SSLA, **21**, 421–452. Printed in the United States of America.

TESTING THE OUTPUT HYPOTHESIS

Effects of Output on Noticing and Second Language Acquisition

Shinichi Izumi, Martha Bigelow, Miho Fujiwara, and Sarah Fearnow

Georgetown University

This study addresses one of the functions of output proposed by Swain (1993, 1995, 1998). In particular, the activity of producing the target language may, under certain circumstances, prompt L2 learners to recognize some of their linguistic problems and bring to their attention something they need to discover about their L2. Two research questions were posed: (a) Does output promote noticing of linguistic form? and (b) Does output result in improved performance on the target form? In treatment phase 1, participants reconstructed a short passage after being exposed to it, followed by a second exposure to the same input material and a second reconstruction opportunity. In phase 2, participants wrote on given topics, followed by the presentation of a model written by a native speaker. Participants wrote a second time on the same topic. To test the noticing function of output, participants underlined parts of the sentences they thought were "particularly necessary" for subsequent (re)production. The control group was exposed to the same input materials but was not

The study reported in this paper was conducted as a project for a seminar in second language acquisition at Georgetown University. An earlier version of this paper was presented at SLRF '97 (Michigan State University) by Shinichi Izumi and Miho Fujiwara. The authors would like to thank Catherine Doughty for her assistance throughout the entire process of the study. Thanks are also due to Michael Long, Michael O'Neill, Deborah Mifflin, and Christine Glover for their many insightful suggestions and comments. We would like to thank the students who participated in this study and the ESL instructors who allowed us to pilot tasks and tests in their classes. Thanks are also due to Yoshiho Shibuya and Takae Tsujioka for drawing the pictures needed for the tests and tasks. Finally, the authors thank the anonymous *SSLA* reviewers for their many helpful suggestions.

Address correspondence to: Shinichi Izumi and Martha Bigelow, Linguistics Department, Georgetown University, Washington, DC 20057, e-mail: izumis@gusun.georgetown.edu and bigelowm@gusun.georgetown.edu.

© 1999 Cambridge University Press 0272-2631/99 $9.50

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

required to produce any output. Although phase 1 tasks resulted in noticing and immediate incorporation of the target form, the posttest performance failed to reveal their effects. In contrast, phase 2 tasks resulted in improvement on posttest 2. The results are discussed in terms of the efficacy of output in promoting noticing and learning and the conditions that may be required for output to be useful for SLA.

In Canadian immersion programs, teachers provide students with a rich source of comprehensible input. Decades of research on these programs have shown that they are among the most successful large-scale L2 programs in existence. However, over the years, research on these programs has accumulated evidence indicating that merely providing large amounts of comprehensible input is not sufficient for the L2 learner to attain a high level of L2 proficiency (Hammerly, 1987; Harley, 1993; Harley & Swain, 1978). Findings from such studies have shown that L2 learners in immersion classes often demonstrate weaknesses in grammatical accuracy, despite gaining high-level listening comprehension skills and communicative fluency.

Swain (1985) argued that one important reason why immersion learners display numerous grammatical errors in their L2 is that they actually engage in a small amount of language production. This characteristic of their language-learning environment may prevent them from going beyond their current level of L2 proficiency. Observational studies of interaction in French immersion classrooms have described typical examples of teacher-student interactions in which teachers did most of the talking and few of the utterances made by students were longer than a single clause (Allen, Swain, Harley, & Cummins, 1990; Swain, 1985). This kind of interaction permits students to operate successfully with their incomplete knowledge of the language because they are rarely pushed to be more accurate.

Observations such as these have led Swain to conclude that "comprehensible input," although invaluable to the acquisition process, is not sufficient for learners to fully develop their L2 proficiency. Specifically, she argues that what students need is not only comprehensible input but also "comprehensible output" if they are to be both fluent and accurate in the target language. The construct of comprehensible output posits that, when learners experience communication difficulties, they will be pushed into making their output more precise, coherent, and accurate. According to Swain, "producing the target language may be the trigger that forces the learner to pay attention to the means of expression needed in order to successfully convey his or her own intended meaning" (1985, p. 249).

Since the Output Hypothesis was first proposed, Swain (1993, 1995) has extended the scope of the hypothesis and identified the following three functions of output. First, output has a hypothesis-testing function. Producing output is potentially a way of testing one's hypothesis about the target lan-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

guage. Learners can test comprehensibility and linguistic well-formedness of their interlanguage (IL) against feedback obtained from their interlocutors. This function of output relates directly to the notion of comprehensible output that Swain claimed to be necessary for immersion students. By producing output, learners can test their hypotheses and, by being pushed in the process of negotiation of meaning, they can be more accurate in their production.

Second, output may have a metalinguistic function. It is claimed that "as learners reflect upon their own target language use, their output serves a metalinguistic function, enabling them to control and internalize linguistic knowledge" (Swain, 1995, p. 126). In other words, output may cause the learner to engage in more syntactic processing than is necessary for the comprehension of input. This syntactic processing may lead to modified or reprocessed output—a possible step toward language acquisition. Reflection on language may not necessarily involve metalinguistic terminology. "Metatalk" produced with or without metalinguistic terminology may, according to Swain, serve the function of deepening the students' awareness of forms, rules, and form-function relationships, provided that the contexts are such that the language produced by the learners serves some genuine communicative function.

Finally, output may serve a "noticing/triggering" (or consciousness-raising) function. That is to say:

> [I]n producing the target language (vocally or subvocally) learners may notice a gap between what they want to say and what they can say, leading them to recognize what they do not know, or know only partially, about the target language. In other words, under some circumstances, the activity of producing the target language may prompt second language learners to consciously recognize some of their linguistic problems; it may bring to their attention something they need to discover about their L2. (Swain, 1995, pp. 125–126)

This function of output is consistent with the claim made by Schmidt and Frota (1986) that "a second language learner will begin to acquire the target-like form if and only if it is present in comprehended input and 'noticed' in the normal sense of the word, that is consciously" (p. 311). The role of output here is proposed to facilitate the process of noticing of both problems in one's IL and the relevant features in the input. This noticing will then stimulate the processes of language acquisition.

It is important to note that Swain does not claim that output is the only source of L2 acquisition. Her claim, rather, is that output can, under some conditions, promote language acquisition by allowing learners to recognize problems in their IL capabilities. Recognition of problems may occur because of either internal feedback (the very process of producing output may prompt syntactic processing of language, not just semantic or pragmatic processing, which is often invoked in processing language for the purpose of comprehension) or external feedback (feedback obtained from interlocutors, teachers, etc., which informs the learners as to the well-formedness of their IL produc-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

tion). This recognition of problems is believed to prompt learners to generate alternatives by searching their knowledge, assessing alternatives, applying existing knowledge to known contexts or new contexts, and then using the resulting knowledge. Alternatively, recognition of problems through output may prompt learners to seek out relevant input with more focused attention (Swain & Lapkin, 1995).

Several studies have been conducted to test the first and second functions of output—namely, the hypothesis-testing function and the metalinguistic function. The results from the studies related to the hypothesis-testing function of output (Nobuyoshi & Ellis, 1993; Pica, 1988, 1992; Pica, Holliday, Lewis, & Morgenthaler, 1989; Takashima, 1994) show that learners often modify their output in response to the linguistic demands of NS signals of comprehension difficulty and also suggest that pushing learners to produce more comprehensible output may have a long-term effect. Results of the studies focusing on the metalinguistic function of output (Donato, 1994; LaPierre, 1994; Swain, 1995) lend some support to the claim that producing language and reflecting on it in an attempt to create meaning has positive effects on language-learning processes.

In contrast to the hypothesis-testing function and the metalinguistic function of output, there has been little research done to test the third hypothesized function of output—the noticing function. One of the few studies that has investigated this function of output is that of Swain and Lapkin (1995). They examined whether output would lead to conscious recognition of problems and whether this recognition would activate the kinds of cognitive processes that are believed to be conducive to SLA. Think-aloud protocols were used as a technique to elicit information about internal processes taking place in the learners. The learners in this study wrote an article for a newspaper about environmental problems and were asked to think aloud as they wrote their articles. From the think-aloud protocols, Swain and Lapkin abstracted what they called "language-related episodes," in which the learners either spoke about a language problem they encountered while writing and solved it correctly or incorrectly, or simply solved it correctly or incorrectly without having explicitly identified it as a problem. The results revealed that about 40% of these language-related episodes consisted of the students' paying attention to the grammatical form, which suggests that these learners do recognize problems in their IL during production. Swain and Lapkin (1995) also found that recognition of problems often triggers cognitive processes that have been implicated in second language learning: processes that either generate linguistic knowledge that is new for the learners or consolidate their existing knowledge. Based on these findings, Swain and Lapkin argue that output not only leads to noticing of the gaps in IL knowledge but also facilitates language acquisition by triggering various internal processes conducive to SLA.

Swain and Lapkin's (1995) study addressed whether output leads to recognition of problems in IL and whether certain internal processes that are be-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

lieved to be relevant to language acquisition can be triggered. However, this study did not address the question of whether the awareness of problems during production can prompt the learner to seek out subsequent input with more focused attention. That is, the question remains: Does learners' recognition of linguistic problems prompt them to notice relevant features if input is subsequently provided to them? The current study attempts to investigate this question by providing learners with opportunities for output, which are then followed by opportunities to receive relevant input, to see whether they would notice and learn the targeted feature in the input.

The general research questions to be investigated in this paper are the following:

1. Does learner output promote the noticing of linguistic form when relevant input is subsequently provided?
2. Does output result in the acquisition of the linguistic form?

RESEARCH HYPOTHESES

The Output Hypothesis generates the following predictions concerning noticing and SLA. The first hypothesis is derived from the first research question and is concerned with the noticing issue, and the second and third hypotheses originate in the second research question and are concerned with the acquisition issue.

Hypothesis 1

The experimental group, which is allowed to produce output, will show more noticing of the target grammatical form contained in the input than the control group, which does not produce output but instead answers true/false comprehension questions.

Hypothesis 2

The experimental group will indicate immediate incorporation of the target form in their output during the treatment tasks after being exposed to the input containing the target form.

Hypothesis 3

The experimental group will show greater gains than the control group in the accuracy of the use of the target form on posttests.

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

METHOD

Participants

Participants for the study ($N = 22$) came from two college-level academic writing classes in a community college ESL department. The participants came from the second level of three writing courses in an ESL program. Students were placed into the course by a departmentally administered placement test or achieved placement by passing the previous writing course. Most of the students in this program plan to enter a two- or four-year degree program after completing their required ESL courses. Participants in the study came from the following first language backgrounds: Spanish (8), Vietnamese (6), Amharic (2), Chinese (2), Turkish (1), Serbo-Croatian (1), Tagalog (1), and Somali (1). Ages of the participants ranged from 18 to 55. They varied in their amount of formal education, time in the United States (1.7–8.5 years), and in number of EFL or ESL courses taken. All but two spoke their native language at home. After the administration of the pretest, participants were assigned to an experimental group (EG, $n = 11$) and a control group (CG, $n = 11$) using a stratified random assignment procedure with the pretest results serving as the basis of stratification.

Target Form

The past hypothetical conditional in English was the target form for this study (e.g., *If Ann had traveled to Spain in '92, she would have seen the Olympics*). Celce-Murcia and Larsen-Freeman (1983) noted that conditional sentences in general, and hypothetical or counterfactual conditionals (imaginative conditionals) in particular, cause problems for many ESL learners. Conditional sentences consist of two clauses: an *if*-clause and a main clause. Therefore, the structure of such sentences is complex. Mastery of this structure requires a good grasp of the English tense system and the modal auxiliaries.

The pretest, administered prior to the treatment, showed that participants in the study had gaps in their ability to use the target form. Although their attempted use of the structure suggested that this form was not completely new to them, all of them still had problems using the form accurately. Some examples of the IL sentences produced by the participants are as follows:

(1) *If Kevin got up early in the morning, he would eat breakfast.*

(2) *If Jack travels to China, he would saw the Great Wall.*

(3) *If Kevin have caught the bus, he must have been on time at school.*

Such emergence of the target form may indicate a certain readiness to learn the form, because many of the participants already attempt to use it in a meaningful way.

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 1. Experimental sequence

Day	Activity	Time
1	Pretest	40 min
2	Treatment phase 1, Task A Paragraph about Christopher Reeve	30 min
3	Treatment phase 1, Task B Paragraph about Clinton's welfare bill	30 min
4	1st posttest	40 min
5	Treatment phase 2, Task A Essay on job offers	1 hr
6	Treatment phase 2, Task B Essay on scholarships	1 hr
7	2nd posttest	40 min

Experimental Design

The design of the study was a true experiment with pre- and posttests and a control group. As an overview, Table 1 shows that the experimental sequence was carried out over a period of approximately 1 month, totaling approximately 5 hours.

As can be seen from Table 1, the treatment consisted of two phases, with two tasks in each phase. Each treatment was followed by a posttest. In an attempt to minimize the test effects, the first phase of the treatment began a full week after the pretest, and the second phase of the treatment began a full week after the first posttest.

Treatment

All the materials used in the study were designed specifically for adult, immigrant, ESL learners such as those in this specific ESL program. The written modality was used throughout the experiment for both input (i.e., reading passages) and output (i.e., writing tasks). Piloting of all the materials in the study was carried out with 65 participants from other classes of the same level in the same program.

Both phases of the treatment were intended to give the EG participants the opportunity to produce output, hence providing them with the chance to notice a gap in their ILs (see Figure 1). Exposure to targetlike input was intended to allow participants to compare their ILs to the target forms in the input. The CG did not produce, but instead answered true/false comprehension questions.

In order to assess noticing of the target form, participants were required to underline the passage each time they were provided with the input. The directions given to the EG were to "underline the word, words, or parts of words that you feel are particularly necessary for your subsequent (re)production."[1] The CG was also required to underline their passage for comprehension (i.e.,

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

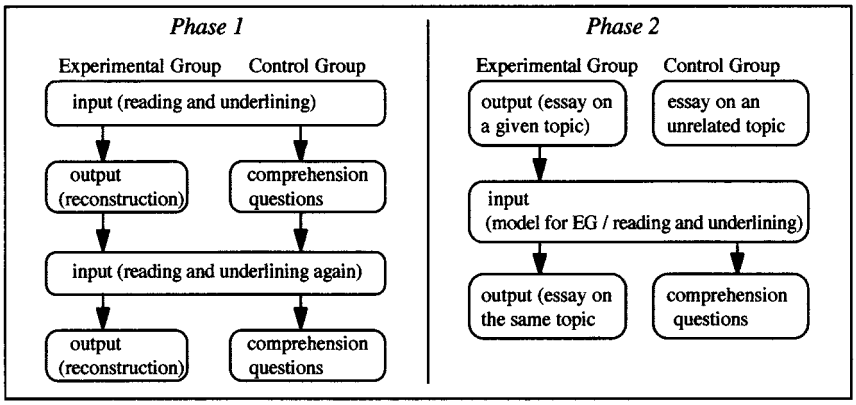


Figure 1. Instructional procedures.

to answer questions about the passage). Before participants carried out the tasks, the underlining portion of the activity was modeled for both groups. The researchers, using a passage that did not contain the target form, showed participants examples of underlining, to illustrate the options of underlining the "word, words, and parts of words" of the passage and to stress the importance of precise underlines. This was done to enhance participants' familiarity with the underlining procedure and the preciseness of this measurement of noticing. The entire process was role-played with one researcher giving the directions and another researcher carrying out the task. This was to help the participants realize that they would not be able to look at their underlining passage while reproducing or to check their first production while doing the second. It was necessary to clarify these two points to collect the most accurate data possible, as was discovered when the tasks were piloted with a similar class. Both the EG and the CG were informed in advance of the sequence of activities they would perform throughout the task.

In phase 1, EG participants were asked to read and underline a short passage with approximately 70% of its sentences containing the past hypothetical conditional, in order to immediately reconstruct the passage as accurately as possible (4 min).[2] After the passage was collected, the EG participants attempted to reconstruct it (6 min). Their reconstruction was collected. Then EG participants received a second exposure to the same input material and were asked again to read and underline the input (3 min). This was again collected, and EG participants were given a second reconstruction opportunity (5 min). CG participants were exposed to the same input for the same amount of time and instructed to read and underline, just as the EG were, except that they were reading and underlining to answer true/false comprehension questions.[3] A set of pictures was provided in the input passage sheets and output sheets in order to lighten the memory load, particularly for the EG. The phase 1 treatment was followed by posttest 1.

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Although these phase 1 reconstruction tasks have the advantage of controlling, to some extent, for the content and form that learners produce, one may argue that they may not provide learners with opportunities for more natural, generative use of the target form. To compensate for this disadvantage of the phase 1 tasks, phase 2 tasks were also provided. These tasks allowed the participants to use the target form in a freer, essay-writing format.

In the phase 2 tasks, experimental participants received a writing topic prompt, which was given to them before the writing task. This prompt did not use, but called for, the past hypothetical conditional. The participants were informed that they would write an in-class essay during the next class. In class, the participants wrote essays on the given topics (20 min), and the essays were then collected. The CG wrote on a topic that did not require the past conditional. The CG's writing on an unrelated topic was done for logistical and pedagogical purposes so that both groups would be doing a writing activity that day. This was followed by the presentation of a model essay written by a native speaker, of which approximately 80% of the input sentences contained the past hypothetical conditional (the essay was provided as a reading exercise for the CG). Participants read and underlined this input (10 min for the EG and 13 min for the CG). The CG was given a longer amount of time because they read an additional passage contextualizing the input model, much like the prompt the EG participants were given. EG participants were then asked to produce a second version of the essay on the same topic. CG participants answered true/false comprehension questions. This phase was followed by posttest 2.

As mentioned above, these guided essay-writing tasks have the advantage of providing opportunities for a more natural context in which to use the form. However, a disadvantage of these tasks may be that they pose some difficulty in controlling for the content and form that learners produce. Particularly, a difficulty may arise if the learners' IL production and the TL model essay are too different from each other in content or organization. Such differences may make it difficult for the learners to effectively compare their IL output attempts and the TL model. Although this problem may be partially solved by providing learners with some guidelines to follow in writing their essays, as was done in this study, we believed that providing the learners with both types of tasks (i.e., the phase 1 reconstruction tasks and the phase 2 guided essay-writing tasks) would maximize the possibility for the learners to notice the gap between their IL production and the TL input. For this reason, we decided to use both types of tasks in our study, with a posttest following each phase of the treatment in order to examine the relative effects of each type of treatment (see the conclusion section for the implications of this design in interpreting the results of the study).

Although these two phases of the treatment involved different kinds of tasks, it was assumed that the same internal processes would be involved in both cases. The output hypothesis predicts that the following internal processes would be triggered by each stage of the instructional sequence. When

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

receiving the input the first time, participants were expected to process input primarily for meaning. In the first (re)production, a semantic and syntactic analysis would be invoked, and participants would notice an inadequacy in their IL capacity. During the second input, it was expected that participants would process input for both meaning and form, which would result in noticing of the gap between their IL and the TL. In the second (re)production, participants were expected to attempt to incorporate what was noticed upon receiving the second round of input (see Appendixes A and B for an example of an input passage used for the tasks in each phase).

In order to obtain some information about what kinds of problems our participants had while producing output and what they paid attention to while processing input, we also conducted a brief interview with some randomly selected experimental-group participants after each task. In conducting interviews, great care was taken to ensure that the interview would not be obtrusive or heighten the participants' awareness of the target form or the purpose of the experiment. We did so by avoiding any remark on our part of the target form or attempt on our part to draw the participants' attention to form. Questions asked of the participants were the following: (a) What did you underline while reading the passage the first time? And why did you underline it? (b) Describe all difficulties or problems you had in (re)producing the passage/essay the first time. (c) What did you underline while reading the passage/model essay the second time? And why did you underline it? (d) What did you try to do differently when you (re)produced the passage/essay the second time? (Note that for phase 2 the first question was not relevant because it started out with essay writing.)

Testing Instruments

The pre- and posttests consisted of a grammaticality judgment test and a picture-cued production test. There were two forms of both tests, and a split-block design was used in administering them.[4] The grammaticality judgment test took about 15 minutes and the production test took about 20 minutes.

The grammaticality judgment test consisted of 28 sentences, of which 20 used the target form and 8 served as distracters. Of the 20 using the target form, 8 were correct and 12 incorrect, with 6 incorrect sentences starting with an *if*-clause and 6 with a main clause. An audiotape was created containing grammaticality judgment test instructions and the reading of the sentences. Participants both listened to and read each item in the test. A pause of 9 seconds was given between each item. The grammaticality judgment test was timed this way to increase the likelihood that the test would tap participants' intuition, rather than their analytical skills. Doing so also helped to assure that participants finished at the same time and completed all of the items. Participants were asked to judge a sentence correct or incorrect and, if incorrect, to circle the incorrect part of the sentence.

The production test consisted of eight picture-cued sentence items calling

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

for the past hypothetical conditional and four distracters. For every four sentences, an instruction page contextualized the pictures in the hypothetical past. Participants were instructed to use the verb written below each picture and combine the meaning of the two pictures into one sentence. The pictures were joined with an arrow to attempt to indicate that one would be the *if*-clause and the second would be the main clause. A prompt under each item (e.g., *If Anne*) was intended to encourage participants to use the target form. See Appendix C for sample test items.

Scoring and Analysis

The corpus of analysis in this study included data from underlining made during the treatment, written productions produced by the EG during the treatment, and the test results. The following is a description of how each data set was analyzed.

Underlining Scoring. Underlines were scored by counting total items underlined and then by obtaining a percentage of conditional-related underlined items. This procedure was used to balance individual variation arising from the differences in the absolute quantity of underlines made by the participants. Conditional-related words in this study were operationalized as the following words or morphemes contained in the conditional clauses (both the main and the *if*-clauses): modals (*would, could*), aspectual auxiliaries (*have, had*), copula in the past participle form (*been*), complementizer (*if*), and the past participle ending (*-ed* and *-en*).[5] These items were considered to serve grammatical functions and carry little semantic content as compared to major lexical categories such as nouns, verbs, and adjectives. In this sense, if these words were underlined, we took it as an indication of the participant's paying attention to the conditional forms.

Because underlines were made both before and after output (i.e., on the first and second inputs) in the phase 1 tasks, underlines in these tasks were examined to see whether there was any shift of noticing of the form from the first input to the second input for both groups, as well as to see whether the EG and the CG differed with respect to the proportion of their conditional-related underlines on both inputs. In the case of the phase 2 tasks, only the differences between the two groups in their proportion of conditional-related underlines were compared because each of these tasks involved only one round of input. The analysis excluded those participants who either failed to underline any words or underlined all words in the passage on the first input, the second input, or both. In order to check for interrater reliability, 30% of the underlines were scored by two researchers, which resulted in 98% agreement.

Production Scoring (Tests and Task Productions). The production data obtained for the EG from the treatment tasks were analyzed to examine whether the form noticed during the exposure to the input would be incorpo-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

rated into the participants' second (re)production attempts. This stage of immediate incorporation may be described as what Chaudron (1985) called a stage of "preliminary intake." The data obtained from tests were used to examine whether each phase of the treatment resulted in the acquisition of the target form. This stage was assumed to correspond to Chaudron's "final intake" stage.

The production test was scored using both a targetlike use (TLU) analysis and an IL analysis. The TLU analysis was designed to indicate targetlike performance of each clause. It was carried out by totaling the number of targetlike answers divided by the number of target items. This was done for each clause, to measure differences in progress for each clause. A combined score was also obtained. Incorrect morphology (e.g., *catched* for *caught*) was taken to be correct for the purpose of this study. Additionally, the use of a modal in the *if*-clause (e.g., *If Jack would have gone to Italy*) was also considered targetlike because this was acceptable to some native English speakers. (It should be noted, however, that this construction was never included in any of the materials in the study.) Production data from the tasks were analyzed in a similar fashion, except that only participants with a minimum of four clausal attempts in phase 1 and three attempts in phase 2 were included in the analysis. This was done to minimize distortion of percentage scores resulting from the inclusion of small denominators.

The IL scoring system was aimed to examine the specific IL features of the nontargetlike attempts made on the production tests throughout the study. This IL analysis was used to complement the strict coding system of the TLU analysis. Seven component features of the past hypothetical conditional were determined. The *if*-clause consists of the past tense, the perfect aspect, and the past participle form. The main clause consists of a modal, the past tense, the perfect aspect, and the past participle form. For each attempt at the form, plus or minus was given for each component feature to record suppliance (+) or lack of suppliance (–) of that component. It was decided that the modal could carry the past tense. Thus, modals including *would, could, should, might*, and *must* were coded as [+modal] and [+past]. Other modals such as *will, can, shall*, and *may* were coded as [+modal] but [–past]. It was also decided that any suppliance of the auxiliary *have*, in any form, along with another verb following it, would be coded [+perfect]. Therefore, if the auxiliary was not followed by a past participle, the phrase was still coded [+perfect] but [–past participle]. The possible error types include those in:

(4) *if*-clause
   a. [–past] errors; for example, *If Jack joins*
   b. [–perfect] errors; for example, *If Jack joined*
   c. [–past participle] errors; for example, *If Jack had go*

(5) main clause
   a. [–modal] errors; for example, *Jack had gone*
   b. [–past] errors; for example, *Jack joins*

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

c. [–perfect] errors; for example, *Jack would go*
d. [–past participle] errors; for example, *Jack would have go*

Every combination of these errors occurred.

Upon actual coding of the data, it was discovered that coding of the above seven features was not sufficient to account for all error types. Thus, an additional category, [+extra element], was added. For example, some participants marked the past tense twice in a clause, such as in the phrase *he would had gone*. Here, the past was marked in the modal, as well as in the auxiliary. For cases such as this, the extra-element category was added to penalize for this type of IL use of the structure. Using these grammatical features, IL scores were calculated for each participant by assigning 1 point when each of the target grammatical features was present in the learners' production and 0 points when it was absent. When a [+extra element] feature was present in the learners' productions, 1 point was deducted for each IL feature. To assure reliable scoring of the production data, interrater reliability was checked, yielding a 96% agreement rate.

Grammaticality Judgment Test. Because of the problems that grammaticality judgment tests can pose (see Sorace, 1996), such as guessing or the uncertainty of why an examinee judged a sentence ungrammatical, a conservative system of scoring was devised, as seen in Figure 2. Any item that was not judged, or if judged incorrect and no part of the sentence was circled, was excluded from the analysis (Decision 1). If a non-conditional-related (inappropriate) part of the sentence was circled, such items were also excluded from further analysis (Decision 2). Conditional-related words were defined here as the verb or modal parts of each clause. A circle indicating an error could include up to two irrelevant words that were adjacent to the target words (i.e., locus of the error), as long as it included the target words. This restraint was applied to exclude items with large circles, which made it difficult to ascertain exactly what the participant was focusing on. The judgment of the grammaticality of the sentence was then examined (Decision 3). If the response was "no" when the sentence was indeed ungrammatical, a further decision was made as to whether the circle was made in the right clause (Decision 4). We decided that this fourth decision was important in light of the fact that native speaker participants who completed the grammaticality judgment test during piloting stages of our study were highly consistent in judging which clause of the sentence contained an error. Finally, all "yes" responses were corrected for guessing to give yet a more conservative final score. The total points for each participant were then calculated by adding the scores corrected for guessing and the scores obtained from "no" responses, which were then divided by the total number of the items included for the analysis. Each participant, therefore, received a percentage score for this test.

Statistical Procedures. Because of the small sample size and the failure to obtain a normal distribution, the median was used as a measure of the central

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

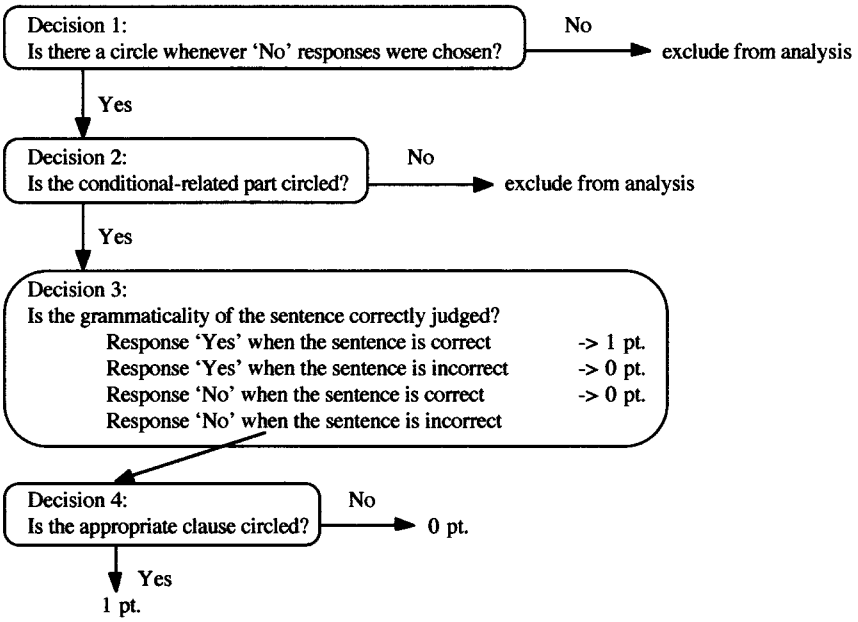


Figure 2. Scoring decision of grammaticality judgment test.

tendency of the groups in all the results reported. The Wilcoxon Signed Ranks test was used to assess the significance of the differences between the testing occasions, between the first and second inputs on the underlines, and between the first and second outputs for the experimental group during the treatment phases. The Mann-Whitney U test was used to assess the significance of the differences between groups on the underline scores and the test scores.

RESULTS

In this section, we first present the results of the underlines, which address the noticing issue. Then, the results of the data produced by the EG during the treatment are examined to address the issue of immediate incorporation of the target form. Finally, the test results, which address the acquisition issue, are presented.

The Results of Underlines: The Noticing Issue

Phase 1 Underlines. In the phase 1 tasks, underlines were made on the first input (i.e., before the first output) and the second input (i.e., after the output). Therefore, our interest here is in whether there was any increase of noticing of the form from the first input to the second input, as well as whether the EG

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 2. Median underline scores for experimental and control groups in phase 1, task A

	1st input		2nd input		
Group	*Mdn*	*SD*	*Mdn*	*SD*	Significance (1st → 2nd input)
EG ($n = 11$)	10	11	17	17	$p = .012$
CG ($n = 10$)	8	8	14	5	$p = .028$

Note. The scores are presented in percentages: the number of conditional-related words underlined divided by the total number of words underlined. The differences between the EG and CG on both the first input and the second input are not statistically significant.

Table 3. Median underline scores for experimental and control groups in phase 1, task B

	1st input		2nd input		
Group	*Mdn*	*SD*	*Mdn*	*SD*	Significance (1st → 2nd input)
EG ($n = 11$)	14	23	19	30	n.s.
CG ($n = 10$)	5	4	13	7	n.s.

Note. The scores are presented in percentages: the number of conditional-related words underlined divided by the total number of words underlined. The differences between the EG and CG on both the first input and the second input are not statistically significant.

and the CG differed with respect to their noticing of the conditional-related words. Tables 2 and 3 present the results of the underlines of tasks A and B in phase 1, respectively. For the EG, the group median increased significantly from the first input to the second input in task A. This significant increase indicates a shift of noticing from more content words noticed on the first input to more form words noticed on the second input, which could have been triggered by the output. Table 3 shows that the EG group median also increased in task B. However, this increase was not significant. This nonsignificance of the increase may be attributed to the fact that the median of the first input in task 2 was already high for this group. The CG also shows a significant increase from the first input to the second input in task A. There was also an increase in the proportion of the conditional-related words underlined for this group from the first input to the second input in task 2. However, the difference was not statistically significant. The differences between the EG and the CG on either the first input or the second input in both tasks were not statistically significant either.

Phase 2 Underlines. In each task of phase 2, the input passage was given to both groups only once. Thus, our focus here is on the comparison of the underline scores obtained by the EG and the CG. Tables 4 and 5 show the results of the phase 2 underline scores for both groups. As can be seen in Table 4, the EG obtained higher underline scores in task A than did the CG. This was reversed for task B (see Table 5), where the CG obtained slightly

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 4. Median underline scores for experimental and control groups in phase 2, task A

Group	Input		Significance (EG vs. CG)
	Mdn	*SD*	
EG ($n = 10$)	12	25	n.s.
CG ($n = 9$)	4	7	

Note. The scores are presented in percentages: the number of conditional-related words underlined divided by the total number of words underlined.

Table 5. Median underline scores for experimental and control groups in phase 2, task B

Group	Input		Significance (EG vs. CG)
	Mdn	*SD*	
EG ($n = 11$)	4	31	n.s.
CG ($n = 11$)	8	7	n.s.

Note. The scores are presented in percentages: the number of conditional-related words underlined divided by the total number of words underlined.

higher underline scores than did the EG. In either task, however, the differences between the two groups did not reach statistical significance. It should also be noted that the EG always had larger standard deviations than the CG in their underlining scores, even when the median scores for both groups are similar (see Tables 2–5). This suggests that the output treatment (for the EG) is, for some reason, more susceptible to individual learner variables than the input treatment (for the CG). Such potential susceptibility to individual learner characteristics appears to be stronger in the phase 2 essay-writing tasks than in the phase 1 reconstruction tasks.

Task Results: The Immediate Incorporation Issue

Phase 1 Reconstruction. Tables 6 and 7 present the results of the reconstruction by the EG during treatment phase 1. As can be seen in Table 6, the median percentage of the correct past conditional sentences increased significantly from the first reconstruction to the second reconstruction in task A. Clausal analysis of the production indicates that a significant increase of the correct use of the target form was observed in the *if*-clause but not in the main clause. These results indicate that there was an immediate incorporation of the target form by the EG in task A. Table 7 also shows an increase in the percentage of correctly formed conditional sentences from the first to the sec-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 6. Median reconstruction scores by experimental group in phase 1, task A ($n = 11$)

Group	1st reconstruction		2nd reconstruction		Significance (1st → 2nd reconstruction)
	Mdn	*SD*	*Mdn*	*SD*	
Main and *if*-clauses combined	0	8	25	26	$p = .017$
if-clause	0	16	33	32	$p = .016$
Main clause	0	0	0	23	n.s.

Note. The scores are presented in percentages: the number of correctly formulated past conditional sentences divided by the total number of sentences attempted.

Table 7. Median reconstruction scores by experimental group in phase 1, task B ($n = 11$)

Group	1st reconstruction		2nd reconstruction		Significance (1st → 2nd reconstruction)
	Mdn	*SD*	*Mdn*	*SD*	
Main and *if*-clauses combined	25	29	40	26	n.s.
if-clause	50	35	75	34	n.s.
Main clause	0	34	0	34	n.s.

Note. The scores are presented in percentages: the number of correctly formulated past conditional sentences divided by the total number of sentences attempted.

Table 8. Median production scores by experimental group in phase 2, task A ($n = 11$)

Group	1st production		2nd production		Significance (1st → 2nd production)
	Mdn	*SD*	*Mdn*	*SD*	
Main and *if*-clauses combined	0	31	19	32	$p = .017$
if-clause	0	32	27	35	$p = .016$
Main clause	0	32	10	32	n.s.

Note. The scores are presented in percentages: the number of correctly formulated past conditional sentences divided by the total number of sentences attempted.

ond reconstruction in task 2. However, the difference was not statistically significant. This might be due to the fact that, in task B, the participants already started out with a higher percentage of correct forms in the first reconstruction, benefiting from previous activities in task A.

Phase 2 Production. Tables 8 and 9 display the results of the production by the EG during treatment phase 2. Table 8 shows that the percentage of the correct target sentences increases from the first production to the second production attempts in task A. Again, a significant increase was observed in

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 9. Median production scores by experimental group in phase 2, task B ($n = 11$)

Group	1st production		2nd production		Significance (1st → 2nd production)
	Mdn	*SD*	*Mdn*	*SD*	
Main and *if*-clauses combined	0	36	13	36	n.s.
if-clause	0	39	8	43	n.s.
Main clause	0	34	6	38	n.s.

Note. The scores are presented in percentages: the number of correctly formulated past conditional sentences divided by the total number of sentences attempted.

Table 10. Grammaticality judgment test scores

Group	Pretest		1st posttest		2nd posttest		Significance		
	Mdn	*SD*	*Mdn*	*SD*	*Mdn*	*SD*	Pre → Post 1	Post 1 → Post 2	Pre → Post 2
EG ($n = 10$)	34	17	18	34	37	26	n.s.	n.s.	n.s.
CG ($n = 10$)	20	28	55	34	37	35	$p = .038$	n.s.	n.s.

Note. The scores are presented in percentages. One participant from each group was excluded from the GJT analysis because of their failure to conduct the tests as administered. The differences between the EG and CG on all three sessions of the tests are not statistically significant.

the case of the *if*-clause but not in the case of the main clause. Table 9 shows that there was an increase in the correct use of the conditional form from the first production to the second production in Task B. However, this was not statistically significant. Although some improvement can be observed for both the *if*-clause and the main clause, neither reached statistical significance.[6]

Test Results: The Acquisition Issue

Recall that the participants in this study were assigned to the EG and the CG based on the stratified random-assignment procedure with the pretest results used as the basis of stratification. In order to confirm the pretreatment equivalence of the two groups on each of the tests, the data for each test were submitted to the Mann-Whitney U test. The results indicated nonsignificance between the two groups on either of the tests, which suggests that the two groups were indeed equivalent on both tests prior to the treatment.

Grammaticality Judgment Test. Table 10 and Figure 3 show the results of the grammaticality judgment test (GJT). For the EG, there was a decrease in scores from the pretest to the first posttest. In contrast, the scores showed an increase from the first posttest to the second posttest. These differences, however, were not statistically significant. The difference between the pretest and the second posttest was not significant either. For the CG, there was an

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

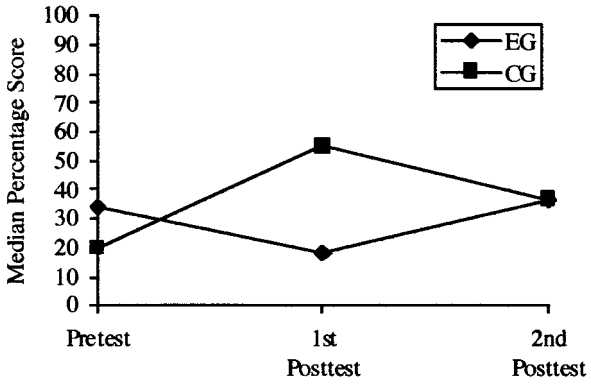


Figure 3. Grammaticality judgment test median scores (% scores).

increase in scores from the pretest to the first posttest, and the scores decreased from the first to the second posttests. The increase from the pretest to the first posttest was statistically significant, but the difference between the first posttest scores and the second posttest scores, as well as the difference between the pretest scores and the second posttest scores, was not significant. Although the difference between the EG and the CG on the first posttest seems large, this difference was not statistically significant. The differences between the two groups on the other two sessions of the tests were not statistically significant.

To confirm the results of the GJT shown in Table 10 and Figure 3, a gain score analysis was conducted. This analysis showed that, whereas the EG exhibited a minus gain from the pretest to the first posttest (–9%), the CG showed some positive gains (14%). This pattern was reversed for the gains made from the first posttest to the second posttest: The EG showed some gains (8%), but the CG showed some loss (–6%). The difference in the gain scores from the pretest to the first posttest between the two groups, however, was not statistically significant. The difference in the gain scores from the first posttest to the second posttest did not attain statistical significance. Thus, the superiority of the EG over the CG on the acquisition of the target form was not observed on the GJT.

Picture-Cued Production Test. Table 11 and Figure 4 present the results of the picture-cued production test, using TLU scoring, as explained above. The EG exhibited a steady increase in the score of correctly formulated conditional sentences from the pretest to the first posttest and from the first posttest to the second posttest. Although the first increase from the pretest to the first posttest failed to reach statistical significance, the increases from the first to the second posttests and also from the pretest to the second posttest were significant. A significant difference was observed in the case of the *if*-clause but not in the case of the main clause.[7] For the CG, the median score

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 11. Picture-cued Production Test: TLU median scores

Group	Pretest		1st posttest		2nd posttest		Significance		
	Mdn	*SD*	*Mdn*	*SD*	*Mdn*	*SD*	Pre → Post 1	Post 1 → Post 2	Pre → Post 2
EG ($n = 11$)									
Main and *if*-clauses combined	0	5	3	4	5	5	n.s.	$p = .017$	$p = .011$
if-clause	0	3	3	3	5	4	n.s.	$p = .034$	$p = .027$
Main clause	0	2	0	2	0	3	n.s.	n.s.	n.s.
CG ($n = 11$)									
Main and *if*-clauses combined	2	4	8	7	1	5	$p = .028$	n.s.	n.s.
if-clause	0	3	8	4	0	4	n.s.	n.s.	n.s.
Main clause	0	3	0	4	1	3	$p = .043$	n.s.	n.s.

Note. The differences between the EG and CG on all three sessions of the tests (on the *if*-clause, main clause, or combined scores) are not statistically significant.

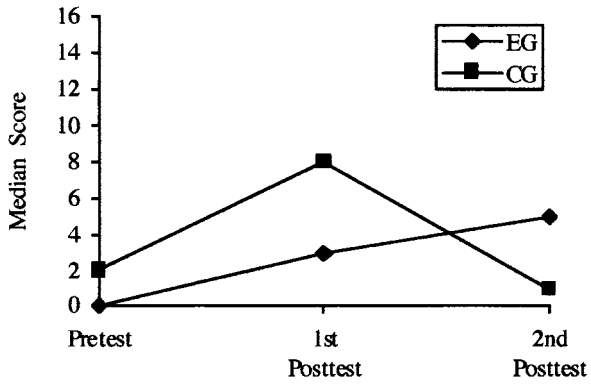


Figure 4. Picture-cued production test TLU median scores: *if*-clauses and main clauses combined (max. = 16).

of correctly formulated sentences increased from the pretest to the first posttest, but the score dropped on the second posttest to the same level as on the pretest. Although the difference between the pretest and the first posttest for this group was significant, the differences between the first posttest and the second posttest and between the pretest and the second posttest were not significant.[8]

The IL scoring procedure of the production test showed some different results from that of the TLU scoring. The main difference was that both the CG and the EG improved significantly on the main clause from the first posttest to the second posttest and from the pretest to the second posttest, although the TLU scoring procedure did not show this difference. Another difference

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Table 12. Picture-cued Production Test: IL median scores

Group	Pretest		1st posttest		2nd posttest		Significance		
	Mdn	*SD*	*Mdn*	*SD*	*Mdn*	*SD*	Pre → Post 1	Post 1 → Post 2	Pre → Post 2
EG ($n = 11$)									
Main and *if*-clauses combined	25	10	26	9	38	11	n.s.	$p = .007$	$p = .010$
if-clause	8	7	12	5	20	8	n.s.	$p = .012$	($p = .052$)
Main clause	15	6	16	6	16	6	n.s.	$p = .007$	$p = .012$
CG ($n = 11$)									
Main and *if*-clauses combined	30	11	36	14	31	15	n.s.	n.s.	n.s.
if-clause	8	8	17	8	8	8	n.s.	n.s.	n.s.
Main clause	18	6	24	8	24	8	n.s.	$p = .032$	$p = .008$

Note. The differences between the EG and CG on all three sessions of the tests (on the *if*-clause, main clause, or combined scores) are not statistically significant. There is a statistical difference between the EG and the CG in gain scores between the second and third session of the test.

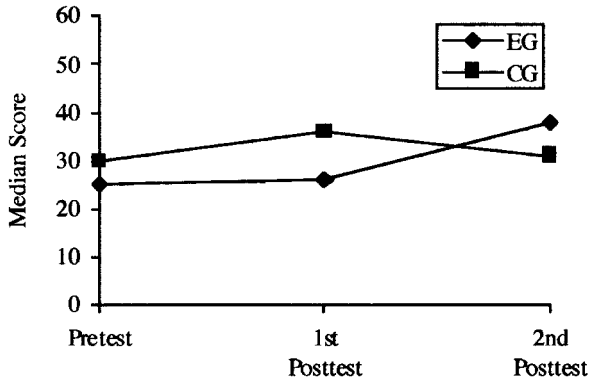


Figure 5. Picture-cued production test IL median scores: *if*-clauses and main clauses combined (max. = 56).

was that, according to the TLU scoring procedure, the CG gained significantly in the main clause from the pretest to the first posttest, whereas the IL scoring did not show this difference. Overall, the CG did not maintain its statistically significant difference between the pretest and the first posttest as it did in the TLU scoring. Table 12 and Figure 5 show these results.

As in the case of the grammaticality judgment test, a gain score analysis was conducted for the picture-cued production test in order to confirm the results obtained above. The EG exhibited a zero gain, and the CG showed a small positive gain (6%) from the pretest to the first posttest. This difference between the two groups was not statistically significant. Gains from the first

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

to the second posttests showed the reverse pattern: The EG obtained a positive gain (13%), but the CG had a zero gain. The difference between the two groups on these second gains was statistically significant. This indicates that the EG outperformed the CG after the phase 2 treatment on this production measure.

DISCUSSION

To summarize the major findings of this study in terms of the research hypotheses set at the outset, the first hypothesis predicting greater noticing of the target form for the EG than for the CG was not confirmed in this study. This was because not only the EG, but also the CG, increased significantly in their noticing of the target form on the second input in the first task of phase 1. Furthermore, the two groups did not differ significantly in their proportion of conditional-related underlines on any of the tasks. The unique effects of output in promoting noticing of the form, therefore, were not confirmed in this study. The second research hypothesis predicted that the EG would indicate immediate incorporation of the target form in their output during the treatment tasks. This hypothesis was partially confirmed, in that the EG participants were able to incorporate the target form presented in the input in their immediately subsequent production. A significant increase in accuracy scores was obtained for the first tasks in both phases of the treatment. The third hypothesis predicted that the EG would show greater gains in the accuracy of the use of the target form than the CG. This hypothesis was also partially confirmed. Although the phase 1 treatment failed to enable the EG to improve on the first posttest, the phase 2 treatment led to significant improvement of this group on the production test. The gain score analysis of the production test also revealed that the EG made significantly larger gains after the second phase of the treatment than did the CG. The results of the grammaticality judgment test, however, did not show significant gains for the EG, although they did for the CG from the pretest to the first posttest (but not from the pretest to the second posttest).

It was an unexpected result that not only the EG but also the CG showed significant increases in their conditional-related underlines in a phase 1 task when they were provided with the input passage a second time. One possible explanation for this result is the lightened processing load experienced by both groups at the second exposure to the input. Based on the notion of the language learner as a limited capacity processor, VanPatten (1995) argued that learners process input for meaning before they process it for form. More specifically, learners process content words in the input before less meaningful, grammatical morphology. And it is only when they can process informational or communicative content at little cost to attention that they are able to process form that is not meaningful. Applying VanPatten's model of input processing to the current study, it may be that participants in both groups processed meaningful elements in the input while reading the input passage

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

the first time. This may account for their relatively low rate of conditional-related underlines on their first input. When they were exposed to the same input passage the second time, their attentional capacity was freed up to process less meaningful and more grammatical elements such as modals like *would* and auxiliaries like *have* and *had*.

This explanation of freed processing capacity, however, does not necessarily explain why the participants' attention was drawn to the conditional form in the second input. Their attention could have been drawn to other content words that may not have been fully processed during the first reading. A closer analysis of the input passages and the comprehension questions provided to the CG reveals that about half of the comprehension questions in both phase 1 tasks could not be answered correctly without comprehending the counterfactual meaning of the conditional sentences in the input. For example, in order to judge whether *Reeve continues riding horses* is true or false, the learner has to know that the following sentence that appeared in the passage is counterfactual: *If he* [*Reeve*] *had stayed healthy, he would have continued riding horses*. In other words, to answer questions such as this one, attention to and careful processing of the conditional form in the input may have been required. Encountering such questions after the first input may, thus, have prompted the CG learners to pay closer attention to the form in the second input, leading to increased conditional-related underlines on the second input for this group.[9]

Although the above explanation may hold true for the CG, it is reasonable to assume that the increase in the conditional-related underlines by the EG was promoted by the output, because an output activity was what intervened between the presentation of the two inputs. As the output hypothesis predicts, producing output may have enhanced the learners' awareness of the inadequacies of their IL, which then prompted them to seek the relevant input at the subsequent input exposure. In comparing the underline scores for the EG and CG, it is noteworthy that, although the differences between the two groups on the underline scores were not statistically significant, a general trend was that the EG tended to obtain a higher proportion of underlines on the conditional-related words. Although the unique effects of output in promoting noticing of the form were not confirmed in this study, it is possible that future research conducted with a larger sample size will demonstrate the unique role played by output in facilitating noticing.

With regard to the acquisition issue, noticing of the form and immediate incorporation of the target form observed for the EG in the phase 1 treatment did not lead to their improved posttest performance. These results suggest that noticing and immediate incorporation may not be so directly related to subsequent learning. In this regard, Chaudron (1985) argued that the initial stages of perception of input (i.e., noticing) and immediate incorporation of the linguistic feature into one's IL do not necessarily guarantee that such an item will be incorporated into one's IL grammar. He argued that a series of stages is involved by which learners fully integrate and incorporate the lin-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

guistic information in input into their developing grammars (i.e., converting "preliminary intake" into "final intake"). Schmidt also noted that noticing may not be sufficient for learning to take place. Citing his own learning of Brazilian Portuguese as a second language (reported in Schmidt & Frota, 1986), he stated that:

> there are cases in which my interlocutor used a particular grammatical form in a question on tape and I repeated it in my answer, suggesting that I noticed its use momentarily, but never used it again. In such cases, a plausible suggestion is that I did not process the form sufficiently deeply to ensure retention. (1990, p. 141)

The same explanation might apply to the results obtained from the phase 1 tasks. The fact that the EG learners did not show improvement on the first posttest despite their indication of both noticing and immediate incorporation during the phase 1 tasks suggests that these learners did not process the input (i.e., the target form) sufficiently.

What prevented the EG participants from processing the form sufficiently? The intervening variable seems to be the task demands placed on these participants. That is, the EG was under the heavy cognitive demand of remembering and reproducing the given passages accurately in the phase 1 tasks. The high memory load demanded by the tasks was probably too taxing for these participants and did not allow them to conduct careful analysis of the form. It is likely that many of the experimental participants merely stored the form in their short-term memory and, upon reproducing it, lost it altogether (or at least no further analysis of the form was conducted). This may in fact be a reasonable strategy because, as far as the task is concerned, reconstruction of the sentences would suffice. The reconstruction requirement of this treatment phase, therefore, may have blocked further processing of the target form.[10] If this explanation turns out to be a valid one, it would point to the importance of considering not only how the task facilitates noticing and immediate intake of the form but also how further processing of the form beyond noticing and immediate intake can be encouraged by the task.[11]

In comparison with the EG, the CG was under relatively low cognitive demand because they only read the passage to answer comprehension questions. This relatively light cognitive demand for the CG presumably allowed them to process the form features in the input more deeply than did the EG. As stated above, the possible requirement of some comprehension questions to attend to the conditional form in the passage may have prompted them to actively process the form while being exposed to the second input. Additionally, it is possible that the pictures provided under the passages, which were used to ease memory load particularly for the EG, were also used by the CG participants to help establish or reinforce the form-meaning mapping.[12] Therefore, both comprehension questions and the pictures may have facilitated the CG participants' enhanced attention to the form and successful establishment

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

or reinforcement of the form-meaning connection of the past hypothetical conditional.

Interestingly, the pattern of improvement was reversed for phase 2. On the second posttest conducted after the phase 2 treatment, the EG, but not the CG, showed significant improvement. In this phase of the treatment, the EG was given the topic one class in advance, wrote on the topic later in class, read a model essay, and wrote on the same topic once again. In this sequence of activities, reading the model essay may not have been especially taxing for these participants because they had already been thinking about the topic and had written on it before they were provided with the model essay. This was a favorable condition for the EG. In contrast, the CG participants did not have a chance to think or write about the topic beforehand because they wrote on an unrelated topic prior to receiving the model essay. This condition required the CG to process the input passage without their schema on the topic activated to aid their comprehension of the passage, which may have made the task more cognitively demanding. Furthermore, the passage that they read was longer than that provided to the EG because both the context (i.e., topics and prompts for writing for the EG) and the model were included in their passage. The combination of these factors may have added to the increase in the cognitive demand placed on the CG participants and may, at least partially, account for why the CG did not benefit as much from the second treatment.[13]

There still remains a question regarding why the grammaticality judgment test failed to reveal significant effects of the phase 2 treatment for the EG. With careful scoring procedures for this test as outlined in the analysis section, we have a certain amount of confidence in the results we obtained from this test. Nevertheless, the fact that it did not show significant results in contrast to the production test might be attributed to lack of transferability of knowledge gained by tasks to this type of test. In future research, it is recommended that other types of tests that tap the learner's intuition and comprehension abilities be used for investigation of the acquisition issue.

Although the above-mentioned processing-related factors may have given the phase 2 tasks some advantages over the phase 1 tasks, it is interesting that the phase 2 tasks tended to produce much larger standard deviations (in relation to the median scores) in the underlining and production scores for the EG (but not for the CG) than the phase 1 reconstruction tasks did (see tables 2–9). Notice that phase 2 tasks allowed for greater freedom in what the participants could produce, whereas phase 1 reconstruction tasks are much more tightly controlled in this respect (see the section on treatment). These characteristics of the phase 2 tasks, although perhaps providing more meaningful contexts for production, might have made it more difficult for the EG participants to directly compare their IL output with the model input than in the phase 1 tasks. A closer examination of the underlines made by the EG participants on the phase 2 tasks revealed that different participants appeared to have focused on many different aspects of the input: Some learners under-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

lined more grammatical forms (i.e., the conditional form), which suggests that their attention was focused on these aspects of input, whereas other learners tended to underline organizational words or phrases such as *on the other hand* and *for example,* or useful content words or phrases such as *good for the academic record* and *fair student selection,* which indicates that these participants' attention was drawn more to these features of the model essay.

An analysis of the results of the interview conducted with some of the experimental group participants after each task confirmed this observation. The interview data suggested that, although output prompted some learners to recognize problems in their IL and to pay greater attention to form when provided with the input, individual learners varied greatly as to what they paid attention to when producing output and when processing input. Although few participants in this study appeared to pay attention to the grammatical form before output, at least some participants paid attention to it after output. Others, however, seemed to pay little attention to the grammatical aspects of the given passage even after output. These participants appeared to have been more preoccupied with generating or organizing their ideas than with finding inadequacy in their IL grammar during output. It appears, then, that these participants' awareness of the grammatical form was not particularly heightened even after output.

One may argue that such variability in learner focus does not necessarily have to be seen negatively because it may indicate the effects of output in drawing the learners' attention to what they perceived to be problematic aspects of their IL production. However, it poses a methodological challenge for the researchers and a pedagogical challenge for the teachers, if they wish the learners to focus on certain specific aspects of the input as was attempted in this study. Future research, therefore, needs to consider ways to adequately control for the variability in what learners focus on in the input after their output attempts.

CONCLUSION

The purpose of this study was to investigate the effects of output on noticing and second language acquisition. The results of this study provide partial support for the output hypothesis. In particular, although phase 1 tasks resulted in noticing and immediate incorporation of the target form, the posttest performance failed to reveal any effects. The phase 2 tasks, in contrast, resulted in improvement on the subsequent posttest.

We have discussed two implications of these results for the noticing function of output. First, in order for output to lead to final intake, it would be necessary for the task not to place heavy cognitive demands on the learners as they engage in output activities. Second, although guided essay-writing tasks used in this study allowed, to some extent, for the learners to have potentially more meaningful and generative use of the target grammatical form, they appear to have caused difficulties for them to effectively compare the IL

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

output and TL input, with respect to the targeted form. It would, therefore, be useful in future research to devise tasks that could adequately control for the learners' focus of attention, while ensuring that their processing capacity is not overloaded. These requirements are especially important if output is to help draw the learners' attention to particular grammatical forms.

Finally, some cautions are in order. First, it must be pointed out that because the phase 2 tasks followed the phase 1 tasks, it is unclear what kind of priming effects the phase 1 tasks had on the results of phase 2. It is entirely possible that the obtained effect of output in this study is a cumulative one and may not be attributable solely to the phase 2 tasks. This issue is addressed in our second study, in which we replicated this original study and examined the ordering effects of the tasks (Izumi & Bigelow, 1998). Second, it is also possible that having to underline the input passage while processing it might have enhanced noticing of the target form. In other words, accompanied with the input flood, underlining may have worked as some form of focused input flooding. To adequately address whether and to what extent underlining had an effect on noticing and learning a linguistic form, it is necessary to compare a group that is required to underline while processing input with a group that is not. Again, this is an issue to be investigated in future research. Finally, we admit that the generalizability of the results obtained in this study is severely constrained by the small sample size of the study, and more studies are needed to substantiate any claims made here.

(*Received 15 April 1998*)

NOTES

1. Because underlining is likely to involve at least a minimum level of awareness, this measure is considered to tap noticing in Schmidt's sense (1990, 1995) and not detection as discussed by Tomlin and Villa (1994).

2. This requirement of exact reconstruction was important because, as in the dictogloss task reported by Kowal and Swain (1997), it would help avoid the problem of the learners relying solely on their already-established IL knowledge—that is, the requirement aimed to implicitly push the learners for accurate production.

3. An anonymous reviewer pointed out that the CG participants might have had the expectation that they would have to write sentence-level responses to the comprehension questions and, therefore, read and underlined the text as though they were preparing for this task rather than the true/false questions they were actually given. Because we did not tell the control group participants the form that the comprehension questions would take, this reviewer's point remains a possibility. It should be noted, however, that this explanation most likely applies only to the first task in phase 1 because we believe that the control group participants developed an expectation of the task type after their first experience with it.

4. Both forms of each test were piloted on an equivalent pool of participants. Correlating the two forms of the grammaticality judgment test gave a Pearson r of .89, and correlation between the two forms of the production test yielded a Pearson r of .90. Both were significant at $p < .001$.

5. The directions of the task stated that not just words, but also parts of words, could be underlined. Because this point was emphasized through modeling, some participants underlined only the endings of verbs. Such underlines, as opposed to underlines of the whole verb, were taken to indicate the participant's attention to the form and, as such, counted as conditional-form-related underlines. No other bound morphemes were underlined by any of the participants.

6. The EG reconstructed the given passages during the phase 1 tasks and wrote guided essays during the phase 2 tasks, and the CG answered comprehension questions based on the input pas-

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

sages. The analysis of the answers to the comprehension questions by the CG participants indicated that they scored a mean of about 8 (of 9 questions) for the phase 1 tasks, and a mean of about 13 (of 15 questions) for the phase 2 tasks. These results suggest that the CG participants comprehended the given input passages fairly well.

7. In general, our IL analyses indicate that the EG and the CG benefited from their treatments somewhat differently. Such a difference in gains for the two groups may be attributed to the influence of the production requirement made for the EG participants. An analysis of the EG participants' production during both phases of tasks indicated that the EG participants started their sentences predominantly with the *if*-clause. This was perhaps, at least in part, caused by the frequency in input that contained more sentences starting with the *if*-clause than with the main clause. Seen from the perspective of output processing, however, it can also be argued that the EG participants simply followed temporal ordering in their production of the conditional sentences (i.e., in the time sequence, the condition expressed by the *if*-clause comes before the results expressed by the main clause). This production bias may have prompted these learners to pay closer attention to the *if*-clause, possibly more so than to the main clause, in their subsequent input processing. On the other hand, the failure of the CG participants to improve on the *if*-clause in the production test may suggest that, without an output requirement, these learners may not have paid as careful attention to the *if*-clause as the EG participants when processing the input.

8. An analysis of the individual data revealed that, although the increase of the score for the CG from the pretest to the first posttest was caused by the improvement by 8 participants (out of 11), the decrease of the score from the first posttest to the second posttest for this group was caused by 3 participants, all of whom showed some improvement on the first posttest. The other participants in this group either maintained the same scores on the second posttest as on the first posttest (n = 5) or improved from the first to the second posttests (n = 3). This individual variation is reflected in the rather large standard deviation for this group, as seen in Table 11.

9. In future research, in order to prevent inadvertently drawing the learners' attention to form, comprehension questions may need to be replaced by extension questions that ask the participants for their opinions about the input passage. We are indebted to C. Chaudron (personal communication, November, 1997) for this suggestion. Alternatively, one may use L1 recall procedures to assess learners' comprehension. We are indebted to B. VanPatten (personal communication, October, 1997) for this idea.

10. It should be noted, however, that this argument does not necessarily entail that the reconstruction task cannot enhance learning. The problem with the reconstruction tasks used in this study may have more to do with the length of the passages to be reconstructed. A shorter passage or a long passage divided into shorter sections might help circumvent the problem of the excessive cognitive load placed on the learners.

11. One way to encourage further processing of the form beyond noticing and immediate intake may be through collaborative dialogue as advocated by Swain (1998).

12. For instance, the picture of Christopher Reeve riding a horse was provided to help the participants to remember the sentence *If he* [*Reeve*] *had stayed healthy, he would have continued riding horses*. This picture was drawn in a wavy balloon to indicate that it is an imaginary (i.e., counterfactual) event. Therefore, the form and the meaning could be effectively connected by comparing the sentence and the corresponding picture.

13. It is interesting to note that, even with a higher cognitive demand, a slight change in scores of the CG from the first posttest to the second posttest suggests that further destabilization was taking place for these learners. It would have been interesting to examine their improvement after exposure to more relevant input.

REFERENCES

Allen, P., Swain, M., Harley, B., & Cummins, J. (1990). Aspects of classroom treatment: Toward a more comprehensive view of second language education. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 57–81). New York: Cambridge University Press.

Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book*. Rowley, MA: Newbury House.

Chaudron, C. (1985). Intake: On models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition, 7,* 1–14.

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33–56). Norwood, NJ: Ablex.

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

Hammerly, H. (1987). The immersion program: Litmus test of second language acquisition through language communication. *The Modern Language Journal, 71,* 395–401.
Harley, B. (1993). Instructional strategies and SLA in early French immersion. *Studies in Second Language Acquisition, 15,* 245–260.
Harley, B., & Swain, M. (1978). An analysis of the verb system by young learners of French. *Interlanguage Studies Bulletin, 3,* 35–79.
Izumi, S., & Bigelow, M. (1998). *Does output promote noticing and second language acquisition?* Manuscript in preparation, Georgetown University, Washington, DC.
Kowal, M., & Swain, M. (1997). From semantic to syntactic processing: How can we promote it in the French immersion classroom? In K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp. 284–309). New York: Cambridge University Press.
LaPierre, D. (1994). *Language output in a cooperative learning setting: Determining its effects on second language learning.* Unpublished master's thesis, University of Toronto, Canada.
Nobuyoshi, J., & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal, 47,* 203–210.
Pica, T. (1988). Interlanguage adjustments as an outcome of NS-NNS negotiated interaction. *Language Learning, 38,* 45–73.
Pica, T. (1992). The textual outcomes of native speaker–non-native speaker negotiation: What do they reveal about second language learning? In C. Kramsch & S. McConnell-Ginet (Eds.), *Text and context* (pp. 198–237). Cambridge, MA: Heath.
Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition, 11,* 63–90.
Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129–158.
Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Tech. Rep. No. 9, pp. 1–63). Honolulu: University of Hawai'i Press.
Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn* (pp. 237–326). Rowley, MA: Newbury House.
Sorace, A. (1996). The use of acceptability judgments 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.
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* (pp. 235–253). Rowley, MA: Newbury House.
Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50,* 158–164.
Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144). Oxford: Oxford University Press.
Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom SLA* (pp. 64–81). New York: Cambridge University Press.
Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16,* 371–391.
Takashima, H. (1994, October). *How helpful is output enhancement in promoting accuracy in communicative activity?* Paper presented at the Second Language Research Forum, Montreal, Canada.
VanPatten, B. (1995). Cognitive aspects of input processing in second language acquisition. In P. Hashemipour, R. Maldonaldo, & M. van Naerssen (Eds.), *Studies in language learning and Spanish linguistics: In honor of Tracy D. Terrell* (pp. 170–183). New York: McGraw-Hill.

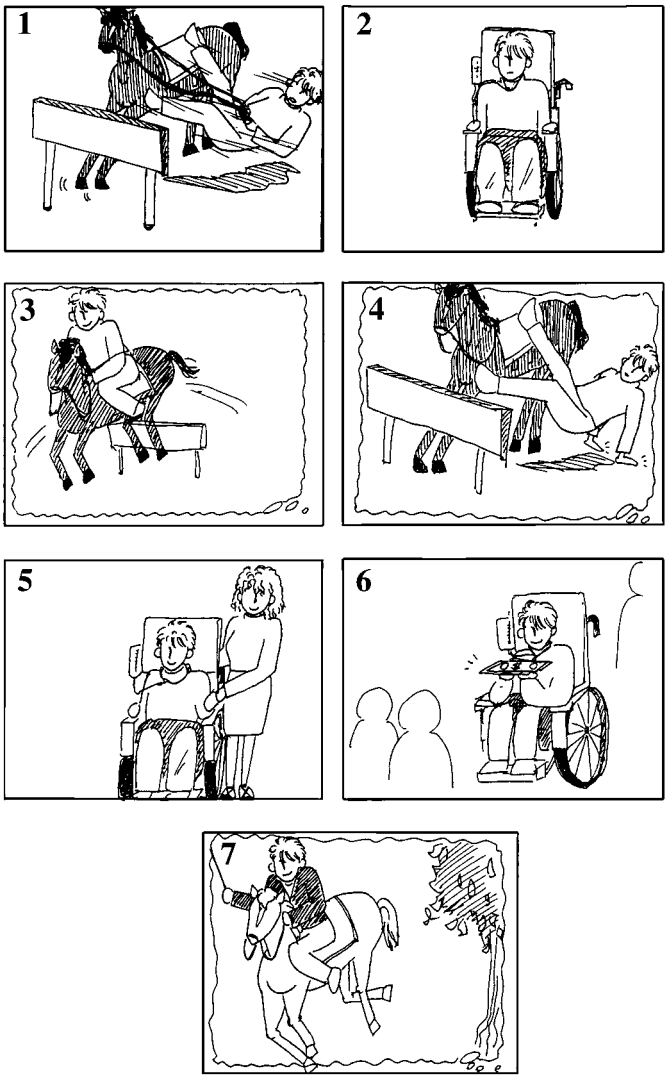
https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

APPENDIX A

PHASE 1 EXPERIMENTAL GROUP WRITING PROMPT

Reading Passage:

Last year, Christopher Reeve fell off his horse. The accident left him paralyzed. If the horse had jumped over the hurdle, Reeve would not have fallen. If his hands had been free, he would have landed safely. He would have given up all hope to live if his wife had not encouraged him to be strong. If he had felt discouraged, he would not have recognized his ability to raise money for medical research. Reeve says that if he had stayed healthy, he would have continued riding horses.



https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

APPENDIX B

PHASE 2 EXPERIMENTAL GROUP WRITING PROMPT

Scholarship Committee—You Decide!

Imagine that there had been a large amount of school funding designated last year to help part-time ESL students with tuition. A committee made up of students, teachers and administers decided who to give the scholarships to. Some of your classmates received enough money to pay for their whole tuition bill.

The committee members had different opinions about how the scholarship money should be distributed. These were some of their ideas:

$ Some people thought that the money should go to students who receive excellent recommendations from their teachers. They said, "The opinion of one teacher is enough."

$ Others considered grades to be the main factor, instead of financial need. These people said, "Academic potential is the main factor here. We have to give money to those we know will be successful students."

$ Some of the members felt that involvement in college activities should be considered primarily. They said, "Students who care about college take the time to get involved. They show community service qualities."

$ Various members of the committee suggested that students who work and study should be the only ones considered for the scholarship. They argued, "We have to appreciate the effort our part-time students make when they work all day and study all night. They deserve a break."

Imagine that you wanted to be on the committee last year. To develop your opinion, consider and write about each alternative and their potential outcomes. Then decide which alternative you prefer.

- Write at least 7 sentences (more if possible) on this topic.
- Use the following beginnings for at least 4 of your sentences.
- Try to use different beginnings each time.
- Remember that this happened last year.

If I . . .

If there . . .

If the/a student . . .

If the/a teacher . . .

I think that if . . .

For example, if . . .

On the other hand, if . . .

If the committee . . .

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press

APPENDIX C

GRAMMATICALITY JUDGMENT TEST EXAMPLES

1. I would have graduated from college two years ago if I have more time and money. YES/NO
2. I would be on time for class yesterday if I had caught the 7 o'clock train. YES/NO

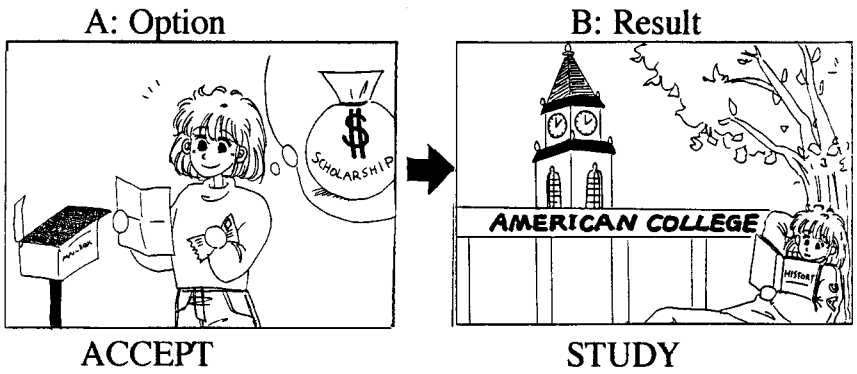
PICTURE-CUED PRODUCTION TEST EXAMPLE

Please read all of the following instructions *very carefully* and then create your own sentences. Note that all of the situations relate to the past.

Lisa had many options when she finished high school in 1992; however, she decided to work for an insurance company. The following pictures illustrate Lisa's other options and their results. She did not choose these options.

In each pair, *Picture A* shows Lisa's option when she finished high school in 1992, and *Picture B* shows the result of the option.

- Create a sentence which describes what would have happened, if she had chosen each of the following options.
- Start each sentence with "If Lisa . . . ".
- Use the verb given below each picture when you make a sentence.
- Change the verb form, if necessary.



If Lisa _____

_____

https://doi.org/10.1017/S0272263199003034 Published online by Cambridge University Press