Articles

Exploring Gaps in Teacher and Student EFL Error Evaluation

Sean Mahoney
Fukushima University

Most studies in EFL error gravity to date have not elicited evaluative decisions from the learners themselves. This paper addresses university-level students' (n = 183) and their native- and nonnative-speaking teachers' (n = 5) perceptions of written errors by comparing their marking of a contextualized, in-class dictation quiz. Results indicate that while student and teacher rankings of questions according to difficulty correlated very well, both native- and nonnative-speaking teachers awarded fewer marks than their students and made very similar judgements about similar error types. The most significant teacher-student gaps were apparent when judgements of error involved phonic distinction, intelligibility, and context breakdown. Student evaluators provided written comments on marking strategies used, which served as a basis for further analysis of their awareness of error gravity. It was found, somewhat paradoxically, that students who had indicated sensitivity to degrees of error did not consistently produce more teacher-like evaluations than those who had not.

本研究は、英語の誤りについての重み付け（error gravity）を調査したもので、大学の授業中に行ったディクテーション活動中に見られた誤りに対する評価の比較である。参加者のネイティブ・スピーカーまたはノンネイティブ・スピーカーの教師（n = 5人）と、英語学習者（n = 183人）がどの誤りに注目するのか、また見つけた誤りに対する評価の度合いを比較した。出された問題の難しさによる順位付けがほぼ同じであった一方、教師の評価はすべての問題において学生より低く、類似したエラーの判断もほぼ一致した。教師と学習者の間で有意差のあったものは、総合理解度、音の区別、文脈上の評価であった。英語学習者からの自分の評価基準についての自由記述的なコメントの分析からは、「誤りの重み付けに対する意識」のある学習者の評価がそうではない学習者より必ずしも教師の評価に近くはないことが示された。

JALT Journal, Vol. 33, No. 2, November 2011

This paper is an inquiry motivated by class activities in which students without peer-editor training evaluated each other's English abilities, namely written English output. From the outset, it became clear that student and teacher notions of what constituted serious or negligible errors often varied greatly. A number of studies in error gravity (e.g., Johansson, 1978; McCretton & Rider, 1993; Roberts & Cimasko, 2008) have attempted to explain not only the types of interlanguage errors made, but also the evaluative gaps between various groups of language users, focussing chiefly on those between (a) native and nonnative speakers, (b) native- and nonnative-speaking evaluators, and (c) teaching and nonteaching judges. The present study will examine both error types and evaluator gaps, in particular those between a group of English teachers in Japan and their student L2 learners. It is hoped that studies of these gaps may provide a basis for a more informed, deliberate, and efficient training of L2 peer-editors.

Attempts to examine the process of evaluating L2 learners' output, and efforts to render the results scientific, come fraught with complexity. As Ellis (2008) notes in his description of error gravity studies to date, most focus on producing hierarchies or scales for error judgements and tend to ignore communicative contexts, especially when they present judges with isolated examples of errors. He adds that any findings based on such questionable project design can result in "spurious" conclusions despite the "appearance of rigour given by the use of descriptive and qualitative statistics" (p. 60).

Rifkin and Roberts (1995), building on the work of Schairer (1992), contrast the conclusions of 28 studies in error gravity and reveal both "inconsistent findings" and striking contradictions in evaluation "which make it difficult to point confidently in any one direction and proclaim it the route for improving native/nonnative interaction" (p. 512). Rifkin (1995) sheds light on still more complexities with a critique of investigations conducted on L2 learners studying an array of languages, the results of which suggest that priorities in determining error gravity vary with different cultural norms and native speaker expectations. Santos (1988) examines the problems involved in the very act of categorizing error types and particularly criticises the tendency among scholars to edit raw learner-produced data. He remarks that while "artificially prepared passages allow for maximum control of the variables, they sacrifice the natural quality of unaltered connected discourse" (p. 74). The experiment discussed in this paper, while admittedly prepared and variable controlled, will attempt to preserve data authenticity and the connectedness or context of discourse, with a view toward producing valid claims about evaluative tendencies among its three groups of participants:

native-speaking teachers (NSTs), nonnative-speaking teachers (NNSTs), and university-level L2 learners.

Literature Review

Findings from research in error gravity have revealed tendencies by native-speaking (NS) and nonnative-speaking (NNS) users of English. Beginning in the mid-1970s, Johansson conducted a variety of experiments which contrasted error judgements of native- and nonnative-speaking teachers of English, finding that while NSTs were less tolerant than NNSTs of learners' written mistakes, they appeared more flexible in regard to orally produced errors (1978, pp. 121-23). A debate has since continued over whether native or nonnative speakers of English exhibit significant patterns in their corrections of students' work. Schairer (1992) summarizes studies that have similar concerns, generally producing hierarchies of error types, each bound to L2 speakers of French, German, and Russian. Vann, Meyer, and Frederick (1984) produced what is arguably the best hierarchy of English error types for the English language, one that still serves, for example, as a guideline for ESL tutors at Iowa State University.

Unfortunately, the majority of thorough investigations into error gravity have in some way tampered with original learner output, either by having evaluators mark researcher-contrived but purportedly typical learner sentences (e.g., Davies, 1983; Magnan as cited in Rifkin, 1995), or by the editing of learners' written or spoken data in order to facilitate empirical analysis, a process that produces what Rifkin and Roberts (1995) dub an "'R-text,' a text that NSs *react* to and/or evaluate" (p. 516). Some (e.g., Hughes & Lascaratou, 1982; Khalil, 1985; Rifkin, 1995; Sheorey, 1986) reduce the number of errors to one per R-text. But this practice could limit the accuracy and authority of individual and overall evaluations of the interlanguage samples. That is, too much information about learners' abilities may in such cases be lost or at least diminished depending upon which error or errors have been amended and which have been allowed to stand. Compare the levels of English competence as conveyed in the following original and edited sentences:

- Now I not entered any club. (original, from student journal)
- I not in any club now. (edited to one error)

Reductions in the number of errors in the above example prevent an accurate, third-party evaluation of the learner's ability. Editing has here cre-

ated problems at several levels. First, an evaluator with access only to the edited version may judge the absence of the verb *am* lightly, as a mere slip of the tongue (or pen) that does not reflect the author's true ability. The original sentence however lays bare several more complex and fundamental problems. Second, it can be argued that the editing of one or more words in a sentence, no matter how carefully done, greatly affects reader expectations. Third, unless the researcher interviews the learner, the latter's original intended meaning may be misrepresented. Fourth, the recasting of syntax and lexis as above may be deemed so extensive as to have created an entirely new sentence, bereft of any link to the learner. Last, and more generally, any findings based on filtered data cannot necessarily be applied to actual communicative events.

Brief abstracts and thorough categorizations of works based on both edited and unedited data appear in Rifkin and Roberts (1995). Chief among those that employ edited output include Guntermann (oral data, 1978), Piazza (written and oral data, 1980), Rifkin (written data, 1995), and Sheorey (written data, 1986). Yet notable, early exceptions to the alteration of raw output appear in Varonis and Gass (in one of four experiments based on oral data, 1982), and in Roberts (who modifies punctuation only, as cited in Rifkin and Roberts, 1995). Most recently, Roberts and Cimasko (2008) limit themselves to correcting only spelling and punctuation in written data, in accordance with Sheorey's (1986) and Vann, Meyer, and Frederick's (1984) findings that such mistakes have little relative bearing on evaluative decisions. One may speculate on whether Roberts and Cimasko have instigated a phasing out of editing raw data, a shift that would further validate experimental results.

While NS and NNS participants involved in assessing learner output need not be given strict criteria on which to base judgements, they must be able at least to assume some set of parameters within which to evaluate. They should be informed of what would be expected of learners in order to achieve, for example, a mark equivalent to 100%. Rifkin and Roberts (1995) caution that "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" (p. 532). With data based on oral interviews, these minimum parameters are not easily determinable since social and contextual variables including pronunciation, intonation, word stress, body language, and pragmatics, which enhance or inhibit communicative efficacy, may profoundly influence individual listeners' judgements of appropriateness and correctness. Further, Albrechtsen, Henriksen, and Færch

(1980) add that while "interlocutors are capable of determining when an interlanguage text is incomprehensible, they can only guess whether the interpretation they give an interlanguage utterance is in fact the interpretation intended by the interlanguage user" (p. 367). These differences in individuals' abilities to guess their L2 interlocutors' meaning can produce a variation in levels of comprehensibility reported. For example, in Galloway (1980), a group of eight nonteaching native Spanish speakers rated the same learner's utterances everything from "understood nothing or very little" to "perfect or near-perfect" (p. 430).

An analysis of error gravity based on written output, on the other hand, can eschew problems that inevitably attend the analysis of interview-based data. Among the variety of means typically employed to evaluate individual nonnative speakers' overall English language proficiency skills, scores on dictation test sections in particular have been found to correlate best with more comprehensive batteries of tests (Oller & Streiff, 1975) and, as Mislevy and Yin (2009) note, dictation tests continue to be used in broad English proficiency tests. For these reasons, this inquiry employs a dictation quiz as a means to measure differences between students' and teachers' evaluations of written output. In doing so, the evaluators (a) are privy to the entire intended message, (b) are less likely to have their attention diverted from the message to the code (possibly causing irritation[1]), and (c) have had their expectations primed to a greater degree than that possible with oral interview-based data.

This paper will attempt to locate and compare differences in teacher versus student group perceptions of what constitute an error by noting observable gaps in degrees of error severity accorded by each group. The research questions to be addressed are:

1. Do native-speaking and nonnative-speaking English teachers share perceptions of error gravity?
2. Do teachers on average deduct more or fewer points than their students for the same written errors?
3. For what reasons are the various point deductions made?

Most importantly, it is hoped that an examination of participants' evaluations of various errors will afford a glimpse at the otherwise abstruse, raw template of principles upon which error judgements are based.

Method

This investigation endeavours to overcome the two limitations of inauthentic (or edited) data and noncontextualized communication in error gravity studies to date with an easily-replicable experiment in dictation. In order to preserve data authenticity, student writing collected and evaluated in this project was not edited, and was left in the learners' original handwriting. To provide a situated listening context that elicits and mimics the real-life need for learners to understand spoken orders, the dictation consists of a series of brief, sequential instructions (see Appendix). These dictated instructions in fact describe the procedure that listeners are to follow during and after the dictation itself, creating an immediate, situated, and relevant listening context.

Participants

The participants were a convenience sample of 183 humanities and science students from various subject majors, including English, of various levels of English ability. The teachers involved were these students' regular EFL professors from four universities and colleges in northern Japan. The three native English-speaking professors hailed from Australia, Britain, and the United States; the two Japanese professors were already familiar with several varieties of English, and had each studied in the UK and the US. All students had had previous experience listening to their professors speak English in the 8 to 9 months of class time preceding the dictation. Each English instructor incorporated the dictation into overall student evaluations in the form of a minor quiz.

Procedure

In light of Brodkey's (1972) conclusion that "familiarity with the voice of the specific speaker is a major variable in comprehension" (p. 216) and in order to be fair to student participants, the five professors were each requested to read the dictation to their own class. Further, as Rost (2002) notes that "a normal speaking rate has about eight words per every two- to three-second burst of speech" (p. 21), efforts were made to have utterances resemble those of this type of naturally occurring discourse, with seven sentences of five-to-nine words each read at natural speed. No attempt was made to employ vocabulary beyond that of everyday classroom-level English. Professors were asked to read each sentence twice, and to wait 8 seconds before proceeding to the next sentence. Steps were taken to render the

dictation as context-rich as possible, and to include common contractions and a degree of word repetition. One question, number 6, required listeners to distinguish between the minimal pair of *collect* and *correct*, either of which would make sense in the sentence.

After the dictation, students were told to pass their completed dictation quiz to the student behind them. Students were given an answer key and then asked to mark each question as they felt appropriate (with the instruction "You be the teacher") on a scale of 0 to 5; again, professors were asked to allow students to decide for themselves how to deduct points for errors. An additional request was made for students to provide an explanation (in their L1 or English) of why they allotted points as they did. Most (79.8%) of the students in four of the five classes in this study provided comments on their marking strategies (students from the smallest class, of just four students, did not provide information on how they made evaluations). Their teachers then collected all the papers and assigned final scores beside those of their students, and in a different colour. They employed the same 0 to 5 point scale for each of the seven questions, allowing for a maximum score of 35. Comparisons were then made between the marks of the students and the marks of the teachers in each of the five classes.

Results

It is vital to bear in mind that since each set of teacher-student evaluations for each class is based on a different set of data, it is impossible to make direct comparisons across classes in pursuit of locating significant differences among markers. It is, however, possible to compare the *gaps* between teacher and students in each class. These individual class-level gaps have been converted into percentages in this paper. While such percentages may appear to represent parametric (i.e., equal-interval based) data, the reader must keep in mind that the process of evaluating errors shall be assumed in this paper to produce nonparametric data. For example, a student-produced sentence that received a score of four from an evaluator will not be considered exactly twice as good as one that receives a two.

In all, teachers were found to have evaluated students 10.3 percentage points lower than students had. The NSTs marked an average of 11.6% lower than their students did, and the two NNSTs similarly gave 8.3% lower scores overall. Excluding the comparatively large gap (of 17.9%) in the class of only four students, the average difference between NSTs and their students' marks averages 8.5%, and that between all teachers and their stu-

dents drops to 8.4%. Taken individually, NSTs' average marks for the entire quiz were lower than those given by their students by 5.3% (*n* = 46), 11.7% (*n* = 80), and 17.9% (*n* = 4), while the NNSTs' average marks were 3.9% (*n* = 17) and 12.7% (*n* = 46) lower than those of their students. Figure 1 shows an overview of student and teacher mean scores for each question, and for total scores on the dictation given by each.

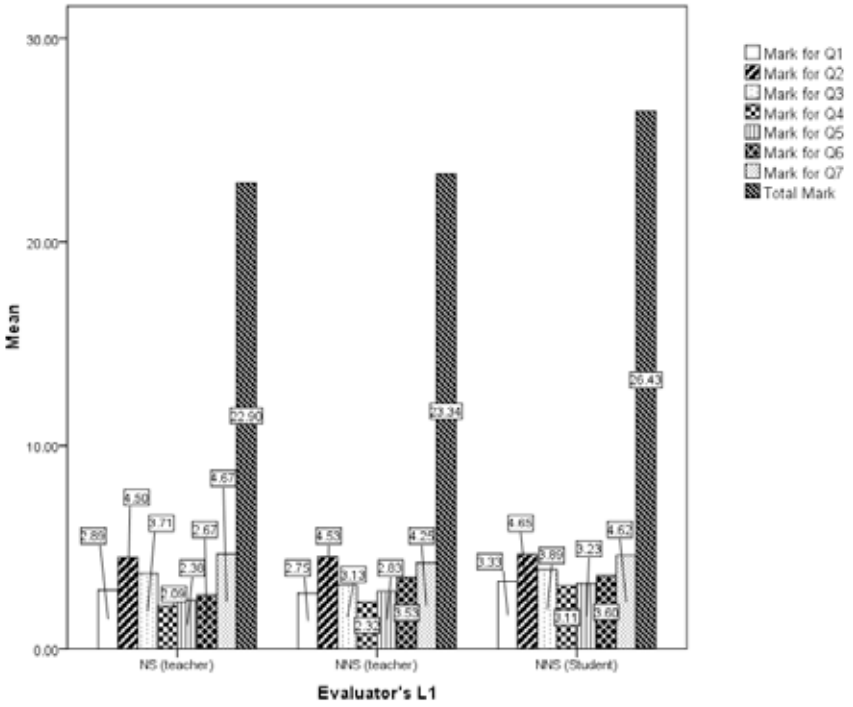

Figure 1. Mean Scores (out of 35) Among NSTs, NNSTs, and Students for Each Question

While students exhibited a tendency to award higher scores than their professors, teachers and students alike generally concurred upon which of the seven questions students fared best in (Question 2, then 7) and to a lesser degree upon which they fared worst in (Question 4, then either 5 or 1) as in Table 1.

Table 1. Overall Rankings of Questions Arranged from Highest to Lowest Mean Mark

NS teachers:	Q7, 2, 3, 1, 6, 5, 4
NNS teachers:	Q2, 7, 6, 3, 5, 1, 4
Students:	Q2, 7, 3, 6, 1, 5, 4

Teachers and their students gave identical marks 54.1% of the time on individual questions, ranging from 42.9% to 76.5% among the five classes. Marks from NS teachers coincided with their students in 49.1% of cases and those from NNS teachers and their students in 61.6% of cases. Students gave fewer points than their professors on less than 7.1% of individual questions, with NST classes averaging 7.9% and NNST classes 5.9%. Teachers, however, awarded fewer points on 40.1% of questions (NS mean: 43.1%; NNS: 37.1%). Little difference was observed between NS and NNS groups of teachers in these two respects. Further, evaluations from teacher groups exhibited a similarly broader spread than those of their students, with the former at an observed average standard deviation of 1.37 (NS: 1.46; NNS: 1.24), as compared with the student average of 1.01, for a difference of 0.36. Again, if we exclude the relatively large gap (0.92) in standard deviations between students and their teacher in the smallest (n = 4) class, the average difference between teachers' and students' standard deviations would be 1.24, with NST and NNS nearly identical (1.25 and 1.23 respectively).

Yet there were few questions on which scores between teachers and their students differed with statistical significance, as measured by the Mann-Whitney test (see Table 2). This nonparametric test was employed in the analyses because (a) the 5-point evaluative scale on which sentences were rated cannot be assumed to produce interval data, "in which all the points on the scale are equally distant from one another" (Larson-Hall, 2010, p. 394); (b) the data from the largest class (n = 80) failed the Kolmorgov-Smirnov test for normal distribution (with a 2-tailed Asymp. Sig. value of 0.01), a prerequisite for conducting parametric t-tests; and (c) the data included several outliers: Nonparametric tests technically make use of the *median* in sets of data, which is insensitive to outliers and can thus carry more power (*Ibid*., p. 374).

Additionally, according to Hinton, Brownlow, McMurray, and Cozens (2004), the correlation coefficient and effect size for this type of nonparametric data with a large number of tied ranks, as here, can best be represented by the Kendall tau-b, as listed. The greatest and most significant gap

in individual classes between student-teacher mean marks was observed in Question 4, with two professors' sets of marks, one NST and one NNST, differing from their students at the 0.01 level of significance (U = 1109.5; U = 326.5), and one NST showing a considerable difference (U = 2.0, *ns* (0.11)). The second largest gap was observed in Question 5, on which three of five classes showed important differences, with two (one NST, one NNST) of these being significant at the 0.05 level (U = 2555.0; U = 456.5), and one NST very nearly significant (U = 820.5, *ns* (0.059)). Further, the teacher and students from only one (NST) class differed significantly on Question 6 (U = 2419.0, p ‹ 0.01), though two classes did produce very different mean scores (U = 872.0, *ns* (0.142); U = 502.0, *ns* (0.087)) from NS and NNS teachers respectively. Lastly, one NST class showed a significant difference in their evaluations of Question 1 (U = 2416.5, p = 0.01).

Table 2. Summary of Greatest Differences in Scores from Teachers and Students

Question No.	Evaluator (No. of students)	Mann-Whitney *U*	Asymp. Sig. (Two-tailed)	Kendall tau-b
1	NST (80)	2416.5	0.006	0.763
4	NST (80)	1109.5	0.000	0.439
4	NNST (36)	326.5	0.000	0.714
4	NST (4)	2.0	*ns* (0.065)	0.800
5	NST (80)	2555.0	0.024	0.889
5	NNST (36)	456.5	0.027	0.700
5	NST (46)	820.5	*ns* (0.059)	0.758
6	NST (80)	2419.0	0.006	0.781
6	NST (46)	872.0	*ns* (0.142)	0.808
6	NNST (36)	502.0	*ns* (0.087)	0.778
1	All teachers: All students	13935.0	0.05*	0.593
4	All teachers: All students	9245.5	0.000	0.620

Question No.	Evaluator (No. of students)	Mann-Whitney *U*	Asymp. Sig. (Two-tailed)	Kendall tau-b
5	All teachers: All students	13064.0	0.000	0.758
6	All teachers: All students	12909.5	0.000	0.808

*At a value of 0.004656

Statistically significant differences are found in Mann-Whitney comparisons between all student and all teacher evaluations performed on Questions 1, 4, 5, and 6. According to these measures, students and teachers differed most in their evaluations of Question number 4 ($U = 9245.5$, $p < 0.01$), followed by Question 6 ($U = 12909.5$, $p < 0.01$), Question 5 ($U = 13064.0$, $p < 0.01$), and Question 1 ($U = 13935.0$, $p < 0.05$).

Discussion

Again, it must be remembered that comparisons of scores across groups of teachers and students from different classes cannot be made directly, as each teacher and student evaluated only his or her own class's set of student-produced data. For example, it may appear that, since NNST scores happen to have coincided with those of their students 12.5% more often than NSTs' did, NNSTs were somehow less strict or more sympathetic to students who shared their L1. But it could easily be the case that the students in the NNST classes happened to be more advanced in English, or that mere chance accounted for the slightly higher correlation (since, for every question evaluated on a 0 to 5 point scale, there is a 16.7% probability that marks from any two evaluators will be identical). Thus, the focus of this discussion will be limited to tendencies observed in the most remarkable of gaps between teachers and students in the *same* class, namely differences of two points (40%) or more. The larger questions as to whether and how perfectly the tendencies observed in this experiment reveal universal ones lie beyond the scope of this paper, as does the issue of whether the teachers with whom students are compared mark consistently (Ferris, 2006).[2]

First, an analysis of mean marks (see Figure 1) shows that students did not give fewer marks than their teachers on any of the dictation questions in any of the five classes. One of the professors in this study noticed this tendency toward leniency and remarked that his students, even in a large and assumedly un-intimate class of 80, were "overly generous to each other."

A student from this class, in a chat with the researcher hours after the quiz, mentioned she had felt "some pressure" sitting next to the person she was asked to mark. Friendliness toward classmates and sympathy toward fellow language learners has been shown to influence evaluation (Galloway, 1980), and may have accounted for some teacher-student differences. These were demonstrated most vividly in students' comments at the bottom and in the margins of quiz papers, such as:

1. "Yes, too bad these [*rr* and *ll*; *fish* and *wish*; *wonder* and *another*] sound so similar!"
2. "I'll give you this one!" (Written after having awarded full points for "We will do seven sentence in know" in Question 3.)
3. "I can see your tenacity." (Written after giving two points in Question 6 for "Our collect [space] paper.")

However, these kinds of remarks appeared infrequently. Still more, expressions of emotion written in margins did not necessarily translate into marks that differed from those of teachers. One student evaluator even expressed overt annoyance with a fellow learner, writing "NOPE!!" in Japanese under a mistaken transcription, and giving it one point; the NST allotted three.

More importantly, it is certainly possible that any trends toward generosity among students marking students may not have been intentional: They may simply have been basing their judgements on evaluative paradigms that do not account for variation in error gravity. The main problem seems to be that students are not equipped to judge the intelligibility of learner output. This observation concurs with findings in other studies, which tabulated error counts as either lexical, or verb-related or semantic difficulties, and in which instructors displayed a strong overall tendency to mark content over form (Guntermann, 1978; Olsson as cited in Khalil, 1985; Piazza, 1980; Roberts & Cimasko, 2008; Sheorey, 1986). Examples of this abound in data that illustrated the widest evaluative gaps, when students awarded responses two or more points (out of five) more than their teachers. For instance, a response to Question 2 of "Please like down what you are say" was given three points by a student and just one point by the NNST. The student evaluator noted in Japanese only that *like* should have been *write* and that *are you* should have been *I am,* and appears not to have noticed the missing *ing*. Again, the evaluator does not appear to have considered whether the phrase *like down* would be comprehensible to a prospective reader, or whether a

shifting of the grammatical agent from *I* to *you* radically alters the sentence's core meaning.

Another, even clearer, example of this gap in the weighting of context-based flaws can be seen in a response to Question 1, *We'll now focus on some listening*, mistakenly reproduced as "We will now forget some listening." The student evaluator explained a marking scheme (in Japanese) at the bottom of the page as "up to 2 errors gets 4 points" and "3-5 errors gets 3 points," and subsequently allotted the sentence four out of five points, while the professor granted just one. The same (NS) teacher later informed the researcher that he had originally attempted to draw up and adhere to a marking scheme of his own, with rules such as "word missing = -1 point," but soon found himself compelled to abandon it.

A comment from a Japanese colleague before beginning this investigation stressed the gaps in confidence and capability between teachers and students in regard to the task of evaluating English output. She mentioned that few if any students in Japan have had experience in marking another student's work; even the few who have such experience would not likely have any in judging sentence intelligibility in their native language, not to mention in an L2. It is thus to be expected that this new challenge could leave many student evaluators uncomfortable and perplexed; and it is understandable that introducing the very idea of marking each other's papers (mentioned in Question 4) as students see fit (Question 5) may have caused immediate difficulties.

Yet despite student evaluators' lack of experience and dearth of confidence, a number of them managed both to describe and apply a marking scheme. While 59 of the 183 students made (unsystematic) comments on specific point deductions made on individual questions, 87 outlined general rules they followed in evaluation, and five did both. Of the 87 descriptions of marking schemes, 46 accounted for some degree of error gravity. For example, one student indicated in Japanese at the bottom of the page "I basically deducted one or two points according to my sense of the error's size." Another, again in Japanese, wrote "I deducted one point for spelling mistakes and just one point in sentences containing only one mistake. In cases where the student had completely misheard what was said, I deducted two points (where this dramatically altered the sentence's meaning)." This particular student, who seems to have arrived at a comparatively advanced approach to marking, gave the quiz a 31 (out of 35) overall, and the NST gave a 28.

Students' remarks at times accounted for error gravity more subtly. One student chose to give three points to a rendering of Question 6 as "[space]

collect everyone's papers later," indicating not only that the verb had been misrepresented but that a grammatical subject was missing ("*Shugo ga nuke,...*"). While the pronoun could (and would) be dropped in this sentence if spoken or written in Japanese, the evaluator's point deduction and brief comment on the lack of subject underscores an acquired, error gravity-based sensitivity to pronoun dropping that mimics that of the proficient English user. The NST, by comparison, also awarded the sentence three points. The student evaluator's and teacher's total scores for this quiz were 31 and 29 respectively, out of a possible 35 (a 5.7% difference).

With these similarities in mind, an investigation was then conducted to determine whether students who had described marking schemes that reflected some sense of error gravity subsequently awarded overall marks that matched or nearly matched those of their teachers, as in the above examples. A comparison of the total scores on tests marked by these participants and their teachers, however, revealed that these students' sense of error gravity in practice accorded either (a) very closely with that of their professors or (b) very little. Out of the 46 students identified as error-gravity sensitive. 25 seem to have fared well, differing from their teachers by just 3/35 points (8.6%) or fewer in quiz score evaluations, while 11 students differed by 8/35 (22.9%) or more, with only 10 falling between these extremes. Thus, in this study, the learners' application of self-determined, error gravity-based marking schemes generally yielded scores either very similar to or very different from those of the teacher, with the former tendency observed most frequently. While one cannot state with certainty whether a learner's consciousness of error gravity would help close the overall gap between teacher and student evaluations, the data here imply that in most cases it would. Having said this, however, it must be noted that the difference in mean overall quiz scores between students who showed an awareness of error gravity and their teachers (12.4%) was more than that between all students and their teachers (10.3%).

Students and teachers differed most in their evaluations of responses to Question 4, the most challenging of the seven. Interestingly, students fared worst on this sentence despite its being embedded in a context, as fourth in the series of seven, through which learners might have made informed guesses at what was being said. More interestingly, teachers and students disagreed by two or more points in their evaluations of this question in 55 instances, more than on any other question. The expression *one another's papers* challenged almost all students, with confusing and confused renditions observed even in responses from students of apparently advanced

proficiency levels. One student, who had accurately transcribed every word of every other question on the quiz, wrote "After this quiz, you will mark one of my papars" for Question 4, a mistake that (with the minor spelling mistake) cost her four points in the eyes of her NS professor, but only one in those of a fellow student.

Close examination of Question 4 reveals numerous examples of student evaluators also encountering difficulty. One who had detailed a comparatively error gravity-sensitive marking strategy (in English) stating "I marked 4 to the sentences which I understood the meaning even though there were some mistakes," granted the following response significantly more points than the NS professor:

- After this quiz, you mark in the another paper. (Teacher's mark: 1; student's: 4)

The student evaluator appears to have misunderstood the semantic cost of the writer's replacing *one another's* with *in the another*. On the surface, the two sentences differ little in word count (nine and eight); the insertion of the word *in* has a negligible effect on meaning; and the omission of the abbreviated auxiliary verb (*'ll*) can be compensated for through the logic of chronology.

The student evaluator indicated in the above comment that the sentence produced was flawed but understood. This leaves two possible explanations for the mark of four out of five, namely that: (a) the evaluator had actually misunderstood both the correct and produced sentences to mean that quiz takers were later to mark another paper, in which case the single point was deducted simply for the several omissions committed; or that (b) the evaluator deducted the single point after having only lightly regarded the impact of all errors combined. Further, since the evaluator's L1 does not require distinction between singular and plural, the very possibility of confusion as to the meaning of *the another* would not likely occur to any but the most advanced Japanese L1 learner of English.

Yet the question remains whether there were fewer gaps between students who indicated an awareness of error gravity and their teachers on, for example, the easiest and most difficult questions (i.e., Questions 2 and 4 respectively). Since overall mean scores on these questions as given by the evaluators hide gaps between each teacher and student on every individual quiz paper, student–teacher differences were here based on each student–teacher instance of evaluation. The 46 error-gravity conscious students,

those who stated they treated errors in terms of perceived severity, allotted marks for Question 2 that differed from their teacher by 16 points in all. Thus, the mean difference between teachers and error-gravity conscious students on Question 2 was 16 ÷ 46, or 0.35 points out of five (7.0%). Yet when differences on individual instances of Question 2 were examined for all 183 students and their teachers, an even smaller mean difference of 0.2 points out of five (4%) was found. Thus, students who displayed awareness of error gravity did not fare better than those who did not.

Similarly, in the more challenging Question 4, the same 46 students did not evince evaluative decisions that brought them in any more concordance with their teachers than the entire student group. They differed from their teachers by 59 points in total, or by 1.28 out of five points (25.7%), whereas students collectively differed by only 1.1 points (22%). Perhaps surprisingly, the data does not support the notion that students who specifically indicated an awareness of error gravity were more likely to make evaluations that agreed with those of their teachers on either the easiest or most difficult questions.

Salem (2004) suggests that local teachers who share their learners' L1 can grow tolerant of their typical errors, a tendency that would lessen the extent to which NNST scores in particular would differ from those of students. However, the Japanese L1 teachers in this study did not display sympathy, for example, with students' mistakes based on misinterpretations of sounds generally considered similar by native Japanese speakers; NNSTs penalized errors that significantly altered meaning. For example, on the second most challenging question, number 5 (*Feel free to mark these as you wish*) a student allotted "Feel free to mark this fish" three points. The student evaluator's sympathy was revealed in the marginal comment "These sounds are similar—too bad!" The NNST displayed no such sympathy and awarded the response a zero.

Similarly, the other NNST involved in this study deducted marks in the dictation for sounds frequently mistaken by Japanese L1 learners (here, *l* and *r*). The NNST gave the Question 6 responses "I'll correcte everyone's paper later" two points and "I'll correct [space] every one's paper later" three, compared to student evaluator points of five and four respectively. While much research has claimed that nonnative-speaking teachers tend to be stricter than native speakers on students overall (Davies, 1983; Fayer & Krasinski, 1987; James, 1977; McCretton & Rider, 1993; Nickel as cited in Johansson, 1978; Sheorey, 1986), and that NNS teachers are particularly harsh in regard to errors based on well-cautioned pitfalls (Davies, 1983; Hughes

& Lascaratou, 1982), neither of the Japanese professors here demonstrated any such pattern. By way of comparison, the only sentence produced in a NST's class with just the single /l/-/r/ confusion, "I'll correct everyone's paper later," received three points from the professor and five from the student evaluator, revealing a gap similar to that between NNSTs and students.

The majority of errors in transcribing Question 6 were the result of listeners' mishearing *I'll* and *collect* as *All/Our* and *correct*, and substituting *sentences* or *answer* in the place of *papers*. Of course these confusions alone do not account for every gap between teacher and student evaluators, as the latter group also tended to ignore singular-plural errors and to treat missing words, including key words, more leniently. The following examples represent typical discrepancies based on misinterpretations of words that sound similar to Japanese listeners:

1. * All corect everyones [space] later. (NST's mark: 0; student's: 2)
2. * All corrects everyone's paper [space]. (NST's mark: 0; student's: 4)

Further, although student evaluators in both of the above examples describe marking schemes that account for whether the sentences make sense, the same students do not seem able to determine, as their teachers did, which type of errors render sentences incomprehensible. While both the more conscientious students and the teachers in this study do not appear to believe that "an error is an error," students' perceptions of correctness in terms of overall intelligibility occasionally failed them on the more difficult questions.

The scores for questions on which student-writers had fared best generally indicated what student and teacher evaluators concurred upon. Recall that 54.1% gave identical marks on individual questions, and that both groups of evaluators agreed overall on which of the seven questions presented not only the most but also the least difficulty. This tendency suggests that most evaluators do share, to some degree, a set of standards upon which error and correctness is identified. McCretton and Rider (1993) found this kind of overall agreement in their study of NST and NNSTs' evaluations, and noted that because the "order in which both groups ranked the errors was remarkably similar, [the authors were led to] consider the validity of establishing a 'universal hierarchy of errors'" (p. 177). One may make similar speculations based on this study, and add that a small group of students also shares in a perception of a hierarchy of error. However, the same caveat McCretton and Rider cautioned against would hold, in that an elemental sense of what error

and correctness mean may be a by-product of most teachers' and students' educational training, as opposed to an inevitable result sprung from an innate, universal hierarchy of errors within every language user. The contention that education plays the greatest role in shaping learners' perceptions of error and correctness seems particularly valid in EFL environments such as those where this study was based (northern Japan), where contact with English remains largely conscribed to school classes and private lessons.

The sense of what constitutes linguistic *correctness* for evaluators was of course most observable in questions awarded the highest marks. The only case in which a class and their teacher gave identical marks occurred on the question that received the highest mean score (97.6%) of any single question in any class. The students in that particular class, the only class of English subject majors and minors investigated, had also scored higher than any other class on the entire quiz (at 71.1%), with teacher and student marks coinciding 76.5% of the time overall.

Contrarily, the class that fared worst on the dictation (at 55.7%) also differed most with their professor in evaluations, which coincided with those of their professor only 42.9% of the time (cf. the overall student–teacher correspondence of 54.1%). This tendency was also observed in two of the remaining three classes. Based on this evidence, students who exemplified better performance as a group on the dictation quiz also appeared to hold perspectives on error that most frequently coincided with those of their professors. Yet one class of students that had performed second worst on the dictation (at 59.5%) differed with their professor by only 5.3% in mean quiz scores allotted. Thus, while one may say that students of higher proficiency as measured by this particular quiz generally gave marks that resembled those of their professional counterparts, the trend cannot be considered a rule.

Lastly, research on error gravity has shown that native speakers of English do not consider spelling errors to be as important as most other error types (e.g., Johansson, 1978; Hughes and Lascaratou, 1982; Vann et al., 1984; Sheorey, 1986). In a ranking of error gravity by Vann et al., spelling errors appear along with article problems and comma splices in the top three most tolerable error categories (p. 431); further, Johansson chose to disregard spelling errors in his analyses "provided that the words were clearly recognizable" (p. 81). While participants in this study were not requested to provide information on spelling issues in particular, several points in regard to spelling can be observed in the data. One student, for example, circled the misspelling of *quiz* as "qiuiz" and docked one mark for it, perhaps not realising this spelling error runs no risk of being interpreted for any other word

in English. The NST appears to have overlooked the mistake and awarded full marks to the otherwise perfect transcription. However, later on the same paper, the teacher circled and deducted one point for the misplaced second apostrophe in "I'll collect everyones' papers…" whereas the student evaluator gave full marks. This difference reveals that at least some students cannot rate the gravity of spelling mistakes in context, and perhaps that the teacher chose to draw the learner's attention to a lexical problem.

Only seven of the 183 student evaluators mentioned any guiding principle in regard to spelling in particular. Two noted that spelling mistakes were to be ignored, and five indicated they would deduct points for them. While this sample size of seven cannot supply conclusive proof, analyses show that students' consideration of spelling errors as either *important* or *not important*, based on their comments, did not bring their overall quiz evaluations more or less in accordance with those of their teachers. The mean teacher–student gap among students who considered spelling important was found to be 10%, with that in the other group at 15%.

Conclusion

In the context of typical language-teaching courses and classroom comprehension checks via dictation, Davis and Rinvolucri note:

> There is no call for the teacher to take on responsibility for correcting dictation scripts. Such work requires care, of course, but it does not require the kind of linguistic judgement that only the teacher can make. Correcting a dictation is a straightforward task which students are quite capable of doing for themselves. . . . (1988, p. 4)

This study shows, however, that while students are generally able to give token-like penalties for the errors they noticed, and while peer review itself can be seen as beneficial to learner reviewers in improving their own writing (Lundstrum & Baker, 2009), many students are not capable of correcting a dictation quiz in a manner similar to that of their teachers. When given an answer key, students perceive what is wrong but not to what degree. Students' marks resemble those of their professors when sentences lack words or contain misspellings, but they stray when evaluative decisions involve context and intelligibility.

One last example will help illustrate the point above. The following student awarded the answer to Question 4 of "After this is quiz one anther's

papers" two points compared to the NST's score of zero. Yet the student evaluator demonstrates an awareness of error gravity in a comment (in almost flawless English) at the bottom of the quiz paper that deserves to be quoted in full:

> She made some mistakes in using verbs such as will → are, which I believe shouldn't be happening. This is the priority when I was marking, if there were mistakes in using verbs, I gave her low scores. Spelling and putting periods came second highest.

Again, the student noticed the significant problems with the sentence and deducted three points, but stopped short of judging it incomprehensible.

One of the NST participants in this experiment raised the issue of whether differences in student and teacher confidence levels as English users may have caused students to hesitate in deducting many marks for intelligibility. It would certainly have been easier for students to judge the correctness of sentences in terms of grammaticality alone. The traditional EFL classroom culture of translation and grammar-based, test-focussed language classes may also have contributed to students' awareness (and ignorance) of what constitutes a serious error. If so, what McCretton and Rider (1993) suggest influences NSTs and NNSTs may well also apply to student evaluators of this dictation: Evaluators' perception of error may not depend as much on a universal hierarchy of errors as on their own educational training. Habits of marking that tend to assign priority to grammar over communicability, for example, may be overcome by local teachers with lengthy experience in the language outside the classroom. Perhaps many EFL learners have not yet arrived at that stage.

Evaluator anonymity prevented finer comparisons within groups in this study, and in some respects limits the current findings to students and teachers with educational and linguistic backgrounds similar to those in the groups surveyed. It follows that findings from this descriptive study may not be used without reservation to predict the outcomes of evaluations by groups elsewhere. Future inquiries should include more information on each student evaluator; yet researchers must, at the same time, employ measures that avoid causing enmity or loss of face in order to elicit unreserved reactions to error. This experiment may also be revised and expanded to include more ordinal data-appropriate, letter-score evaluations, and to include responses from two extremely important groups: nonteaching native and nonnative English users, who after all must be considered the

target interlocutors for modern EFL learners. Measures could also be taken to ensure that evaluators do not see each other's scores, thereby avoiding a possible influence on their own scoring. Lastly, evaluators from different groups could be asked to mark the same quizzes, or a sample of quizzes from classes other than their own in order to allow for inter-class comparisons.

One more paradox involves the issue of how to determine students' awareness of error gravity. In this study it was found that students who appeared sensitive to degrees of error in their comments on evaluation did not necessarily make better evaluations than students who did not. One may interpret this to mean that not all students who report an awareness of error gravity have understandings that coincide with those of teachers. Or it may also simply be the case that not all students who possess an awareness of error gravity, accurate or not, reported it. Future inquiries should be structured to assess student perceptions of error gravity without compromising their freedom to describe their evaluative processes as they wish.

This project has shown that an investigation of gaps between student and teacher evaluations can help describe and assess students' interlanguage proficiencies while locating significant stumbling blocks faced by student evaluators in interpreting fellow learners' and teachers' writing. In-class attention to the types of evaluative gaps revealed in this study would not only introduce the concept of error gravity to students, but would give teachers and learners an opportunity to refocus their energies. It would doubtless uncover several problem areas to examine in preparing learners for peer editing, and could indeed provide a fast track to improved EFL proficiency.

Notes

1. Ludwig defines irritation as "the result of the form of the message intruding upon the interlocutor's perception of the communication," ranging "from unconcerned, undistracted awareness of a communicative error to a conscious preoccupation with form" (1982, p. 275).
2. In Ferris (2006), four researchers judged teachers' marks of students' written errors to be correct in 89.4% of the cases. However, teachers' marks were judged as incorrect 3.6% of the time, and unnecessary in 7% of the study sample.

Acknowledgements

The author dedicates this paper to those who suffered in the Tohoku earthquake of 11 March 2011. Special thanks are due to all those who helped.

The author also wishes to thank Shin'ichi Inoi at Ibaraki University for his suggestions on research design, and Ken Inoue at Fukushima University for advice on statistical methodology.

Sean Mahoney has been an Assistant Professor of English at Fukushima University since 1997, and has written on TOEIC, the JET Programme, and World Englishes.

References

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

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility. *Language Learning*, *22*, 203-220.

Davies, E. (1983). Error evaluation: The importance of viewpoint. *ELT Journal*, *37*, 304-311.

Davis, P., & Rinvolucri, M. (1988). *Dictation: New methods, new possibilities*. Cambridge: Cambridge University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd. ed.). Oxford: Oxford University Press.

Fayer, J., & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning*, *37*, 313-26.

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland, & F. Hyland, (eds.), *Feedback in second language writing* (pp. 81-104). New York: Cambridge University Press.

Galloway, V. (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.

Hinton, P., Brownlow, C., McMurray, I., & Cozens, B. (2004). *SPSS Explained*. Sussex, UK: Routledge.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal*, *36*, 175-182.

James, C. (1977). Judgments of error gravities. *ELT Journal*, *31*, 116-124.

Johansson, S. (1978). Studies of error gravity: Native reactions to errors produced by Swedish learners of English. *Gothenburg Studies in English 44*. Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

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

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

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

Lundstrum, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*, 30-43.

McCretton, E., & Rider, N. (1993). Error gravity and error hierarchies. *International Review of Applied Linguistics*, *31*, 177-88.

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning, 59*, 249-267. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9922.2009.00543.x/full

Oller, J. W., & Streiff, V. (1975). Dictation: A test of grammar-based expectancies. *ELT Journal*, *30*, 25-36.

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

Rifkin, B. (1995). Error gravity in learners' spoken Russian: A preliminary study. *Modern Language Journal*, *79*, 477-490.

Rifkin, B., & Roberts, F. (1995). Error gravity: A critical review of research design. *Language Learning*, *45*, 511-537.

Roberts, F., & Cimasko, T. (2008). Evaluating ESL; Making sense of university professors' responses to second language writing. *Journal of Second Language Writing*, *17*, 125-143.

Rost, M. (2002). *Teaching and researching listening*. Harlow, Essex, UK: Pearson Education.

Salem, I. (2004). Teacher differences in perception of student error. *English Language Teacher Education and Development*, *8*, 48-65.

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

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

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

Vann, R., Meyer, D., & Frederick, 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.

Appendix

In-Class Dictation Quiz

1. We'll now focus on some listening.
2. Please write down what I'm saying.
3. We will do seven sentences in all.
4. After this quiz, you'll mark one another's papers.
5. Feel free to mark these as you wish.
6. I'll collect everyone's papers later.
7. This is the last sentence you need to write.