriting of a learner's erroneous utterances into correct forms. If their responses are identical to the message originally intended by the learner, the message is assumed to be comprehended. If not, it means that comprehension is not achieved. In other words, receiver's interpretation of an IL message is measured on a correct/incorrect scale.

Some studies that adopted the objective design include Guntermann (1978), Tomiyana (1980), Takashima (1987), Kobayashi (1992), and Suenobu et al. (1992). Guntermann (1978) asked Spanish NSs to listen to erroneous sentences recorded by English-speaking learners of Spanish and to restate

what they thought the learners intended to say. Tomiyana asked American graduate students to correct two error types in two different texts. Takashima (1987), in his qualitative study, investigated how two NS teachers and one Japanese EFL teacher corrected a free composition written by a Japanese EFL university graduate. Kobayashi (1992) asked English NSs and Japanese NSs to correct all the errors they identified in two English essays written by Japanese students. Suenobu et al. (1992) focused on a phonological aspect of IL, asking NSs of American English to transcribe sentences pronounced with a Japanese accent exactly as they heard them.

The subjective design, on the other hand, involves a continuous point scale with one end being perfectly comprehensible and the other end being not at all comprehensible. Judges are asked to rate error-laden sentences on this scale. Santos (1988) employed a ten-point scale in her investigation of American professors' judgments of two English compositions written by Chinese and Korean students. Khalil (1985) employed a four-point scale to examine American undergraduate students' judgments of 30 deviant utterances written by Arab freshman students.

A strength of subjective designs is quantification of data. However, a major drawback is that receiver's interpretation of an IL message cannot be checked against the real intention of the message sender. Research shows objective designs yield more accurate data because they reveal directly what happens in the readers'/listeners' minds.

Khalil (1985) checked the validity of a subjective design. In addition to applying point scales to examine comprehensibility of deviant sentences, he administered four-option multiple-choice tests to find if judges correctly interpreted the intended meanings of those sentences. The comprehensibility scores rated on scales and the interpretation test scores were compared, and no or little association was found between them. Khalil suggests that intelligibility judgments obtained by subjective designs should

not be assumed to reflect judges' actual understanding of writers' intended meaning.

With the absence of study which shows strong positive correlation between comprehensibility scores obtained by point scales and scores obtained by interpretation tests, it would be safer to follow Khalil's suggestion and to choose an objective design to measure comprehensibility of deviant utterances. As Rifkin and Robers (1995, p.512) point out, objective approaches provide stronger direct evidence for respondent assessment of what constitutes an error, and thus, provide greater insight into the actual comprehensibility of learner language.

Attitudes Towards Interlanguage Errors

A number of terms have been used to refer to attitudes towards interlanguage errors. Khalil (1985) and Kobayashi (1992) calls it "naturalness." Santos (1988) presumes that it is composed of two subconcepts: acceptability and irritation. Hughes and Lascaratou (1982) combine the notion of attitudes toward IL errors with that of comprehensibility, calling them "seriousness" as a whole.

Two of them, acceptability and irritation, are reviewed in the following subsections because they most commonly appear in the literature. It is followed by discussion concerning whether these two notions, acceptability and irritation, are distinguishable.

Relationships between comprehensibility and acceptability

Acceptability is related to comprehensibility. According to Chomsky (1965), comprehensibility and naturalness are necessary conditions for acceptability. He maintains that "the more acceptable sentences are those that are more likely to be produced, more easily understood, less clumsy and in some sense more natural (p.11)."

Although they are thus closely linked to each other, acceptability is a separate notion from comprehensibility.

Whereas comprehensibility is related to the transmission of a message, acceptability is related to listeners'/readers' attitudes toward IL. In Nunnally's (1978) terms, acceptability does not belong to "judgments" but to "sentiments" in the sense that there is no one correct response to a stimulus. Whereas comprehensibility can be measured by an objective design, acceptability cannot because it is "a matter of personal taste (Chaudron, 1983, p.345)."

Multiple dimensions of acceptability

Acceptability is defined as the degree to which a given sample of IL is perceived by NSs to violate language norms (Ludwig, 1982, p.277). To some degree, the farther an error is from the target language (TL) norm, the lower the acceptability. However, the notion of acceptability is not as straightforward as the definition states for two reasons: first, it involves psycho-social aspects of communication (Ludwig, 1982, p.278), and second, acceptability ratings are relative to context and situation (Enkvist, 1973, p.20).

The psycho-social aspects involved in acceptability has been documented by some researchers. Chastain (1980, p.214) suggests that negative acceptability ratings of particular grammatical errors may be attributed to a tendency to associate negative values with some dialects. Santos (1988, p.84) found that a double negative error ranked the least acceptable and the most irritating although it was completely comprehensible. She speculates that this result was caused by respondents' attitudes toward less educated NSs.

Another example that indicates social aspects of acceptability is related to the possibility that NSs commit the same errors that L2 learners do. Burt and Kiparsky (1975) and Piazza (1980) maintain that certain errors which are committed by NSs are more acceptable than errors typical of NNSs.

The functional relativity of acceptability of errors was described as one of the most important concepts by Enkvist (1973).

He maintains that a given expression may be acceptable in one context and situation, and unacceptable in another. There are two studies that specified situations in which errors occur and examined their acceptability. Vann et al. (1984) and Santos (1988) selected an academic setting and investigated effects of errors that occur in it. Vann et al's (1984) study was motivated by the "chronic dilemma" college ESL instructors face in dealing with structural and mechanical errors. Santos (1988) asked NS professors to read and rate two compositions written by ESL learners as pieces of "academic writing." These two studies, because of specification of context in which errors occur, provide more practical pedagogical implications than other studies that examined acceptability in a general sense.

Measurement of acceptability

Two major techniques for the measurement of acceptability are 1) an operation technique, and 2) a direct question technique. The former is employed by researchers who attempt to draw a line between fully grammatical and fully ungrammatical language (Quirk and Svartvik, 1966). The latter is employed mainly by L2 researchers who seek to examine communicative effects of NNS ungrammatical utterances on the listeners'/readers' side. Whereas the former has been used in order to formulate explicit normative rules of a language by L1 researchers (Greenbaum, 1975, p.167), the latter has been used in order to examine NSs' or language teachers' latitude to tolerate linguistic inappropriateness. The present review describes the direct question technique since its scope is evaluation of errors L2 learners commit.

The direct question technique employs either a paired sentence model (Guntermannn, 1978; Politzer, 1978; Magnan, 1981; Rifkin, 1995) or a point scale system (Vann et al., 1984; Khalil, 1985; Santos, 1988; Teng, 1990). The former involves numerous pairs of sentences loaded with different types of errors, out of which judges select one that is the more acceptable. The

latter adopts three-, or over three-point scales with one end being the most acceptable and with the other end being the least acceptable. Judges are asked to evaluate the degree of acceptability on these scales.

Magnan (1981) contends that the paired sentence model offers an efficient framework because "it presents judges with a forced, binary choice to avoid non-committal response (p.45)," and because it can control many extraneous variables. She designed an instrument containing 105 pairs of French sentences recorded by an American student, in which each sentence contains a single error, representative of one of 15 error types. Rifkin (1995) follows Magnan, constructing an instrument with 75 pairs of Russian sentences recorded by American learners of Russian.

The use of a point scale system to measure acceptability has been identified in four studies in the literature. Khalil (1985) adopted four-point scales, Vann et al. (1984) and Teng (1990), five-point scales, and Santos (1988), ten-point scales. They do not argue the selection of the point scale design over the paired sentence design. My speculation is that a strength of the point scale system is the availability of a wider latitude from which judges choose one which is closest to their reactions. The paired sentence model, on the contrary, does not allow judges to express their reactions in cases when both sentences in a pair are acceptable or unacceptable to the same degree.

Another strength of the point scale system is its applicability in contextualized test instruments. In the paired sentence model, each sentence is separated from context, limiting generalization of the results in realistic communicative settings. However, point scale model can be adopted in a larger discourse universe. Santos (1988) took this advantage, examining acceptability of errors in two 400-word compositions written by a Chinese-speaking student and by a Korean-speaking student.

Irritation

A number of researchers treat irritation as one of the criteria for examining communicative effects of learner language (Johansson, 1978a and 1978b; Piazza, 1980; Fayer and Krasinski, 1987; Hadden, 1991; Okamura 1995). Gynan (1984) offers a definition of irritation as opposed to that of acceptability. According to Gynan, irritation is an affective language attitude, whereas acceptability is an evaluative language attitude. In other words, irritation is "a learned predisposition to state consistently that a given object makes the listener feel good or bad," whereas acceptability is "a learned predisposition to locate consistently a given object on a dimension labeled good/bad (p.316)."

Irritation and acceptability

There are two positions regarding the relationships between irritation and acceptability: 1) irritation and acceptability are distinguishable, and 2) irritation and acceptability are indistinguishable. Santos (1988) is a researcher who takes the first position. She examined comprehensibility, acceptability, and irritation of compositions written by ESL college students.

Gynan (1984) takes the second position. Citing Gynan (1983) and social psychologists Ostrom (1969) and Fishbein and Ajzen (1969), he reports that these two notions are empirically indistinguishable. He notes that, "since affective and evaluative attitudes are not empirically distinguishable, the study of irritation and the study of acceptability, though terminologically separate, have as their object of study the same phenomenon (p.316)."

Gynan's position is supported by some studies which use "irritation" and "acceptability" interchangeably, referring to receivers' reactions to NNS interlanguage errors with two different terms. An example is Rifkin's (1995) study of error gravity in learners' spoken Russian. He had his subjects listen

to 72 pairs of error-laden sentences and asked, "Which utterances is more acceptable [italics added] to you?" In the abstract of his study, however, he says, "Respondents ... were asked ... to select which sentence they found to be less irritating [italics added]."

A similar example is found in Johansson (1978a). He assumes that errors which irritate NS audience are unacceptable and that those which do not are acceptable. Browning (1982) takes the same position. However, there is no justification that affective attitude towards learner errors serve as a measurement of evaluative attitudes. Besides, as Zuengler (1980, p.510) points out, his terminology lacks consistency. Using a five-point scale ranging from "native-like" to "very foreign" in two studies, he discusses the ratings of one study as irritation, and the ratings of the other study as acceptability.

The discussion described above leads us to assume that audience's attitudes toward IL errors may be composed of multiple constructs, but that clear distinction among them has not been established. This poses a methodological weakness. Due to terminological confusion, what is measured and what is supposed to be measured may not be isomorphic. There is no justification that acceptability and irritation serve as separate criteria to measure listeners'/readers' attitudes toward IL errors. Based on the studies to date, it would be reasonable to suggest that attitudes toward IL errors be treated as one notion, not being divided into smaller units of concepts.

Conclusion

There is agreement in the literature that at least two theoretical constructs are involved in error evaluation: comprehensibility of erroneous messages and audience's attitudes toward them. Comprehensibility of messages is of the utmost importance for successful interpersonal communication, but evaluative and affective reactions of message receivers are equally crucial since one of the goals of communication is to establish social/personal relationships.

Comprehensibility is the degree to which the interlocutor understands what is said or written (Ludwig, 1982; Santos, 1988). It is measured either by objective design or a subjective design. The objective design involves judges' restating or rewriting of a learner's erroneous utterances into correct forms. In this design, comprehensibility is measured on a bipolar scale, with deviant utterances being perceived to be either comprehensible or incomprehensible. A strength of this design is that judges' assessment of what constitutes an error is directly reflected in obtained data. The subjective design, on the other hand, involves a point scale on which judges assess the degree of comprehensibility. Although this design enables researchers to obtain quantitative data, there is no evidence that the data obtained by the subjective design reflect judges' understanding of a message sender's intended meaning. Indeed, Khalil (1985) checked the validity of the subjective design and found that there was no or little correlation between judges' comprehensibility scores and their interpretation of intended messages.

Thus, error evaluation studies that investigate comprehensibility of learners' erroneous utterances should be viewed with caution if they adopt point-scales for measurement. The quantitative data obtained by the subjective design may appear to provide objective data. However, they can be spurious because what judges understand an ESL/EFL learner writes/says may not coincide with what the learner intends to write/say.

Attitudes toward IL errors are referred to with a number of terms: acceptability, irritation, naturalness, and seriousness. The most common terms in the literature are acceptability and irritation. There are two views regarding the relationships between acceptability and irritation. Santos (1988) investigated professors' reactions to academic writing of NNS students assuming that acceptability and irritation constitute different constructs. Gynan (1985), on the other hand, presents empirical evidence that acceptability and irritation are intertwined. Gynan's position is accidentally supported by other researchers

who used the terms acceptability and irritation interchangeably.

To date, theoretical constructs of audience's attitudes toward IL errors have not been defined with sufficient rigor. Therefore, we need to exercise caution in interpreting the studies that examine audience's evaluative and/or affective responses towards learner errors. Even though a researcher attempts to measure more than one aspect of readers'/listeners' reactions to IL errors, it is possible that he or she is looking at only one dimension. Future studies need to recognize this precarious nature of theoretical constructs of error evaluation research and operationally define what aspects of learner errors it attempts to measure.

References

Browning, G. L. (1982). A listener-based hierarchy of acceptability for non-native features of oral English. Unpublished doctoral dissertation, University of California Los Angeles.

Burt, M. K., & Kiparsky, C. (1975). Global and local mistakes. In J. Schumann, & N. Stenson (Eds.), New frontiers in second language learning (pp. 71-80). Newbury House Publishers, Inc.

Chastain, K. (1980). Native speaker reaction to instructor-identified student second-language errors. Modern language journal, 64 (2), 210-215.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. Language learning, 33 (3), 343-377.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, Massachusetts: The MIT Press.

Conner-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. World Englishes, 14 (1), 99-115.

Ellis, R. (1994). The study of second language acquisition. Oxford University Press.

Enkvist, N. E. (1973). Should we count errors or measure success? In J. Svartvik (Ed.), Errata: Papers in error analysis (pp. 16-23). Lund, Sweden: CWK Gleerup.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. Language learning, 37 (3), 313-326.

Greenbaum, S. (1975). Language variation and acceptability. TESOL Quarterly, 9 (2), 165-172.

Guntermann, G. (1978). A study of the frequency and communicative effects of errors in Spanish. Modern language journal, 62 (5-6), 249-253.

Gynan, S. N. (1984). Attitudes toward interlanguage: What is the object of study? Modern language journal, 68 (4), 315-321.

Gynan, S. N. (1985). Comprehension, irritation and error hierarchies. Hispania 68, 160-65.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. Language learning, 41 (1), 1-24.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. English language teaching journal, 36 (3), 175-182.

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

Johansson, S. (1978b). Problems in studying the communicative effect of learner's errors. Studies in second language acquisition, 1, 41-52.

Khalil, A. (1985). Communicative error evaluation: Native speaker's evaluation and interpretation of writer errors of oral EFL learners. TESOL Quarterly, 19 (2), 335-351.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. TESOL Quarterly 26 (1), 81-112.

Ludwig, J. (1982). Native-speaker judgments of second-language learners' efforts at communication: A review. Modern language journal, 66 (3), 274-283.

Magnan, S. S. (1981). Native and non-native reaction to grammatical error in French. Unpublished doctoral dissertation, Indiana University.

Nunnally, J. C. (1978). Psychometric theory. New York: McGraw-Hill.

Okamura, A. (1995). Teachers' and nonteachers' perception of elementary learners' spoken Japanese. The Modern language journal, 79 (1), 29-40.

Piazza, L. G. (1980). French tolerance for grammatical errors made by Americans. Modern language journal, 64 (4), 422-427.

Politzer, R. (1978). Errors of English speakers of German as perceived and evaluated by German natives. The Modern language journal, 62 (5-6), 253-261.

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

Rifkin, B. (1995). Error gravity in learners' spoken Russian: A preliminary study. Modern language journal, 79 (4), 477-490.

Rifkin, B., & Roberts, F. D. (1995). Review article: Error gravity: A critical review of research design. Language learning, 45 (3), 511-527.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. TESOL Quarterly, 22 (1), 69-90.

Suenobu, M, Kanzaki, K, & Yamane, S. (1992). An experimental study of intelligibility of Japanese English. International review of applied linguistics [IRAL], 30, 146-156.

Svartvik, J. (1973). Introduction. In J. Svartvik (Ed.), Errata: Papers in error analysis (pp. 7-15). Lund, Sweden: CWK Gleerup.

Takashima, H. (1987). To what extent are non-native speakers qualified to correct free composition? - A case study. British journal of language teaching, 25 (1), 43-48.

Teng, W. (1990). Typical errors of Taiwanese students of English as a foreign language (EFL) as perceived by native English speakers and EFL teachers. Unpublished doctoral dissertation, The University of Texas at Austin.

Tomiyana, M. (1980). Grammatical errors communication breakdown. TESOL Quarterly, 14 (1), 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 (3), 427-440.

Zuengler, J. (1980). Review of S. Johansson's studies of error gravity. Language learning, 30, (2), 509-513.