Language Learning 49:4, December 1999, pp. 549–581

The Effects of Task Repetition on Linguistic Output

Susan Gass
Michigan State University

Alison Mackey
Georgetown University

María José Alvarez-Torres and Marisol Fernández-García
Michigan State University

This article explores form/meaning relationships, focussing on the use learners make of their internal L2 linguistic resources as a function of focus on meaning. Native speakers of English watched video segments 4 times while recording their own on-line rendition in Spanish. One group watched the same video 3 times and the

---

Susan Gass, and María José Alvarez-Torres, Department of English; Alison Mackey, Department of Linguistics; Marisol Fernández, Department of Romance and Classical Languages.

Partial funding for this project was provided by a federal grant to establish a National Foreign Language Resource Center at Michigan State University—Grant #P229A60012. A preliminary version of this article was presented at the Pacific Second Language Research Forum (PacSLRF), Tokyo, 1998. We are grateful for the comments of the participants in the symposium on Task Design and Interlanguage, organized by Jonathan Newton. We also acknowledge the important contributions of India Plough and Charlene Polio who played a significant role with various parts of this research. We also thank Jenefer Philp and David Yarowsky for their help. Finally, we express our gratitude for the astute comments of *Language Learning* reviewers and the editor. All remaining errors are ours.

Correspondence concerning this article may be sent to Susan Gass, English Language Center, A-714 Wells Hall, Michigan State University, E. Lansing, MI 48824. Internet: gass@pilot.msu.edu

other group watched different videos each time. At Time 4 both experimental groups saw a new video. A control group saw videos only at Time 1 and Time 4. Analyses were conducted on the basis of overall proficiency, morphosyntax, and lexical sophistication. The results provide limited support for the prediction of improvement over time for the group that saw the same video, but no support for a "carryover" effect when the content changed.

The recent literature on second language acquisition (SLA), and in particular, SLA within a classroom context, has witnessed a growing interest in focus on form and focus on meaning. This article explores the form/meaning relationship from the point of view of the use that learners may or may not make of their internal L2 linguistic resources as a function of greater or lesser focus on meaning. That is, how do learners extend and refine their L2 linguistic knowledge?

One area of central concern to the study of form-meaning processing is known as input processing. Input processing focusses on the conversion of input to intake during comprehension and on the form-meaning connections that occur while learners are processing the input (VanPatten, 1995, 1996; VanPatten & Cadierno, 1993; VanPatten & Sanz, 1995). In a series of experiments, VanPatten and his colleagues presented a model for instructional intervention that relies heavily on the notion of attention to form and the crucial role it plays as a learner moves from input to intake and finally to output. Within this approach, there are certain operating assumptions: (a) For acquisition to take place learners need to attend to form, (b) humans are limited in their processing capacity, (c) attention has a limited capacity—not everything can be attended to at once, and (d) form competes with meaning for attention.

VanPatten and his colleagues compared two instructional models: a traditional instruction model and a processing instruction model. In the case of the traditional instruction model, grammar was taught via a form of output manipulation; in other words,

information was presented to the learner to then be practiced. In the second model, the processing instruction model, there was an attempt to change the way input was perceived and processed. Rather than allow an internalized system to (begin to) develop, the input processing model attempts to influence the way that input is processed and hence the way the system develops. The results of these experiments (both sentence level and discourse related) suggest a positive effect for processing instruction. Participants in the processing instruction group were better able to understand and produce target structures than participants in the traditional instruction group (VanPatten & Cadierno, 1993; VanPatten & Sanz, 1995).

In another study, VanPatten (1990a) explicitly considered the question of whether learners can consciously attend to both form and meaning as they process input. The results of this study suggest that attending to morphosyntactic form and attending to meaning are often in a competitive relationship.

Another series of studies that considers the role of teacher-induced attention to form is the "garden path" research by Tomasello and Herron (1988, 1989) in which learners (a) are guided down the garden path into making errors of overgeneralization, or (b) receive explicit rule instruction (in the VanPatten framework, this could be considered akin to the input processing mode in which the focus is on processing input before internalization of that input). They found that corrective feedback was more meaningful after learners had been induced to produce an overgeneralization.[1]

Within the second language literature, much focus in recent years has been on the concept of attention and the extent to which it is a necessary condition for learning (cf. Schmidt, 1995, for a review; Ellis, 1994 [implicit/explicit learning]; Leow, 1997, 1998; Robinson, 1994, 1996; Schachter, Rounds, Wright & Smith, in press; Tomlin & Villa, 1994). Within the second/foreign language classroom there are many ways of drawing learners' attention to form, including explicit instruction, structure-based tasks that target specific predetermined structures (cf. Gass, 1997; Mackey,

1994; McDonough & Mackey, in press), and interaction (Gass, 1997; Long, 1996—the latter including well-discussed topics of negotiation and recasts).

In this article, we approach the issue of attention from a different but related perspective. Our study focusses on the ability learners have to utilize their L2 knowledge[2] in production. In particular, we investigate whether there is evidence of greater targetlike production when the need to focus on meaning[3] has been minimized through task repetition, thereby freeing learners to attend to form, not from input, but from their own internal system.

Two studies have direct relevance for our research, both relating to the repetition of tasks and the opportunity for allocating attention to meaning/form. In the first, Bygate (1996), using Levelt's model of processing, suggested that variation in task selection and use can impact oral processing. He argued that the effects of task repetition may be relevant for learning, teaching, and testing. In Bygate's study, 11 participants (no control) orally retold a video story and then retold the same story 10 weeks later. Video segments were about a minute and a half in length. Using various measures of linguistic complexity, Bygate reported (a) an increase in number of arguments per t-unit by half of the group, (b) reuse of certain phrases, and (c) an increase in comments "about the story" or metalinguistic comments, as he called them. In reviewing these comments, he found that learners increased their provision of scene-setting information, providing motivations and intentions for video characters, and their "abstracts" of what was about to happen.

Another relevant study was carried out by Skehan and Foster (1997, 1999), who proposed that more structured tasks would result in greater fluency and/or accuracy, but that linguistic complexity[4] was unrelated to task structure. Using a complex methodology, similar in one respect to the study reported here, as well as the same oral story–retelling tasks used here, Skehan and Foster differentiated Mr. Bean video segments (the prompt for the oral retelling) on the basis of story complexity, categorizing

unstructured tasks as ones without a predictable story line and structured tasks as ones with predictable story lines. In their study, for example, a restaurant scene was classified as a structured task because the story line could be anticipated in terms of the courses served and the waiter and customer's expectations about the progression of the meal. A "crazy golf" scene was classified as unstructured because the story line was unanticipated. The protagonist, in an effort not to disobey the "rules" of the game by handling the ball, hit the ball all around town. Their study consisted of four conditions: (a) watch and tell on-line (this was the condition most related to our study), (b) watch one time and then watch again while telling on-line, (c) watch and retell immediately afterward, and (d) watch and retell later (delayed). Skehan and Foster found that fluency, but not accuracy or complexity, increased when there was predictable structure.[5] Our study is similar to theirs in that we too consider familiarity of content; in our case we induced familiarity, whereas in theirs they assumed it by virtue of world knowledge.

In this study, we are focussing on the effects of task repetition on the linguistic output of second language learners. Task repetition in this study was designed to allow learners to familiarize themselves with the content of a short video extract, thus freeing up their attentional resources so they could focus on form rather than on content, which, as discussed above, we equate with meaning. In other words, we deal with the use that learners make of existing knowledge, particularly the conditions surrounding this use. The question of how knowledge comes to be has been addressed extensively in the SLA literature (see the numerous works on Universal Grammar, e.g., White, 1998, for a recent example; work on attention by Schmidt, 1995, and others; and work by Gass, 1997) and is beyond the scope of this article.

In the SLA literature of the past decade, there have been arguments for the role of output in language learning. Swain (1995, p. 127) argued that through output "learners can 'stretch' their interlanguage to meet communicative goals. They might work towards solving their linguistic limitations by using their

own internalized knowledge, or by cueing themselves to listen for a solution in future input" (see also Swain, 1985; Swain & Lapkin, 1995, 1998). We further assume that among the aspects of language to be accounted for is the control of knowledge that learners have. Bialystok and Sharwood Smith (1985), Bialystok (1994), and McLaughlin (1987) argued for the dual role of (a) knowledge representation and (b) language control (i.e., the control that one has over language knowledge). The latter includes the speed and efficiency with which language information can be accessed. In other words, learners may have knowledge of certain features, but they may not have acquired control over that knowledge. As Lightbown (1998, p. 183) pointed out, "opportunities to use the utterances in discourse-appropriate contexts help learners get this control." It is control that we are concerned with in this article, and the conditions that influence control.

We further explore whether increased targetlike production in second language form carries over to a new context. In other words, if production of a new linguistic form emerges or if there is greater accuracy in production of an existing form, there is reason to believe that there may be generalization from one context to another (cf. Tarone & Liu, 1995). Our research questions are as follows:

> Research question 1: Does task repetition yield more sophisticated language use?
>
> Research question 2: Will more accurate and/or sophisticated language use carry over to a new context?

Our general hypothesis was that familiarity with the content of the story (i.e., having watched and retold a story several times) results in more resources being available for issues of language form. Our specific hypotheses are as follows:

> Hypothesis 1: Task repetition results in greater "overall" proficiency.
>
> Hypothesis 2: Task repetition yields greater accuracy in morphosyntax.

Hypothesis 3: Task repetition results in greater lexical sophistication.

Hypothesis 4: For proficiency and morphosyntactic accuracy, change will carry over to a new context.

Method

In this study, students watched snippets from Mr. Bean videos. Mr. Bean videos involve short vignettes with a protagonist who gets himself into unusual and comical situations. We selected video snippets (approximately 6–7 min in length) in which there was no audio or in which the audio could be removed without interfering with the overall continuity or comprehension of the episode. The videos were played in a language laboratory so that participants could watch the video while simultaneously recording their own on-line rendition in Spanish of what was happening. (See Appendix A for instructions.)

Participants

Participants in the study were 103 students[6] in their fourth semester of Spanish at a large university in the United States.

Procedure

There were two experimental groups and one control group. The two experimental groups watched a Mr. Bean video four times and the control group saw a Mr. Bean video twice. For the experimental groups, each viewing was separated by 2 or 3 days, with the exception of the fourth and final viewing, which took place 1 week after the third viewing. Approximately 2 weeks separated the two control group viewings. Sections of fourth-semester Spanish participants were randomly assigned into one of three groups.[7]

1. Experimental Group 1. This group saw the same Mr. Bean episode three times followed by a fourth Mr. Bean video that

was a different episode. We refer to this as our Same Content group (n = 32).

2. Experimental Group 2. This group saw a different Mr. Bean episode each of the 4 viewing days. Their first video was the same as the Same Content group's first video and their fourth video was the same as the Same Content group's fourth video. This second group we refer to as the Different Content group (n = 33).

3. Control Group. The control group saw the first video when the other groups saw the first video, and they saw the fourth video when the other groups saw the fourth video (n = 38). They saw nothing in between.

The design is schematized in Table 1, which includes the names of the video episodes.

Our criteria for selection of the particular episodes were similar to those used by Skehan and Foster (1997, 1999). First, we isolated approximately equal-length snippets from the full version of the episode. Second, we chose the episodes to be of such a nature that participants would find them amusing and understandable and without a particular British cultural bias. Our selection

Table 1

Design of Present Study

Group	Time 1	Time 2 2–3 days later	Time 3 2–3 days later	Time 4 1 week later
Same Content (n = 32)	Library	Library	Library	Packing
Different Content (n = 33)	Library	The guard	Lunch	Packing
Control (n = 38)	Library			Packing

Note. "Library," "Packing," and so on refer to the names of the Mr. Bean episodes used in the present study. Average number of minutes per trial = 6.8; hours of data = 38.53.

differed from Skehan and Foster's in that we were not considering structured versus unstructured tasks. All of our episodes would probably fit into their unstructured category because most aspects of the sequence of events were not predictable.

Analytic Procedures

Holistic change: Magnitude estimation. There are a number of measures used for analysis in this study; the first measure (utilized to test Hypothesis 1, relating to improvement in overall proficiency) deals with holistic change across testing sessions (Time 1 to Time 3 and Time 1 to Time 4). These holistic assessments were made by native speakers of Spanish who were asked to judge Time 1, Time 3, and Time 4 on-line renditions to see if they would judge the later ones as superior to Time 1.

The methodology used to ascertain holistic change was magnitude estimation. This is a well-established research tool used in a variety of disciplines (e.g., psychophysics: Stevens, 1956; linguistics: Bard, Robertson, & Sorace, 1996; Fucci, Ellis, & Petrosino, 1990; Green, 1987; Grosjean, 1977; Pavlovic, Rossi, & Espesser, 1990; Takefuta, Guberina, Pizzamiglio, & Black, 1986; Toner & Emanuel, 1989) when one not only wants to rank items as one better than another, but also wants to know how much better X is than Y. That is, we can easily rank something, in this case speakers, into an order of 1–9, but we want to know if each of the rankings is equidistant from the others, and if not, the magnitude of the ranking differences. Magnitude estimation, the reliability and validity of which are widely accepted (see references above), was "developed by psychophysicists to make maximal use of participants' ability to make fine judgments about physical stimuli" (Bard et al., p. 32) and in recent years has been extended to other disciplines, including language research. As Bard et al. (p. 41) point out, there are a number of positive aspects to magnitude estimation, among which are: (a) Researchers do not set the number of values that are used to measure the particular property of concern and (b) one can observe meaningful differences

that directly reflect differences in impressions of the property being investigated. This latter point was particularly germane to our study.

We first describe what it was that raters were asked to judge and then discuss the way magnitude estimation judging works. As mentioned, native speakers of Spanish were asked to judge whether Time 3 and Time 4 episodes demonstrated "better Spanish" than Time 1 episodes. Thirty participants, 10 from each of the two experimental groups and 10 from the control group, were selected on the basis of having produced the most speech in each of their respective groups.[8] In Table 2 we present the number of words used for the 30 students about whom judgments were made. Each rater listened to six or nine excerpts, three different trials of three different individuals (two trials from individuals in the control group). These excerpts were randomly ordered. Each participant was rated by three different raters.

In a magnitude estimation procedure, each rater determines his or her own scale for a particular stimulus. Each subsequent stimulus is rated according to the basis established from the previous stimulus. This is perhaps best understood by showing how it works with a physical stimulus. In this study, as in other studies in linguistics using magnitude estimation (cf. Bard et al., 1996), raters were trained on the physical stimulus of line length. To begin, each rater was shown a horizontal line and was asked to assign a number to it. They were then shown another horizontal line and were asked to assign a number to it in comparison to the length of the previous line. This continued for another 10 lines. The 12 lines were between 26 and 152 mm. The first line shown was in the middle range—77 mm. It was suggested that they use a scale larger than 10. This makes subsequent ratings easier to work with because raters are dealing with multiples of a previous number. In other words, it would be difficult if raters were to start with a number such as 555 and then wanted to view the next stimulus as 1.5 times better. (For a lengthy discussion of the procedure, particularly as it relates to language-related issues, see Bard et al.)

Table 2

Word Count for Participants in Magnitude Estimation Analysis

Participant		*M*	*SD*	Participant		*M*	*SD*	Participant		*M*	*SD*
1	SC	436.75	58.59	1	DC	457.75	59.04	1	C	532.50	45.96
2	SC	368.25	101.71	2	DC	380.50	54.02	2	C	405.50	31.82
3	SC	358.50	36.83	3	DC	351.75	47.93	3	C	358.00	14.14
4	SC	343.75	63.96	4	DC	328.25	26.02	4	C	343.50	26.16
5	SC	343.50	58.19	5	DC	310.00	24.37	5	C	336.00	52.33
6	SC	327.25	56.42	6	DC	272.25	49.28	6	C	330.00	63.64
7	SC	303.25	64.07	7	DC	268.50	95.44	7	C	326.50	12.02
8	SC	259.00	103.95	8	DC	266.00	23.76	8	C	278.50	10.61
9	SC	259.00	38.58	9	DC	262.75	24.42	9	C	278.00	1.41
10	SC	257.50	73.31	10	DC	261.25	33.01	10	C	277.50	70.00
		M = 325.68				*M* = 315.90				*M* = 346.60	

Note. SC = Same Content; DC = Different Content; C = control.

After working with line length and before listening to the actual tapes, raters[9] listened to a training tape. They heard three samples of 1 min each which they rated using the magnitude estimation methodology.[10] For the actual rating, raters listened to the first 2.5 min[11] of each participant's tape. (See Appendix B for rater instructions.) For purposes of analysis, to compare the magnitude of improvement judged by each rater it is necessary to convert the unique scales created by individual raters into a logarithmic scale.[12]

Morphosyntax: Ser and estar. The question of change in the area of morphosyntax is rather complex because we did not focus participants' attention on any specific areas of the grammar. Each participant was free to focus on whatever s/he wanted to. Hence, group data may not reveal differences; one student may have focussed on agreement and another on verb tenses, for example. We did, however, select one area for reporting here on the basis of the frequent and pervasive difficulty learners have. To determine change in the area of morphosyntax (Hypothesis 2), we considered two copula verbs in Spanish, *ser*/*estar.* The difference between these two verbs is notoriously problematic for English learners at the level of instruction of our participants given that English collapses the two senses into one copular verb *to be.* We hypothesized that because of the nature of the task, we could expect frequent use of these verbs to describe characters in the video or situations and locations of these characters. The verbs in question are *ser* and *estar*, both with rough translations of *to be.* In Figures 1 and 2 we present a summary of the uses of each of these verbs with examples from our data.[13]

To be included in the analysis used to determine whether there was positive change from Time 1 to Time 3 in the use of *ser* and *estar*, a participant had to have produced at least two tokens of the verb under consideration. If not, that participant was eliminated from consideration for this part of the analysis.

Lexical sophistication. Hypothesis 3 deals with lexical sophistication. Popular measures used in describing lexical change include lexical originality, lexical density, lexical variation, lexical

Ser identifies a subject or an object

Es un lápiz para pintar
it is a pencil for to draw
'It is a pencil for drawing.'

Ser, combined with an adjective, expresses inherent characteristics that define a subject or an object

el señor Bean no **es** muy inteligente
the Mr. Bean not is very intelligent
'Mr. Bean isn't very intelligent.'

Ser refers to the time and location

pero **es** la hora y el guardia va a una posición nueva
but it is the time and the guard goes to a position new
'But it is time and the guard goes to a new post.'

Ser is also used with the preposition de ('s, from) to indicate origin, and possession

Las camisas **son** de Hawaii
the shirts are from Hawaii
'The shirts are from Hawaii.'

Figure 1. Summary of uses of *ser.* (See also Cubillos, 1996, and Gutiérrez, Martínez-Lage, & Rosse, 1995.)

frequency profiles, and lexical sophistication. In what follows we briefly define each one and then specify which measure we used and why.[14]

Lexical originality has usually been operationalized as the percentage of words used in one particular person's writing, but no one else's in the same group. Lexical density has generally been defined as the percentage of lexical words in the text (as opposed to function words). These terms generally correspond to open- versus closed-class words. According to this measure, a text is considered dense if there are many lexical words relative to

Estar is used with adjectives or adverbs of manner to describe the state or condition of the subject which are true at a given moment, but not necessarily permanent

el soldado no **está** contento
the soldier not is happy
'The soldier isn't happy.'

Estar indicates location of persons/things

el señor Bean **está** en el parque
the Mr. Bean is in the park
'Mr. Bean is in the park.'

Some adjectives have different meanings when used with *ser* and *estar*

la biblioteca **es** muy aburrida
the library is very boring
'The library is very boring.'

me parece que **está** un poco aburrido
to me it appears that he is a little bored
'It seems to me that he is a little bored.'

Figure 2. Summary of uses of *estar*. (See also Cubillos, 1996, and Gutiérrez, Martínez-Lage, & Rosse, 1995.)

function words. Criticisms of such measures (see, e.g., Laufer & Nation, 1993, 1995) point out that in written texts, fewer function words in a composition may reflect subordinate clauses or ellipsis, reflecting structural, not lexical, characteristics.

Lexical variation has been operationalized as the type-token ratio between the different words in the text and the total number of all words. Lexical variation measures are often sensitive to the length of a text, and sometimes fixed numbers of words—for example, the last 200 words in the text—have been used as the basis for the measure. Other problems include the fact that if

derivations are counted separately rather than as a "word family," counts may be inflated, depending on what the researcher is looking at. If production of different derivations are to be considered indicative of development, then, of course, there will be no problem.

Lexical frequency profiles proposed by Laufer and Nation (1993, 1995) are claimed to detail increases in the size of the active lexicon. A lexical profile is constructed by specifying where on the list each lexical item would fall, and determining the ratio of use of each entry of the list. Thus, in this system, a learner's profile of production at one time could read: 75%-10%-10%-5%. This would mean that on a composition consisting of 200 word types, 150 belong to the first 1,000 most frequent, 20 to the second 1,000, and so on. Lexical frequency profile results to date, although promising in terms of validity and reliability, are preliminary in terms of their accounts of profile stability and length and type of text (written).

Lexical sophistication, the measure with which we are concerned in this article, has been defined as the number of more advanced or sophisticated words expressed as a ratio of the total words produced. Our measure, based on a list of 200 most commonly used words in Spanish, is similar to the lexical frequency profile. In this study, "advanced" was operationalized as words not on the list of the 200 most commonly used words.[15] To summarize very simply, however, our measure is more basic in nature than the lexical frequency profile, in that we classified lexical items only in terms of whether they did or did not appear on a list of 200 most commonly used words.

In this consideration of the lexicon, our intention was to explore whether the Same Content group would have a more extensive use of vocabulary by Time 4 than would the other two groups. As noted earlier, at Times 1 and 4 all groups saw the same video, and thus the contexts for vocabulary were equivalent and could be compared. We predicted, following Ard and Homburg (1992), that because there would be little need to focus on vocabulary already practiced or known, more resources could be freed up to focus on little-used vocabulary. To summarize, for this analysis

we explored Time 1 to Time 4 comparisons and compared the words used with a standard list of 200 most commonly used words. We considered both open-class words and closed-class words. In the data set as a whole, there were 76,207 words. We used a Unix-based natural language processing tool to classify, tag, and count these words. We looked at increases in production broken down by four classes of counts:

1. increases in open-class word production
2. increases in closed-class word production
3. increases in production of all words
4. increases in production of medium- to low-frequency words (measure of lexical sophistication)

We quantified increase in production as an increase in the number of unique word types per fixed number of tokens. These type-token ratios were created in order to explore any changes in production of different types of words across the different groups. We carried out an analysis of type-token ratios rather than simple word counts because these reflect learners' increase in lexical diversity. Simple word counts do not take into account the fact that continued use of the same word is not illustrative of a wide range of word choice. For example, many learners produced large quantities of "El Señor Bean" (Mr. Bean).

Results

As noted above, three measures were used to address the research questions and our specific hypotheses.

1. A holistic measure (magnitude estimation) was carried out between Times 1 and 3 and Times 1 and 4 (Hypotheses 1 and 4).
2. An analysis of morphosyntax between Times 1 and 3 and between Times 1 and 4 was carried out to assess whether more accurate production of morphosyntax could be observed, and if so, whether it generalized to new contexts (Hypotheses 2 and 4).

3. An analysis of lexical sophistication was carried out between Times 1 and 4 to ascertain whether the learners' lexicons improved across generalizable contexts. Because lexical use is highly context dependent, the lexical comparison was not made between Times 1 and 3, because at Time 3, the two experimental groups saw different videos (Hypothesis 3).

Holistic Judgments

For holistic ratings, we first examine Time 1–Time 3 differences.

Figure 3 presents the percentage of improvement for the 10 participants of the two experimental groups at Times 1 and 3. There are two important points to note. First, the Different Content group started out higher than the Same Content group. Second, and perhaps more importantly for our purposes, the Same Content group showed greater improvement than the Different Content group. Their scores at the end were higher than those of the Different Content group, despite the fact that they started lower. Differences on a Mann-Whitney *U,* however, were nonsignificant.[16]

We next wanted to explore the extent to which the overall trends seen for grouped data reflected individual evaluations. These can be seen in Figure 4, which presents the results for both the Same Content and Different Content groups. As can be seen, overall, the magnitude of improvement was greater for individuals in the Same Content group (sum of differences = 4.56) than for individuals in the Different Content group (sum of differences = 3.39). This is the same trend shown in the grouped data reported above.

As noted earlier, one of our hypotheses (Hypothesis 4) relates to the extent to which improvement that was evidenced during treatment would carry over to a new context. Figure 5 shows a comparison of grouped data for Times 1, 3, and 4 and Times 1 and 4 for the control group. As can be seen, the Same Content group decreased considerably and the Different Content group decreased

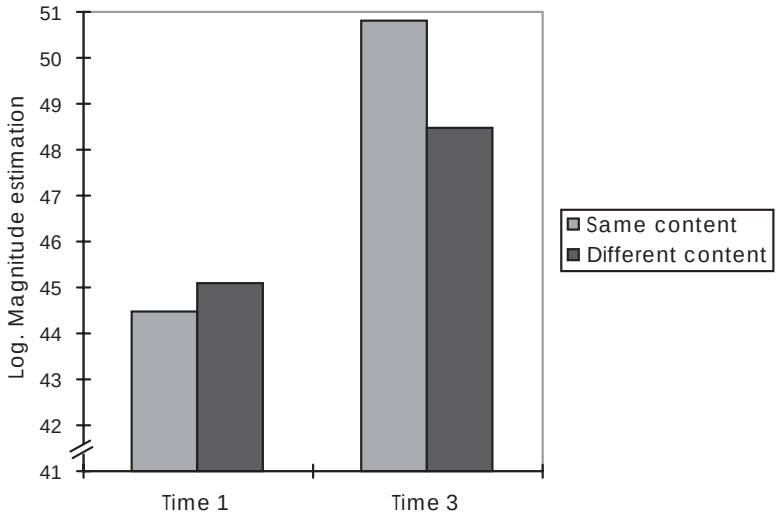

Figure 3. Change on holistic judgments (magnitude estimation) from Time 1 to Time 3 for the Same Content and Different Content groups

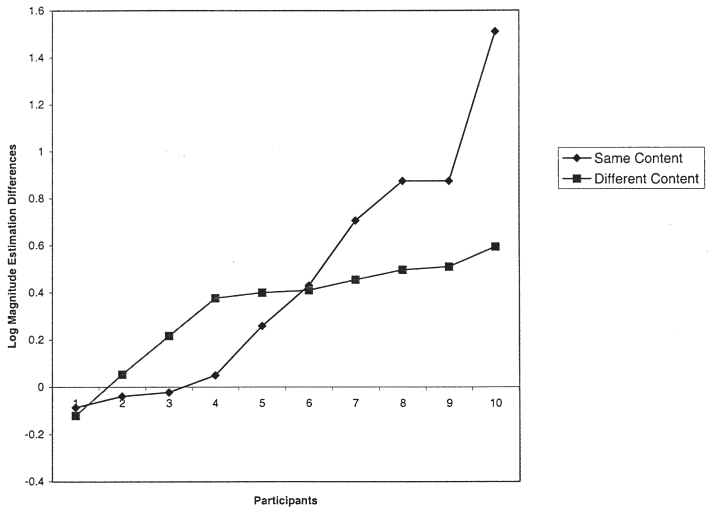

Figure 4. Comparative results between Time 1 and Time 3 for the 10 greatest producers in the Same Content and Different Content groups

slightly in the holistic ratings.[17] There was a slight increase for the control group from Time 1 to Time 4.

What we have done thus far is examine holistic judgments of improvement. The next results relate to specific areas of improvement in morphosyntax and the lexicon.

Ser/Estar

Table 3 (*ser*) and Table 4 (*estar*) show that there was a ceiling effect for both groups in their accurate use of *ser*. Because there was so little opportunity for improvement, it is unclear how to interpret the positive change that did occur. However, the situation with *estar* is different.[18] Here there was a greater percentage of participants who improved in the Same Content group as opposed to those who improved in the Different Content group (44% versus 30%). No statistical analyses were performed due to the fact that upon close examination, the starting points of the two groups who showed improvement (13 participants in the Same Content group and 9 in the Different Content group)

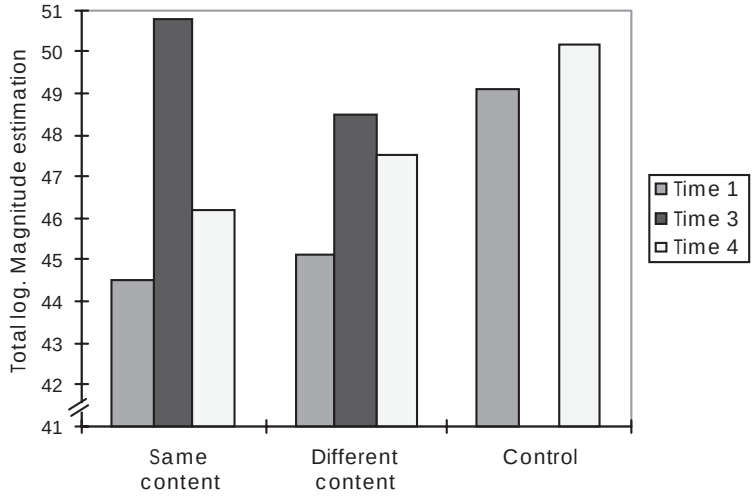

Figure 5. Comparative results over time for the Same Content, Different Content, and control groups

were not equal. The average percentage correct for the 13 Same Content participants at Time 1 was 40%, whereas the average percentage correct for the 9 Different Content participants at Time 1 was 12%. This discrepancy and the ceiling of 100% made any reasonable statistical comparison impossible.

Tables 5 (*ser*) and 6 (*estar*) show the percentage of participants who improved from Time 1 to Time 4. The purpose of this comparison was to determine whether the improvement trend that we saw from Time 1 to Time 3 for the Same Content group carried over to Time 4 (Hypothesis 4). Even though the Same Content group did show a slight improvement for *ser* (28% versus 24% for the Different Content group), with only 8% improvement for the control group, it is a complex issue to decide how to interpret these results given the high accuracy rates at Time 1.

The results for *estar* indicate a greater improvement for the Different Content group as opposed to either the Same Content group or the control group. So, whatever benefit there may have been at Time 3 stemming from less need for participants to focus on the content of the video, that benefit did not carry over to a time when a different video was seen, thus addressing the second research question and Hypothesis 4: Improvement did not generalize to a new context on the morphosyntactic measure.

Table 3

Percentage of Participants Who Improved From Time 1 to Time 3 (Ser)

Groups	Subset who improved (%)	% correct (of subset) Time 1	% correct (of subset) Time 3	% correct (all participants) at Time 1
Same Content (n = 32)	28	50	94	83
Different Content (n = 33)	21	42	100	74

Note. Criterion for improvement: at least two contexts at Time 3.

Table 4

Percentage of Participants Who Improved From Time 1 to Time 3 (Estar)

Groups	Subset who improved (%)	% correct (of subset) Time 1	Time 3	% correct (all participants) at Time 1
Same Content (n = 32)	44	40	75	46
Different Content (n = 33)	30	12	74	38

Note. Criterion for improvement: at least two contexts at Time 3.

Table 5

Percentage of Participants Who Improved From Time 1 to Time 4 (Ser)

Groups	Subset who improved (%)	% correct (of subset) Time 1	Time 4	% correct (all participants) at Time 1
Same Content (n = 32)	28	48	89	83
Different Content (n = 33)	24	42	92	74
Control (n = 38)	8	17	85	88

Note. Criterion for improvement: at least two contexts at Time 4.

Lexical Analysis

Of the total words in the database, 51% were open-class words (39,028), 44% of the total words were closed-class words (33,167), and 5% were English words (3,672). The learners often used English words where they presumably did not have the Spanish lexical item. For example, *white-out* was consistently referred to as such in English by the learners. We also measured

Table 6

Percentage of Participants Who Improved From Time 1 to Time 4 (Estar)

Groups	Subset who improved (%)	% correct (of subset) Time 1	% correct (of subset) Time 4	% correct (all participants) at Time 1
Same Content (n = 32)	31	43	88	46
Different Content (n = 33)	46	19	79	38
Control (n = 38)	34	29	84	43

Note. Criterion for improvement: at least two contexts at Time 4.

and categorized production of mid- to low-frequency words. These were words that were not on the list of 200 most commonly used words in Spanish.[19]

Table 7 shows the increase from Time 1 to Time 4 (we focussed on this comparison because at Time 4 both experimental and control groups watched the same video snippets), in open-class, closed-class, and all words. In terms of open-class words (type-token ratios), there was an increase of 11.6% for the Same Content group, whereas the Different Content group only increased by 5.6%. The control group, however, increased by 9.6%, so this finding was less interesting than the results for closed-class words. The control group increased in production of closed-class words by only 5.9%, the Different Content group by 12.3%, and the Same Content group by 26.8%. These findings for closed-class words are primarily responsible for the overall increase in all words.

Table 8 shows the mid- to low-frequency words that were not on the list of 200 most common words in Spanish, and again the results were interesting. The Same Content group increased by 12.3%, the Different Content group by 2.7%, and the control by 1.6%.

Discussion

This study tested the following four hypotheses:

1. Task repetition results in greater "overall" proficiency.
2. Task repetition yields greater accuracy in morphosyntax.
3. Task repetition results in greater lexical sophistication.
4. For holistic judgments and morphosyntactic accuracy, change will carry over to a new context.

Table 7

Percentage Increase in Type-Token Ratio From Time 1 to Time 4

Groups	Open-class words (%)	Closed-class words (%)	All words (%) (OC + CC + L1)
Same Content (n = 32)	11.6	26.8	17.6
Different Content (n = 33)	5.6	12.3	10.4
Control (n = 38)	9.6	5.9	6.8

Table 8

Percentage Increase in Type-Token Ratio of Mid- to Low-Frequency Words From Time 1 to Time 4

	Time 1		Time 4		% increase
Groups	M	SD	M	SD	Time 1 → Time 4
Same Content n = 32	53.2	12.6	59.7	27.5	12.3
Different Content n = 33	52.0	10.8	53.4	9.9	2.7
Control n = 38	53.2	10.9	54.1	6.7	1.6

Our results suggest that there is some validity to each of these hypotheses. In terms of Hypothesis 1, greater overall proficiency, the findings of better evaluations of the Same Content group at Time 3 (when the content was the same) correspond to the findings of Skehan and Foster (1997, 1999), who dealt with predictable sequences of events. In our case the predictable sequences came from having seen a story at an earlier point; in the Skehan and Foster study, it came from world knowledge, that is, events that were predictable on the basis of one's knowledge of particular situations. In general, the mere fact of repeating the task yielded improvement. In both the experimental groups as well as in the control group, Time 4 was better than Time 1. Content familiarity resulted in a greater increase, as can be seen by the large increase for the Time 3 Same Content group, although this greater increase was not sustained, as can be seen by the Time 4 results (Hypothesis 4). This is not surprising given that fluency is likely to be partially context dependent (Riggenbach, 1989, 1991).

In terms of Hypothesis 2, greater accuracy in morphosyntax, again there was evidence that task repetition led to improvement. More participants in the Same Content group showed a trend toward more targetlike production of *estar* at Time 3 than did participants in the Different Content group (44% versus 30%). In terms of Hypothesis 4, this improvement did not generalize to Time 4. One possible explanation for the lack of generalization of improvement at Time 4 has to do with the notion of task repetition. Plough and Gass (1993) suggested that when carrying out task-based work in a classroom, learners can easily become somewhat disinterested in the tasks given to them when those tasks have been carried out repeatedly. At a certain point in this study, the novelty of the task may have ended and disinterest settled in.[20]

In terms of lexical sophistication (Hypothesis 3), the findings show clearly that less common words (type-token ratios) were used in much larger numbers by the Same Content group. Knowing content thus also appears to have positively affected lexical sophistication in this study. This increase, which mostly took place in closed-class words, may be due to an increase in structured,

ordered storytelling devices, representing a more controlled retelling. This was also a finding reported by Bygate (1997).

Conclusion

In summary, using a range of measures, we found some evidence that task repetition resulted in improvement in overall proficiency, selected morphosyntax, and lexical sophistication. There does appear, then, to be limited support for the claim that freeing up attention to meaning allows learners to gain greater control over their linguistic knowledge. However, where we were able to measure generalization, using judgments of overall proficiency and morphosyntax, these findings of improvement did not generalize to a new context. Our approach to the study of the linguistic effects of task repetition was not to constrain participants to focus on any particular part of their grammars in the task repetition exercise, but to allow them complete freedom in carrying out the tasks. We thus cast a wide net in looking for any interlanguage change, and we found clear indications that such change was taking place. Our study suggests that future research might now benefit from focussing more closely on specific areas of the grammar in task repetition,[21] in order to further explore the cognitive processes of attention to meaning, attention to form, and L2 acquisition.

Revised version accepted 04 January 1999

Notes

[1]It should be noted that criticisms have been allayed against the Tomasello and Herron studies in the form of both the methodology and the analysis (see Beck & Eubank, 1991, and the response by Tomasello & Herron, 1991).

[2]In terms of the L2 knowledge of these learners, we made the assumption in this study that because all learners had similar amounts and types of instruction and had been placed into the same level, L2 knowledge would be broadly similar. For example, one of the features that we examine in this study is the Spanish copula (*ser* and *estar*). All classes had had significant amounts of instruction on these verbs prior to the onset of this study.

[3]In this study we operationalized meaning as knowledge of content. Hence, throughout, we use the terms *content* and *meaning* interchangeably.
[4]In the analysis presented in this article, we will not be specifically considering measures of complexity or fluency (although these may have entered into our holistic judgments of improvement, particularly in the case of fluency). There were two reasons for not considering complexity. First, the level of the students' knowledge of Spanish was low and thus the number of complex sentences was low from the outset. Second, given the on-line nature of the task, learners tended to produce short simple sentences, and very often incomplete ones, as they abandoned what they were saying to keep up with the events on the screen.
[5]Crookes (1989) also found that learners, as a result of planning, improved on measures of complexity, but not on measures of accuracy. As will be seen from the description of our study, our study differs from the planning literature in that planning was not a variable, and, in fact, until the beginning of the task, learners were unaware of what the task content would be, although they could predict the task procedure (watch and retell).
[6]The study began with a larger data pool. However, because we required that all students participate in every data collection session, there was attrition from beginning to end. In addition, participants who produced fewer than 50 words in two sessions were not included.
[7]Steps were taken to ensure that, for example, not all of the 8:00 A.M. sections or not all of the night sections were assigned to the same group.
[8]We selected the top 10 in each of the two experimental groups and the control group as a way of limiting the number of tapes each rater had to listen to. Listening to more than nine segments (at 2.5 min each) might have reduced the reliability of the judgments. In addition, 10 tapes from each group were deemed sufficient to provide information on change over time.
[9]There were seven raters in all: Five were university-level instructors of Spanish, one was a graduate student, and one was a bilingual secretary.
[10]Because only three samples were rated by seven raters, interrater reliability between pairs of raters could not be calculated. Instead, the seven raters' rankings of pairs of samples were correlated. The reason for this was that if all seven raters ranked Sample 1 first, for example, and Sample 2 second, there would be a high correlation between them (i.e., 1.0). As it was, the Spearman correlations between pairs of the samples were .93, .96, and .98.
[11]We selected the first 2.5 min of the tape (as opposed to other parts of the tape) because the beginning was the only place where we could be assured of common ground; all participants began talking at the beginning of the film. Had we selected any later portions, we were not guaranteed equivalent uniformity.
[12]Conversion to a logarithmic scale is standard procedure when using magnitude estimation. Because the methodology allows for unique scales to be created by each rater, there must be a way to standardize the scales across raters to obtain a meaningful comparison.

[13]Two native speakers of Spanish coded these data. For both verbs, interrater reliability was .97.

[14]Most of these measures come out of the L2 writing literature and all have their critics. Criticisms have focussed on issues of reliability, validity, or both (Arnaud, 1984; Laufer, 1995; Laufer & Nation, 1993; Read, 1988).

[15]After advertising on Linguist list, we obtained from an Internet site a list of 2,000 most commonly used words, broken down incrementally. For the analysis presented in this article, we utilized 200 most commonly used words. The patterns reported in this article also held for increments at higher levels with minor variations. See (http://ccr.dsi.uanl.mx/~rchandlr/RCHANDLR.WWW7 and http://ccr.dsi.uanl.mx/~rchandlr/RCHANDLR.WWW6). The first address is for the actual list and the second is for the instructions.

[16]The issue of meaningfulness versus significance (in the statistical sense) is an important one in this and in other L2 studies, particularly those studies that are based on classroom treatments. Ideally, in an experiment such as the one reported here, one would want a longer period to pass in order to allow learners to demonstrate change in language use. Language learning is not instantaneous; it takes time for information to become integrated into one's grammar (Gass, 1997; Lightbown, 1998). However, the realities of many experimental contexts do not allow a more protracted period given the constraints placed on our use of class time (see papers in Schachter & Gass, 1996, for discussions of difficulties and constraints in conducting classroom research). This returns us to the issue of significance. From our perspective, we have trends that are meaningful. However, these trends are not significant in the statistical sense with a probability level of $<.05$, the generally accepted standard in the field.

The need to have all results adhere to a .05 standard may be questionable. Shavelson (1988) noted that the convention of using .05 or .01

> grew out of experimental settings in which the error of rejecting a true H_0 was very serious. For example, in medical research, the null hypothesis might be that a particular drug produces undesirable effects. Deciding that the medicine is safe (i.e., rejecting H_0) can have serious consequences. Hence, conservatism is desired. (p. 248)

Shavelson went on to say that "often in behavioral research . . . the consequences are not so dire" (p. 248). He suggested that a level of .25 might be appropriate in some cases. He concluded his discussion by saying that "some wisdom, then, should be exercised in setting the level of significance" (p. 248) and pointed out that there is a trade-off between the level of significance that one sets and the power of one's conclusions.

Given the essential arbitrariness in setting significance levels and given the constraints in conducting classroom research, we feel that trends are important and at least point to the notion that experiments should be replicated, particularly when it is impractical or impossible for experiments

to cover a long period. We also believe that trends may at times be as meaningful as statistical significance.

[17]Given our instructions to raters to evaluate holistically, we cannot ascertain to what extent the holistic evaluations reflect fluency, accuracy, or complexity. Our goal was not to differentiate among these possibilities; rather, these were taken to represent an overall improvement in "proficiency."

[18]VanPatten (1985, 1987, 1990b) showed that *ser* is learned before *estar.* He provided three reasons for this learning order: (a) *Ser* is more frequent in the input, (b) simplification—learners adopt a "one copula" system in the early stages of acquisition, and (c) language transfer. The greater frequency of *ser* and the natural learning mechanism of one form/one meaning leads learners to rely on the native language "one copula" system.

[19]For learners in this developmental range, words not on the list of 200 most commonly used words can be classified as mid- to low-frequency in terms of use, as opposed to native speakers for whom, obviously, they would not be mid to low. Examples of these words include *dentro* (within), *mitad* (half), *próximo* (next), *lleno* (full), *durante* (during, for), *también* (also, as well as), *bien* (good, well), *grande* (big), and *comenzar* (to start).

[20]As pointed out in Note 6, some students were eliminated from the study if they produced very little (i.e., were not on task). It is also possible that not being on task was evidence of boredom.

[21]One way of achieving more robust results in a study of task repetition would be to use tasks that produce contexts for particular structures (cf. Mackey, 1995, 1999).

References

Ard, J., & Homburg, T. (1992). Verification of language transfer. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 47–70). Amsterdam: John Benjamins.

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing*. Colchester, UK: University of Essex.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72,* 32–68.

Beck, M., & Eubank, L. (1991). Acquisition theory and experimental design: A critique of Tomasello & Herron. *Studies in Second Language Acquisition, 13,* 73–76.

Bialystok, E. (1994). Analysis and control in the development of a second language. *Studies in Second Language Acquisition, 16,* 157–168.

Bialystok, E., & Sharwood Smith, M. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second-language acquisition. *Applied Linguistics, 6,* 101–117.

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136–146). London: Heinemann.

Bygate, M. (1997, March). *The effect of task repetition on language structure and control*. Paper presented at the 1997 American Association for Applied Linguistics Convention, Orlando, FL.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11,* 367–383.

Cubillos, J. H. (1996). *Siempre adelante* [Onwards and Upwards]. Boston: Heinle & Heinle.

Ellis, N. (1994). Implicit and explicit language learning—An overview. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1–32). London: Academic Press.

Fucci, D., Ellis, L., & Petrosino, L. (1990). Speech clarity/intelligibility: Test-retest reliability of magnitude estimation scaling. *Perceptual and Motor Skills, 70,* 232–234.

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

Green, K. (1987). The perception of speaking rate using visual information from a talker's face. *Perception and Psychophysics, 42,* 587–593.

Grosjean, F. (1977). The perception of rate in spoken and sign languages. *Perception and Psychophysics, 22,* 408–413.

Gutiérrez, J. R., Martínez-Lage, A., & Rosse, H. L. (1995). *¡Tú dirás!* [Tell me!]. Boston: Heinle & Heinle.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In L. Eubank, L. Selinker, & M. Sharwood Smith (Eds.), *The current state of interlanguage: Studies in honor of William E. Rutherford* (pp. 265–272). Amsterdam: John Benjamins.

Laufer, B., & Nation, P. (1993, August). *Lexical richness in L2 written production: Can it be measured?* Paper presented at the 10th Congress of the International Association of Applied Linguistics, Amsterdam.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16,* 307–322.

Leow, R. (1997). Attention, awareness and foreign language behavior. *Language Learning, 47,* 467–506.

Leow, R. (1998). The effects of amount and type of exposure on adult learners' L2 development in SLA. *The Modern Language Journal, 82,* 49–68.

Lightbown, P. (1998). The importance of timing in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 177–196). Cambridge: Cambridge University Press.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Mackey, A. (1994). Targeting morpho-syntax in children's ESL: An empirical study of the use of interactive goal-based tasks. *Working Papers in Educational Linguistics, 10,* 67–88.

Mackey, A. (1995). *Stepping up the pace: Input, interaction and second language development*. Unpublished doctoral dissertation, University of Sydney, Australia.

Mackey, A. (1999). Input, interaction and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21,* 557–587.

McDonough, K., & Mackey, A. (in press). Form and meaning: Designing communicative tasks to target grammar in Thai classrooms. *Foreign Language Annals.*

McLaughlin, B. (1987). *Theories of second language acquisition*. London: Edward Arnold.

Pavlovic, C., Rossi, M., & Espesser, R. (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems. *Journal of the Acoustical Society of America, 87,* 373–382.

Plough, I., & Gass, S. (1993). Interlocutor and task familiarity: Effects on interactional structure. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 35–56). Clevedon, UK: Multilingual Matters.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19,* 12–25.

Riggenbach, H. (1989). *Nonnative fluency in dialogue versus monologue speech: A microanalytic approach*. Unpublished doctoral dissertation, University of California, Los Angeles.

Riggenbach, H. (1991). Toward an understanding of fluency—A microanalysis of nonnative speaker conversations. *Discourse Processes, 14,* 423–441.

Robinson, P. (1994). Implicit knowledge, second language learning and syllabus construction. *TESOL Quarterly, 28,* 160–166.

Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search and instructed conditions. *Studies in Second Language Acquisition, 18,* 27–67.

Schachter, J., & Gass, S. (Eds.). (1996). *Second language classroom research: Issues and opportunities.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schachter J., Rounds, P., Wright, S., & Smith, T. (in press). Comparing conditions for learning syntactic patterns: Attentional, nonattentional, and awareness. *Applied Linguistics.*

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* (pp. 1–64). Honolulu: University of Hawai'i Press.

Shavelson, R. (1988). *Statistical reasoning for the behavioral sciences*. Boston: Allyn and Bacon.

Skehan, P., & Foster, P. (1997, March). Language and content planning, task structure, and task performance. Paper presented at the 1997 American Association for Applied Linguistics Convention, Orlando, FL.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narritive retellings. *Language Learning, 49,* 93–120.

Stevens, S. (1956). The direct estimation of sensory magnitudes—Loudness. *American Journal of Psychology, 69,* 1–25.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensive output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford 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.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal, 82,* 320–337.

Takefuta, Y., Guberina, P., Pizzamiglio, L., & Black, J. (1986). Cross-lingual measurements of interconsonantal differences. *Journal of Psycholinguistic Research, 15,* 489–507.

Tarone, E., & Liu, G. (1995). Situational context, variation, and second language acquisition theory. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 107–124). Oxford: Oxford University Press.

Tomasello, M., & Herron, C. (1988). Down the garden path: Inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics, 9,* 237–246.

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

Tomasello, M., & Herron, C. (1991). Experiments in the real world: A reply to Beck & Eubank. *Studies in Second Language Acquisition, 13,* 513–517.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16,* 183–203.

Toner, M., & Emanuel, F. (1989). Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research, 31,* 78–82.

VanPatten, B. (1985). The acquisition of *ser* and *estar* by adult classroom learners: A preliminary investigation of transitional stages of competence. *Hispania, 68*, 399–406.

VanPatten, B. (1987). Classroom learners' acquisition of *ser* and *estar*: Accounting for developmental patterns. In B. VanPatten, T. Dvorak, & J. Lee (Eds.), *Foreign language learning: A research perspective* (pp. 61–75). Cambridge, MA: Newbury House.

VanPatten, B. (1990a). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12,* 287–301.

VanPatten, B. (1990b). Theory and research in second language acquisition and foreign language learning: On producers and consumers. In B. VanPatten & J. Lee (Eds.), *Second language acquisition—Foreign language learning* (pp. 17–26). Clevedon, UK: Multilingual Matters.

VanPatten, B. (1995). Input processing and second language acquisition: On the relationship between form and meaning. In P. Hashemipour, R. Maldonado, & M. van Naerssen (Eds.), *Festschrift in honor of Tracy D. Terrell* (pp. 170–183). New York: McGraw Hill.

VanPatten, B. (1996). *Input processing and grammar instruction: Theory and research*. Norwood, NJ: Ablex.

VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition, 15,* 225–243.

VanPatten, B., & Sanz, C. (1995). From input to output: Processing instruction and communicative tasks. In F. Eckman, D. Highland, P. Lee, J. Mileham, & R. Weber (Eds.), *Second language acquisition theory and pedagogy* (pp. 169–186). Hillsdale, NJ: Lawrence Erlbaum Associates.

White, L. (Ed.). (1998). The implications of divergent outcomes in L2 acquisition [Special issue]. *Second Language Research, 14*(4).

Appendix A

Instructions for Participants

You will watch a story about a man named El Señor Bean. This is approximately a 6-minute silent video called xxxx. When the video begins, your teacher will begin recording your voice. While watching the video, tell the story in Spanish with as many details as possible. That is, talk about what you see as it is happening. Do not wait until the video is over to talk. Try to say as much as you can.

Appendix B

Instructions for Raters

You will hear nine tapes of different non-native speakers of Spanish doing an on-line description in Spanish of a video they were watching. Your task is to rate their Spanish. Assign any number that seems appropriate to you to the first speech sample. This number will be your "base." Then assign successive numbers in such a way that they reflect your subjective impression (use a range wider than 10). For example, if a speech sample seems 20 times as good, assign a number 20 times as large as the first. If it seems one-fifth as good, assign a number one-fifth as large, and so forth. Use fractions, whole numbers, or decimals, but make each assignment proportional to how good you perceive the person's Spanish to be.