Indirect Negative Evidence as Corrective Feedback in Second Language Writing: Comparing Output to Input

Caroline H. Vickers
University of Arizona

This study examines the effect of comparing learner output to input in increasing the accuracy of grammatical form in L2 writing. Plough (1994) discusses indirect negative evidence as the setting up of an expected environment so that when the expected environment is altered, learners notice the difference via comparison between the expected and the unexpected. The study suggests that the expected environment is the learner's own output, and the input is unexpected. Two research questions are addressed: 1) Can learners notice their own output errors by comparing output to input? and 2) Does comparing output to input promote higher levels of sustained accuracy? The experimental group performed significantly better than the control group on both the posttest and the delayed posttest, though both groups improved significantly.

Input and Output in Second Language Acquisition

In second language acquisition (SLA) research, the role of input has been the subject of much debate. Krashen (1985) argues that learners need only comprehensible input to trigger acquisition. Other studies have also demonstrated the important role of input in SLA, showing that second language learners need comprehensible input to gain fluency in the second language whether the input is made comprehensible through modified input (Hatch, 1978) or negotiated input (Long, 1983). However, years of study of French immersion classrooms in Canada have shed light on the reality of the situation that immersion students who receive an abundance of second language input show high levels of fluency but low levels of accuracy in the second language (Hammerly, 1987), making their second language output somewhat stigmatized. According to Swain (1985), one reason for the lack of accuracy demonstrated by Canadian immersion students is that there are not sufficient opportunities in the immersion classes for the students to produce pushed and extended output. Therefore, Swain has argued that "comprehensible input" is necessary but not sufficient in promoting second language accuracy and, importantly, that students need opportunities to produce "comprehensible output" to promote accuracy in the second language.

As stated in the Output Hypothesis, Swain argues that output serves three functions in terms of helping students to become more accurate in the second language: 1) a noticing/triggering function, 2) a hypothesis testing function, and 3) a metatalk function. Nobuyushi & Ellis (1993) have shown the hypothesis testing function to be important because in the process of producing pushed output, learners have the opportunity to confirm or disconfirm their hypotheses about the second language as they receive feedback from an interlocutor. Therefore, the metalinguistic function of output is successful, according to Nobuyushi and Ellis, but critically, feedback is a part of the equation. Moreover, Swain & Lapkin (1998) and Swain (1998) have provided evidence of the benefits of the metalinguistic function in allowing students to think about the language through discussion of language forms, which Swain has called language related episodes. These language related episode studies show that learners gain metalinguistic knowledge through the discussion of language forms, but it is unclear whether the metalinguistic knowledge can become incorporated into the learner's interlanguage system. Finally Izumi et al. (1999) have found that producing written output draws learners' attention to their output problems but does not guarantee attention to a particular grammatical form. Moreover, Izumi & Bigelow (2000) have failed to confirm the noticing function of written

output, which is crucial because of the important role of noticing in acquisition. Schmidt & Frota (1986) argue that noticing is a conscious condition, which is a necessary condition for acquisition. Thus, the role of output in triggering acquisition is called into question.

It is possible that the failure of output to promote noticing lies in the fact that through output, attention is not drawn to particular problematic IL forms. Perhaps learners need more targeted information to help guide them to notice problematic output features and to help them notice relevant input features. Shook (1994) has demonstrated that explicitly drawing students' attention to grammatical items in the input allows greater gains in learner-readers' linguistic information about items than those learners whose attention is not called to an item and that drawing students' attention to a grammatical form leads to higher levels of intake. In Shook's study, enhanced input was sufficient to draw students' attention to target grammatical items. Furthermore, Doughty (1991) has provided evidence to support the effect of visual salience in the form of typographical input enhancement to promote noticing. In this study, groups which were exposed to enhanced input made greater gains in form learning than groups who did not receive enhanced input but focused only on meaning. Importantly, the enhanced input group also made similar gains in meaning to the meaning focused group. However, the focus of Doughty's study was not to investigate the effects of typographically enhanced input per se. White (1998), whose study concentrated exclusively on the effects of typographically enhanced input, could not confirm the effects of typographically enhanced input in promoting language learning. It is important to note that White's study depended on implicit noticing of forms on the part of the subjects, but DeKeyser (1998) argues for a more explicit focus on form, which may explain why White's hypotheses concerning typographically enhanced input were not confirmed.

One advantage of typographically enhanced input is that it allows learners to read that enhanced element in context. Such contextualized exposure to a form could be beneficial in error correction. Since both the student's output and the typographically enhanced input contain grammar used in context, learners might benefit in terms of receiving evidence about their second language hypotheses concerning particular forms. Perhaps learners could notice their own production errors by comparing their output to visually enhanced input.

The Role of Evidence in SLA

The role of evidence in the conversion of input to acquisition has been an important object of investigation in second language research. Three specific types of evidence have been shown to play a role in acquisition: positive evidence, direct negative evidence, and indirect negative evidence (Gass, 1997). Positive evidence is simply defined as exposure to contextualized input. The role of positive evidence has been posited to be necessary and sufficient for SLA as noted by Krashen (1985). According to Krashen's concept, positive evidence must be in the form of comprehensible input for it to have an effect on acquisition. On the other hand, Hammerly (1987) posits that comprehensible input is not sufficient for SLA because of a need to focus on form to promote accuracy in language learning.

Further in an effort to investigate the effect of focusing on form, direct negative evidence, both implicit and explicit has been investigated to understand the role of direct negative evidence on the learning of language forms. Implicit negative evidence has been researched in the framework of interactional modifications in SLA. It has been shown that through such modifications, such as clarification requests, confirmation checks, and comprehension checks, learners receive information that an utterance is the source of some communication problem (Long, 1983; Pica et al., 1987; Gass & Varonis, 1989). Implicit negative evidence can help

learners gain comprehensible input through negotiation, but it is not certain that such negotiated comprehensible input leads to acquisition (Long, 1991). Clearly, implicit negative evidence is important, but not completely supported in terms of its effect on the language learning process.

On the other hand, explicit negative evidence, which occurs when learners are made overtly aware of the inaccuracy of an utterance, has received more support in terms of its role in acquisition. For instance, Carroll & Swain (1993) provide evidence for the beneficial language learning effects of both implicit and explicit negative evidence in an extensive study of the role of feedback in second language learning, but explicit negative evidence in the form of explicit metalinguistic feedback was found to be superior to other implicit and explicit feedback conditions in promoting acquisition. Furthermore, Lightbown & Spada (1990) have demonstrated by comparing different instructional situations that explicit focus on form and corrective feedback are successful in promoting more accurate language use in communicative language teaching. In addition, Tomasello & Herron (1988, 1989) have demonstrated the positive effects of inducing learner production errors, which is then followed by immediate explicit negative evidence to promote rule learning. This so called "garden path technique" sets up a situation that produces a salient contrast between the learner's error and the correct form, thus promoting hypothesis testing. However, Carroll et al. (1992) call into question the results presented by Tomasello & Herron because the "garden path technique" leads to metalinguistic knowledge, but not necessarily restructuring of the learner's interlanguage system. Interestingly also, the results discussed by Carroll et al. show that explicit negative evidence has a positive effect on learning in terms of memorizing specific forms, but that it does not appear to help learners make generalizations about language form. Therefore, direct negative evidence has been shown to promote comprehensible input, metalinguistic knowledge, and memorization of items, but its effect on acquisition is uncertain.

Obviously the roles of implicit and explicit direct negative evidence have received a good deal of research attention in SLA. However, the role of indirect negative evidence in SLA has not been as well researched, and its role in SLA remains uncertain. Plough (1994) recognizes the important role of indirect negative evidence in letting a learner know that a language feature is not possible because it is never present in the expected environment. In other words, if a feature is different from that which is expected, the feature is a candidate for restructuring. Chomsky (1981) has stated that "there is good reason to believe that direct negative evidence is not necessary for language acquisition, but indirect negative evidence may be relevant" (p. 9). Lasnik (1989) also promotes the benefit of indirect negative evidence in parameter resetting. Thus, indirect negative evidence is relevant in the universal grammar (UG) framework.

On the other hand, in Plough's conception, indirect negative evidence relies not necessarily on the accessibility to UG, but rather on the use of inductive inferencing. Specifically, according to Plough, indirect negative evidence involves three sequential stages: "1) scanning what is known (either L1, L2 and/or world knowledge), 2) linking new material with what is known (it is at this stage where the absence of a structure may be noticed), and 3) establishing probably true conclusions or generalizations based on the (mis)match between new material and what is already known" (p. 90). Thus, indirect negative evidence entails noticing in terms of exposure to input and a chance to compare and contrast. Further, indirect negative evidence may help in making generalizations based on evidence.

Moreover, indirect negative evidence may be useful in the classroom to help learners not only to memorize specific forms, but to generalize to new forms since indirect negative evidence requires inductive inferencing, which by nature entails taking specific evidence into account and

making generalizations based on that evidence. It seems that learners could then benefit by making their own IL (interlanguage) output the established material and TL (target language) enhanced input the new material. This matching of the established and the new could set up a situation which would encourage the noticing of errors in the learner's output. Especially in writing instruction, in which the role of direct negative evidence has been questioned (Steinbach et al., 1988), indirect negative evidence could be of benefit in helping learners to notice their own production errors. Importantly, if indirect negative evidence can allow learners to notice errors, it should also promote higher levels of production accuracy and higher levels of competence in the second language because of the generalizing capacity inherent in the notion of indirect negative evidence. This experiment, therefore, intends to investigate two questions to find a better understanding of the role of indirect negative evidence in allowing students to notice their own production errors and whether such noticing of production errors can lead to higher levels of accuracy and higher levels of competence in the second language.

Indirect negative evidence is operationalized for the purposes of this study as comparing output errors to visually enhanced input. In essence, the participant has a chance to match the IL written output against the TL written input. Learners can scan their own output, compare and contrast their own output with the unexpected TL input, and make conclusions based on the comparison and contrast about the nature of the target form. The research questions are concerned with matching output to input as opposed to highlighting and coding type error correction, which is common in L2 writing programs. Highlighting and coding type error correction represents a type of direct negative evidence as it entails explicit information that particular forms in the output contain errors.

RESEARCH QUESTIONS AND HYPOTHESES

1. Can learners notice and correct their own writing production errors by comparing output to enhanced input?
2. Does comparing output to enhanced input promote higher levels of sustained accuracy for second language learners than more direct corrective feedback?

Hypothesis 1
Matching output to enhanced input will allow participants to notice and correct their own writing production errors at least as well as participants who have their errors highlighted and coded.

Hypothesis 2
Matching output to enhanced input will lead to higher levels of sustained accuracy than highlighting and coding corrective feedback.

METHOD

Participants

The participants for this study (n=22) were advanced and high intermediate learners of English in a university composition program in the southwestern United States. They were enrolled in a composition course which is the equivalent of freshman composition for native speakers of English, but this course was designed especially for ESL students. They were placed in the course through a departmental timed essay which was scored by committee. The students in the composition program are generally enrolled as undergraduate students in the university,

and take freshman composition during their first year in residence. The participants came from a variety of native language backgrounds, including Japanese, Spanish, Arabic, Hindi, Portuguese, Mandarin, Indonesian, Greek, and Tamil. The ages of the participants ranged from 17 to 35. Most of the students had been in the United States from six months to two years at the time of the study. Moreover, all of the participants had achieved at least a 500 on the TOEFL and had had at least four years of English language instruction at the time of the study. Following the pretest, participants were assigned to an experimental group (EG, n=11) and a control group (CG, n=11) using a stratified random assignment based on the pretest scores.

The Target Form

The target form was the past hypothetical conditional (e.g., *If I had gone to the baseball game, I would have seen Hank Aaron hit a homerun*). The form was chosen because it either did not appear in the students' written work for the course or it appeared but with obvious gaps in accuracy. The pretest indicated that the participants use of the target form was inaccurate, but that they had knowledge of the conditional in English and an ability to hypothesize. However, there were errors in the use of the past hypothetical conditional. It was clear that the form was emergent in the interlanguage systems of the students, indicating that they were developmentally ready to incorporate the form.

Research Design

This study incorporated a pretest/posttest/delayed posttest design to test the hypotheses. There was one experimental group (EG) and one control group (CG). Table 1 demonstrates the experimental sequence, which was carried out over a period of ten days, and a total of 1 hour and 50 minutes. The pretest was given five days before the treatment began. The treatment consisted of three phases. Phase one was a prewriting phase. Phase two, which occurred seven days after phase one, consisted of error correction tasks. Phase three involved a post-writing task, which was given two days after phase two. Posttest 1 then took place five days later, and the delayed posttest five weeks later.

Table 1.

Treatment	Groups	Time	Day
Pretest	Groups EG and CG	30 minutes	Day 1
Phase 1: Prewriting Task, *The Intolerable Teacher*	Groups EG and CG	25 minutes	Day 2
Phase 2: Correction Phase, *Comparing output to input* *Correcting through codes*	 Group E Group C	 30 minutes 30 minutes	Day 3
Phase 3: Reading and Writing Task, *George's Product Proposal*	Groups EG and CG	30 minutes	Day 3
Posttest	Groups EG and CG	30 minutes	Day 4

Treatment

The materials used in this study were designed specifically for advanced and high intermediate level ESL students in a university composition program. All tasks used in the treatment phases involved the written modality. The materials had been previously piloted in an ESL program among the advanced level classes in that program. Piloting of the materials helped

to assure reliability. The treatment tasks and the testing materials were also tested among native speakers to be confident about the validity of the materials.

Phase one and three of treatment were identical for EG and CG, but phase two differed in the type of corrective feedback offered to each group. The treatment was designed to isolate the differences in the corrective feedback received by EG and CG. Coding and highlighting is a popular method of providing corrective feedback on essays as indicated by ESL writing textbooks and as indicated in an informal survey of fifteen ESL teachers. However, the ESL teachers surveyed expressed frustration because they have used the highlighting and coding technique to provide corrective feedback, and students made the corrections, but the same errors occurred in the students' writing in subsequent essays. Highlighting and coding was, therefore, chosen as the treatment for the control group.

During phase one, the experimental group and control group 1 wrote a short paragraph. The task for this paragraph was intended to elicit the past hypothetical conditional by setting up a scenario for the participants in which they had to imagine being on a committee of students the month before to choose a new teacher, and then they were prompted to write (see appendix A). If-prompts were listed at the end to ensure that participants used the conditional, but the students were supposed to incorporate the if-prompts into the content of a paragraph.

The correction phase for EG and CG differed for the two groups in terms of the type of correction they received. EG was given a reading task that contained a brief explanation of the past hypothetical conditional at the top of the reading passage (see appendix B). Moreover, the past hypothetical conditional was typographically enhanced throughout the reading passage, by making the form darker than the surrounding context. Next, EG participants were returned the paragraphs which they had written in phase one with the amount of errors using the target form totaled at the bottom. They were instructed to make the necessary corrections to their own paragraphs by comparing their use of the past hypothetical conditional to the reading passage containing the typographically enhanced input. EG repeated the same procedure correcting a paragraph prepared by the researcher that had been littered with target form errors. In contrast, CG during phase two read the same passage as EG, but it had not been typographically enhanced. Next, they corrected their paragraphs after they received the same brief explanation of the target form as EG (see appendix B). Their paragraphs contained highlighted errors along with error codes, and they were instructed to correct the target form errors by looking at an index of codes and correcting the error. They repeated the same procedure on a paragraph containing target form errors prepared by the researcher.

In phase 3, both EG and CG completed another short reading task and paragraph writing task. The reading task described the exploits of a failed businessperson, and the past hypothetical conditional was incorporated throughout the passage to describe what would have happened if the businessperson had chosen a different route. Next, the participants wrote a paragraph based on a scenario that put the ill-fated businessperson in the position of another failed business venture (see appendix C). If-prompts were provided to ensure the elicitation of the target form.

Testing Instruments

The pretest and both posttests consisted of a picture based production test and a grammaticality judgement test. A split block design was used to administer the tests, and the tests were administered by the researcher. The picture based production test was given prior to the grammaticality judgement test, and the picture based production test lasted for approximately 20 minutes, while the grammaticality judgement test lasted approximately 10 minutes.

The picture based production test consisted of 10 items eliciting the past hypothetical conditional. The students were instructed that the first picture and the sentence written underneath it indicated a true situation. Next, they were instructed that the following two pictures and the verbs underneath them represented an untrue situation in the past. The pictures were connected with an arrow to indicate a cause and effect relationship. The participants were then instructed to write a sentence based on the untrue situation in the past and to use *if* to begin the sentence. The participants were not allowed to revise their answers. See appendix D for a sample picture based production test item.

The grammaticality judgement test contained 15 items, 8 of which were ungrammatical and 7 of which were grammatical on the pretest and posttest. The delayed posttest contained 10 ungrammatical items and 5 grammatical items. The number of variations in result clause ungrammaticality was increased on the delayed posttest because on the posttest, the result clause form had been the most problematic for the participants. Therefore, boosting the result clause variations allowed more insight into what the participants in both CG and EG would consider grammatical and ungrammatical with regard to that form. Both groups were very accurate with regard to the if clause. The participants were instructed to circle *U* if the statement was ungrammatical and *G* if it was grammatical. They were instructed on the meaning of grammatical and ungrammatical before beginning the test. The instructions indicated that the participants should circle the ungrammatical part of the ungrammatical items and correct them. Having the students correct those sentences that they judged as ungrammatical helped to clarify why a participant judged a sentence as ungrammatical. The participants were not allowed to revise after completing the grammaticality judgements, and the test was timed to increase the chances that the participants were using their intuition rather than their analytical skills. See Appendix E for sample grammaticality judgement test items.

Scoring

Throughout the scoring, conditional related forms were defined as modals (would, could), aspectual auxiliaries (have, had), copula in the past participle form (been), complementizer (if), and the past participle ending (-ed and -en).

Error Identification and Correction Scoring

The experimental group's error identification and correction were scored by 1) counting the number of conditional related errors they identified divided by the total number of errors, and 2) by counting the number of conditional related errors they corrected divided by the total number of errors.

Picture Cued Production Test Scoring

The production test was scored through a target-like use (TLU) analysis to count how many forms were used in a target-like by dividing the total number of correct usages over all possible contexts. The if-clause (*If I had gone to the baseball game*) and the result clause (*I would have seen Hank Aaron hit*) of the past hypothetical conditional were scored separately.

Grammaticality Judgement Test Scoring

The grammaticality judgement test was scored by assigning 1 point to items judged correctly as grammatical and 2 points to items judged correctly as ungrammatical and that were corrected accurately. If an item was judged correctly as ungrammatical and not accurately corrected, but the attempted correction was on the right part of the sentence, that item was

assigned 1 point. An item judged incorrectly as grammatical or as ungrammatical was assigned 0 points. 1.5 points were assigned to an item in which the participants were partially accurate in correcting an ungrammatical item. For instance, it was common to see improper word order in negative constructions, but an accurate correction of the item otherwise. Finally, if the item was rated as ungrammatical and the correction was attempted, but on the wrong part of the sentence, 0 points were assigned. The total score obtained was then divided by the total possible score.

RESULTS

Error Identification

Noticing of errors was operationalized by the participants' attempting to correct the error during phase two of treatment, in other words, error identification. There was not a significant difference in the noticing of errors between EG and CG even though CG noticed slightly more errors than EG as indicated in Table 2. This result indicates that participants in EG, who did not have the errors highlighted for them, identified the same amount of errors as participants on CG, whose errors were highlighted for them. See Table 2.

Table 2

Group	Mean	SD	Significance
EG (n=11)	93.2082	8.1154	p = .303
CG (n=11)	96.2900	5.2592	n.s.

Error Correction

There was not a significant difference between EG and CG on the corrections completed successfully, which demonstrates that participants can make corrections equally well in the experimental condition as in the control condition. See Table 3 in appendix F.

Table 3

Group	Mean	SD	Significance
EG (n=11)	87.1364	11.7232	p = .458
CG (n=11)	82.6727	15.1029	n.s.

Based on the results of the error identification and the error correction scores, hypothesis 1 is supported. It appears that matching output to input allows participants to notice and correct their errors as well as participants who have their errors highlighted and coded.

Picture Based Production Tests

If-Clause

To interpret the results, t-tests were run to determine the significance of the difference between EG and CG on posttest 1 and on the delayed posttest. The posttest 1 scores of EG were not significantly higher at the .05 level than the posttest 1 scores of CG for the if-clause at the .05 level as demonstrated by the t-test results. The delayed posttest scores of EG and CG were not significantly different at the .05 level either, showing that both groups maintained the gains made on posttest 1. See Table 4 in appendix F.

In addition, a One-Way ANOVA was run to determine whether there was a significant difference for EG from pretest to posttest 1 to the delayed posttest and for CG from pretest to posttest 1 to the delayed posttest. EG showed a significant difference ($p <= .001$) and CG also

showed a significant difference (p <= .001). To determine exactly where the significant difference lay, a Tukey's LSD post hoc comparison was run. The Tukey's LSD test determined that both EG and CG improved significantly at the .05 level from pretest to posttest 1, and there was no significant change at the .05 level for either EG or CG from posttest 1 to the delayed posttest (See Table 5 in appendix F). Both groups showed significant improvement on if-clause production on posttest 1, and they both maintained that improvement on the delayed posttest.

Result-Clause

The results of a t-test demonstrated that EG scored significantly higher at the .05 level on posttest 1 than did CG. Further, another t-test showed that EG also scored significantly higher at the .05 level than CG on the delayed posttest. Thus, on both testing occasions, EG's scores for the result-clause were significantly higher than CG's scores. See Table 4 in appendix F.

In addition, One-Way ANOVAs showed that from pretest to posttest 1 to the delayed posttest, EG's scores were significantly different (p <= .001), while CG's scores were not significantly different (p = .153). Thus, for CG, there was not a significant difference from pretest to posttest 1, nor from posttest 1 to the delayed posttest. For EG, on the other hand, Tukey's LSD post hoc comparisons demonstrated a significant difference at the .05 level from pretest to posttest 1, but not a significant difference from posttest 1 to the delayed posttest. Therefore, EG improved significantly on posttest 1 and maintained those gains on the delayed posttest. See Table 5 in appendix F.

The results of the if-clause production and the result-clause production provide partial support for Hypothesis 2, that matching output to input constitutes greater gains in production accuracy than correcting errors that had been highlighted and coded. While the result-clause production results show greater gains for EG, the if-clause results show similar gains for EG and CG. Moreover, both groups made significant improvement from pretest to posttest 1 on the if-clause production, and they maintained those gains on the delayed posttest. However, CG had no significant gains from pretest to posttest 1 to delayed posttest on result-clause production, while EG made significant gains from pretest to posttest 1 and maintained those gains from posttest 1 to the delayed posttest. Therefore, EG's production accuracy gains are generally more substantial than the gains of CG in after going through the different treatment conditions.

Grammaticality Judgement Tests

A significant difference at the .05 level was demonstrated between EG and CG on posttest 1 and on the delayed posttest based on t-tests. These results demonstrate that EG outperformed CG on posttest 1 and that EG also maintained a higher level of performance than CG on the delayed posttest. See Table 4 in appendix F for details.

Further, based on the results of One-Way ANOVAs, there was a significant difference from pretest to posttest 1 to delayed posttest for EG (p <= .001), but for CG , there was not a significant difference from pretest to posttest 1 to the delayed posttest (p = .909). Thus, CG did not make significant gains from pretest to posttest 1 to the delayed posttest. However, the result of a Tukey's LSD post hoc comparison shows that EG improved significantly at the .05 level from pretest to posttest 1, but there was not a significant difference at the .05 level between EG's scores on posttest 1 and the delayed posttest. Therefore, EG improved from pretest to posttest 1, and maintained those gains on the delayed posttest. See Table 5 in appendix F.

The grammaticality judgement results indicate that Hypothesis 3, that matching output to input results in higher levels of competence than correcting errors that had been highlighted and

coded, is supported in this study. EG made greater gains than CG in terms of grammaticality judgements, and they maintained their gains on the delayed posttest. It is interesting that CG made no gains from pretest to posttest 1 nor from posttest 1 to the delayed posttest suggesting that the control treatment had no effect for CG in terms of competence. See Figure 1, Figure 2, and Figure 3 in appendix G for graphical representations of the results.

DISCUSSION

The results of this study indicate that matching output to input is a more effective type of corrective feedback than correction based on highlighting and coding. First, EG was able to notice and correct output errors as well as CG even though CG's errors were identified by the researcher through highlighting while EG's errors were not identified. Further, despite the fact that CG's errors were coded to indicate the nature of the error, EG, whose errors were left without explanation, was able to correct the errors as well as CG. Therefore, evidence in the form of typographically enhanced input was as successful for allowing students to notice and correct their own output errors as the more explicit and direct indication of errors given to CG.

The production accuracy results are intriguing because of the similar gains made by CG and EG on if-clause production as opposed to the superiority of EG over CG on the result-clause production. Perhaps the past perfect, the grammatical structure in the if-clause, is an easier construction for learners to produce than the more complex grammatical structure in the result clause, the conditional plus present perfect. While the if-clause contains only one modal plus the past-participle, the result clause contains two modals plus the past participle. Therefore, different levels of complexity of the structures in the two clauses could be a factor in the differential gains.

Most striking in terms of differences in gains between EG and CG are the results of the grammaticality judgement tests. On these tests, EG made significant improvements while CG's means remained the same from pretest to posttest 1 to the delayed posttest. The results of the grammaticality judgement tests suggest that the experimental treatment promotes higher levels of native-like competence with respect to the target form than does the control treatment. However, this result must be considered with caution because as Sorace (1996) notes, there is concern about the validity of grammaticality judgements because of the possibility of guessing. It is possible that the participants guessed on the grammaticality judgements. It is also possible that EG became better at identifying errors than CG because of the scanning practice that they had during treatment. However, in combination with the production accuracy results, it would appear that the experimental treatment is certainly superior to the control in allowing participants to notice and to correct their own output errors.

It is interesting that increasing the number of result clause errors and reducing the number of correct sentences in the grammaticality judgements did not have an effect on the delayed posttest. However, there were not particular patterns that emerged that could explain with certainty the lower results for CG on the result clause production. However, it is important to note that in the grammaticality judgements, 75% of ungrammatical sentences were ungrammatical due to a result clause error. Therefore, the grammaticality judgement results again support EG's superiority with regard to result clause accuracy.

Perhaps the fact that EG had to identify production errors by scanning their own output meant that they had to do more mental work to notice their own errors than did CG since CG's output errors were identified for them via highlighting. Further, while CG was provided with codes to assist them in deciphering the nature of the error, EG was not provided with such codes,

only with typographically enhanced input with which to compare and contrast their own past hypothetical conditional production. Again, the codes take some of the mental work out of the hands of the learner because the nature of the error is pre-identified. EG, on the other hand, had to do the mental work of comparing and contrasting and drawing conclusions about the exact nature of the output error before proceeding to correct the error.

The experimental treatment was consistently more reliable than was the control treatment in terms of allowing participants to make improvements from pretest to posttest 1 and to maintain those improvements from posttest 1 to the delayed posttest. However, it is important that some learners in CG made substantial gains while others in CG did not benefit from the treatment. Perhaps some of the participants in CG had had greater exposure to the target form in prior instruction than others, so that the control treatment triggered their memory of the form. After all, highlighting and coding as a form of corrective feedback is commonly used in writing instruction programs and is considered good pedagogy. However, the results of highlighting and coding corrective feedback were sporadic, while the results of matching output to input were more consistent.

CONCLUSIONS

Indirect negative evidence, as operationalized through comparing output to input appears to allow greater gains in result-clause production accuracy and for advanced and high intermediate level learners of English than does direct negative evidence, as operationalized through highlighting and coding errors. It would be interesting to conduct further research in other instructional contexts to further understand the role of indirect negative evidence as a corrective feedback tool. For example, would comparing output to input be feasible for younger learners or for learners of lower English proficiency? After all, there might be qualitative differences in the instruction that is useful for different levels and ages of learners. Keep in mind that all of the participants in this study had completed the Test of English as a Foreign Language with a score of at least 500, and they were all between the ages of 17 and 35.

Further research is needed to better understand the three steps involved in inductive inferencing in SLA as defined by Plough (1994): 1) scanning, 2) comparing, and 3) drawing conclusions. What is the most effective way to encourage students to scan their own output? How can we best set up situations for comparisons between learner output and input? Finally, how can we be sure that students draw the right conclusions about the L2 by scanning and comparing? The effectiveness of matching output to input in the classroom should be considered further in terms of the three steps of inductive inferencing and how it is exactly that learners engage in the three steps to promote their own L2 learning. The results of this study do, however, suggest that helping students identify problematic forms in the input, which they can compare to their own output is beneficial. It also suggests that less is more when correcting grammar in student writing, meaning that some of the work of error identification and correction should fall in the hands of students so that students can promote their own L2 learning.

ACKNOWLEDGMENTS

The author would like to thank Seiji Watanabe for his artwork as well as Patrick Bolger and Hang Du for help with piloting the materials for this study.

REFERENCES

Carroll, S. & Swain, M. (1993). An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition,15*(3), 357-386.

Carroll, S., Swain, M., & Roberge, Y. (1992). The role of feedback in adult second language acquisition: Error correction and morphological generalization. *Applied Psycholinguistics,13*(2), 173-189.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study on SL relativization. *Studies in Second Language Acquisition, 13*, 431-469.

Gass, S. (1997). *Input, interaction and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.

Gass, S. & Varonis, E. (1989). Incorporated repairs in NNS discourse. In M. Eisenstein (Ed.), *The dynamic interlanguage* (pp. 71-86). New York: Plenum.

Hammerly, H. (1987). The immersion program: Litmus test of second language acquisition through language communication. *The Modern Language Journal, 71*, 395-401.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings*. Rowley, MA: Newbury House.

Izumi, S. & Bigelow M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly,34*(2).

Izumi, S. Bigelow, M., Fujiwara, M. & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*, 421-452.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Lasnik, H. (1989). On certain substitutes for negative evidence. In R.J. Matthews & W. Demopoulos (Eds.), *Learnability and linguistic theory*, (pp. 89-105). Dordrecht: Kluwer.

Lightbown, P. & Spada, N. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition, 12*(4), 429-448.

Long, M. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics, 4*, 126-141.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, D. Coste, R. Ginsberg & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspectives*. Amsterdam: John Benjamins.

Nobuyushi, J. & Ellis, R. (1993). Focused communication tasks and second language acquisition. *English Language Teaching, 47*, 203-210.

Pica, T, Young, R. & Doughty, C. (1987). The impact of interaction on comprehension. TESOL Quarterly(21), 737-758.

Plough, I. (1994). *Indirect negative evidence, inductive inferencing, and second language acquisition*. In L. Eubank, L. Selinker, & M. Sharwood Smith (Eds.), *The current state of the interlanguage: Studies in honor of William E. Rutherford*, (pp. 89-106). Amsterdam: John Benjamins.

Schmidt, R. & Frota, S. (1986). Developing basic conversation ability in a second language: A case-study of an adult learner. In R. Day (Ed.) *Talking to learn: Conversation in second language acquisition*. Rowley, MA: Newbury House.

Shook, D.J. (1994). FL/L2 Reading, Grammatical Information, and the Input-to-Intake

Phenomenon. *Applied Language Learning*,*5*(2), 57-93.

Sorace, A. (1996). The use of acceptability judgements in second language acquisition research. In W.C. Ritchie & T.K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375-405). New York: Academic Press.

Steinbach, R., Bereiter, C., Burtis, J., & Bertrand, D. (1988). *Student response to feedback on written composition: The role of intentional learning*. Year-end report to the Ontario Ministry of Education, Toronto: Applied Cognitive Science Centre, The Ontario Institute for Studies in Education.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition*. Rowley, MA: Newbury House.

Swain, M. & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Languages Journal*,*82*(3), 320-331.

Tomasello, M. & Herron, C. (1988). Down the garden path: Inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics*,*9*(3), 337-246.

Tomasello, M. & Herron, C. (1989). Feedback for language transfer errors: The garden path technique. *Studies in Second Language Acquisition (11)*, 385-395.

White, J. (1998). Getting learners' attention: A typographical input enhancement study. In C. Doughty and J. Williams (Eds.), *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

APPENDIX A

Imagine you were one of the students on the committee *last month*. Which alternative did you choose regarding the qualities of the best candidate? Which qualities did you think the school should look for in a teacher? Please explain why you chose that alternative and argue against other alternatives.

Write a paragraph of at least 150 words on this topic.

Remember that you were on the committee <u>LAST MONTH</u>. Therefore, this is a <u>PAST EVENT</u>. Try to use at least 6 of the following forms as you write the sentences in the paragraph. Use each only once.

- If I…
- If there…
- I think that if…
- For example, if…
- However, if…
- If the committee…
- If…
- If the teacher…
- If the students…

APPENDIX B

Untrue in the Past Conditionals

1. If you had come to my house, I would have made you dinner.	**1. The true situation is that you did not come to my house.**
2. If they had taken an airplane, they would have arrived faster.	2. Actually, they did not take an airplane. They took a train.
3. If I had not eaten the fish, I would not have gotten sick.	3. In truth, I did eat the fish, and I got sick because of it.

APPENDIX C

Imagine that two weeks ago, George had asked you to plan his trip to Los Angeles. He wanted the cheapest possible transportation, but he also wanted to be comfortable. Most importantly, he wanted to arrive at the meeting on time. He could leave Tucson at 5:30 PM on Monday, and he needed to be in Los Angeles by 9:00 AM on Tuesday. It was important for him to be well rested for the meeting in Los Angeles at 9:00 AM. Look at the following alternatives and choose the best:

 *A bus. The bus left Tucson at 6:00 PM Monday, and arrived in Los Angeles at 8:00 AM on Tuesday. The bus was full. The price was $95.00.*

 *A plane. The plane left Tucson at 5:30 AM Tuesday morning, and arrived in Los Angeles at 7:00 AM Tuesday. The price was $450.00*

 An overnight train with a sleeping car. The train left Tucson at 6:30 PM Monday, and arrived in Los Angeles at 7:00 AM Tuesday. The price was $100.00.

 *A taxi. The taxi left Tucson at 5:30 PM, and arrived in Los Angeles at 5:00 AM. The taxi had a very large back seat. The price was $200.00.*

Write a paragraph of at least 150 words. **Please state the best alternative and argue for it while arguing against other alternatives.**

Remember that this happened TWO WEEKS AGO. *Use at least 6 of the following forms, and use each form only once*.

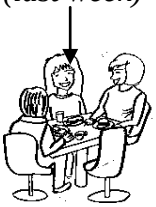
- If….
- If George….
- However, if….
- If the bus….
- If the train….
- If the plane….
- If the taxi….
- I think that if….
- If there…..
- For example, if….

APPENDIX D

Jane did not go out to lunch with her friends
last week because she did not have any money.


make money

(last week)


go out to lunch

If

APPENDIX E

1. Caroline will go to the beach if she had bought a bathing suit.

   G U

2. If Lupe had cooked dinner, she would not have been hungry.

   G U

APPENDIX F

Table 4. t-test Results.

Test	Group	Mean	SD	Significance at .05
If-Clause Pretest	EG CG	66.0818 70.5455	14.4087 13.2649	0.229
If-Clause Posttest 1	EG CG	91.1636 85.6818	15.1923 13.0109	0.187
If-Clause Delay Posttest	EG CG	91.1382 89.9545	15.4013 8.6236	0.413
Result-Clause Pretest	EG CG	69.8909 70.8364	26.8577 17.6247	0.461
Result-Clause Posttest 1	EG CG	99.1273 89.5818	2.2186 13.2599	0.019
Result-Clause Delay Posttest	EG CG	98.1491 85.5082	2.4421 17.3832	0.018
GJ Pretest	EG CG	70.3 72.6909	16.7320 21.0565	0.385
GJ Posttest 1	EG CG	89.8182 73.9091	6.1777 14.8085	0.003
GJ Delay Posttest	EG CG	89.5855 71.4	7.7414 19.7766	0.007

Table 5. Tukey's LSD Post Hoc Comparisons

Group and Test	Pretest Mean and SD	Posttest 1 Mean and SD	Delayed Posttest Mean and SD	Post Hoc Significance Pre to Post 1 at .05	Post Hoc Significance Post 1 to Delayed Post at .05
EG If-Clause	x = 66.0818 SD = 14.408	x = 91.1636 SD = 15.192	x = 91.1382 SD = 15.401	>0.001	0.997
EG Result-Clause	x = 69.8909 SD = 26.857	x = 99.1273 SD = 2.219	x = 98.1491 SD = 2.442	>0.001	0.884
EG Gramm Jugdement	x = 70.3 SD = 16.732	x = 89.8182 SD = 6.178	x = 89.5585 SD = 7.741	>0.001	0.962
CG If-Clause	x = 70.5455 SD = 13.264	x = 85.6818 SD = 13.019	x = 89.9545 SD = 8.624	0.001	0.942
CG Result-Clause	x = 70.8364 SD = 17.624	x = 89.582 SD = 13.259	x = 85.5082 SD = 17.741	0.091	0.096
CG Gramm Judgement	x = 72.6909 SD = 21.056	x = 73.9091 SD = 14.809	x = 71.4 SD = 19.777	0.684	0.739

APPENDIX G

Figure 1. If Clause Production Means

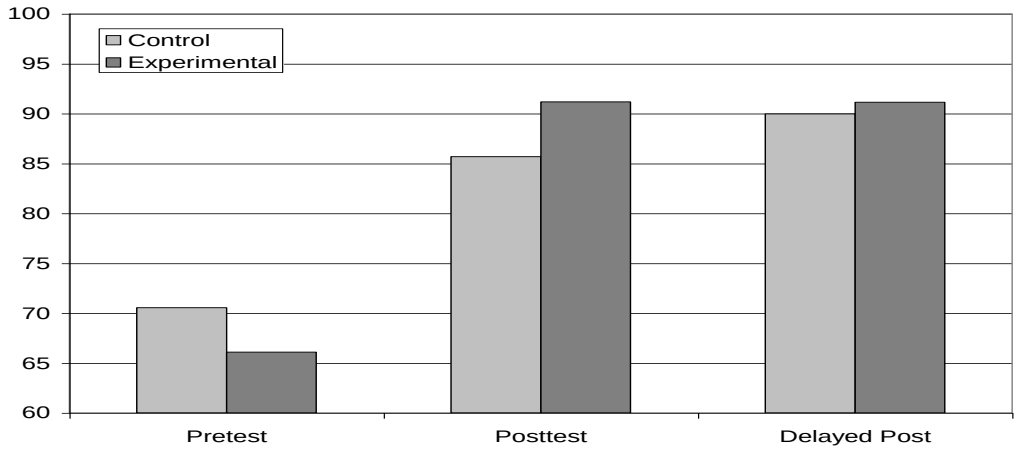


Figure 2. Result Clause Production Means

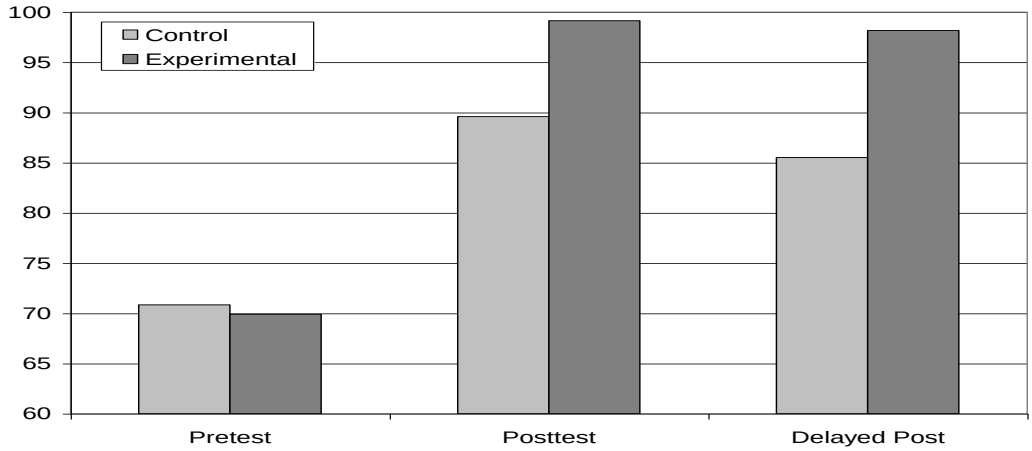


Figure 3. Grammaticality Judgement Means