Effects of Rehearsal on ESL Learners' Responses to Test Tasks

Betsy Lavolette
Second Language Studies Program
Michigan State University
betsy@msu.edu

Second language (L2) testing is often stressful for test takers, especially when they take high-stakes tests such as the TOEFL. On the iBT TOEFL independent speaking questions, for example, test takers are given 15 seconds to prepare to respond to a prompt (Educational Testing Service, 2008) and have only one chance to record their answer, which puts test takers under great pressure. However, the necessity of this stressful situation has not been empirically validated; that is, the effects of giving learners planning opportunities are unclear.

The major types of planning that are generally distinguished are pretask planning and within-task planning (Ellis, 2005). Pretask planning is subdivided into strategic planning and rehearsal, and within-task planning is subdivided into pressured and unpressured planning (Ellis, 2005). Strategic planning is operationalized by giving learners time (often about 5 to 10 minutes) to plan before being asked to perform a task, while rehearsal is operationalized as repetition of a task. Within-task planning is regarded as pressured when a short time limit is set for the performance of a task.

Definition of a Task

The studies reviewed below (and much of the other research on tasks) present their tasks such as video narration (Gass, Mackey, Alvarez-Torres, & Fernández-García, 1999) and poster presentation (Lynch & Maclean, 2000) without any discussion of whether these meet the definition of a task. While tasks have been defined in many different ways, the one that I will adopt here is that of Ellis (2003, pp. 9–10). The key features of a task are (a) a primary focus on meaning, (b) an information gap that learners must communicate to fill, (c) no specification of what linguistic resources must be used to fulfill the task, and (d) a communicative outcome beyond a gratuitous display of language. The poster presentation described by Lynch and Maclean (2000), which simulates the presentation of a poster at an academic conference, fits this definition well, while one could argue that the video narration of Gass et al. (1999) fails on criteria b and d, given that the researchers have also seen the video. However, another way to view the issue is to consider activities on sliding scale, where some possess more task-like qualities than others. The task used in the present study, in which learners responded to a TOEFL prompt, is similar to the video narration in that it has some task-like qualities (i.e., it meets criteria a and c). While not all of the tasks in the studies reviewed below meet all of the criteria, I will refer to them as tasks.

Tasks can also be divided into types based on their purposes. Pedagogic tasks are primarily aimed at enabling student learning, while test tasks are intended to assess student learning.

Pedagogic Tasks

Strategic Planning.

Researchers have extensively examined strategic planning for pedagogic speaking tasks, with the general result that this type of planning improves fluency and complexity, but the results are mixed regarding its effect on accuracy (e.g., Crookes, 1989; Foster & Skehan, 1996, 1999; Mehnert, 1998; Ortega, 1999; Wendel, 1998).

Rehearsal.

Researchers have also investigated how rehearsal affects learners' task performance. Gass, Mackey, Alvarez-Torres, and Fernández-García (1999) asked L2 learners of Spanish to narrate the action in silent video clips in their L2. The participants were divided into three groups: one group (Same Content group) saw the same clip at Times 1 through 3, then a new clip at Time 4; a second group (Different Content group) saw a different clip each of the four times; and a control group saw two different clips, one at Time 1 and one at Time 4. These viewing times were separated by 3 to 4 days and took place in a laboratory setting. Overall, no significant differences were found between the groups on complexity, accuracy, or fluency, with one exception. In terms of lexical sophistication, the Same Content group used more lower frequency words at Time 4 compared to Time 1 than did the Different Content group or the control group. These results indicate that the task repetition had little effect on overall performance, fluency, and accuracy and that only the effect on lexical sophistication carried over to a new task.

Lynch and Maclean (2000) studied two learners at very different levels of English as they repeated a poster presentation task six times during one class session, without instructor feedback. They found that the advanced learner improved in pronunciation accuracy and lexicogrammatical performance. The beginning learner gained in syntactic, lexicogrammatical, and phonological accuracy from the first performance to the last.

Bygate and Samuda (2005) experimentally examined the effects of task repetition, but with a much longer interval of 10 weeks between the first and second performances. They analyzed the performances of 48 learners of English who narrated short cartoon clips and found an effect on fluency and complexity.

The results of these three studies found rather different effects, possibly because of the varied tasks, intervals between repetitions, numbers of repetitions, interactivity (monologic or dialogic), settings (classroom or laboratory), and proficiency levels of the participants. One characteristic that these tasks have in common is that they were all conceived as pedagogical in nature. Test tasks, on the other hand, may affect learner output in a different way, as considered in the next section.

Test Tasks

Despite the many pedagogical studies above that found effects of strategic planning, in testing situations, almost no effects of this type of planning have been found. Iwashita, McNamara, and Elder (2001) found no effect of 3.5 versus 0.5 minutes of pretask planning time on speaking test task ratings or complexity, accuracy, and fluency (CAF) determined by close analysis of the resulting discourse. Similarly, Wigglesworth and Elder (2010) found no difference on ratings or CAF for learners who were given 15 seconds, 1 minute and 15 seconds, or 2 minutes and 15 seconds of planning time before performing speaking test tasks.

No studies have looked at the effects of rehearsal on learners' responses to test tasks. Speculatively, rehearsal of a test task may differ from that of a pedagogic task because of the differing levels of stress in each situation and the learner's focus. The learner taking a test may feel much greater pressure to give his or her best performance than a learner in a normal classroom situation. In addition, to the learner taking a test, "best performance" may mean an accurate performance, rather than one that includes the most complex or fluent language that he or she can produce.

Current Study

Learners are put under tremendous stress to respond quickly to prompts on tests such as the TOEFL, without any evidence that pretask planning and opportunities for rehearsal make a difference in their scores or CAF. No studies have looked at how learners respond to test tasks when they are given the opportunity to rehearse, in addition to pretask planning time. Therefore, in this study, I ask,

> What effect does rehearsing the response to a prompt have on ESL learners' CAF and holistic ratings if they repeat the response once? More than once?

Bygate and Samuda (2005), drawing on Levelt's (1989) model of speech production, claimed that task rehearsal has an effect on both conceptualization (planning the propositions to be expressed) and formulation (choosing the lexical and grammatical elements needed to express the propositions), and thus improves complexity and fluency:

> Hence on the second occasion, formulation is likely to be speedier and more accurate. In addition to these influences, clearly the improvement of speed and accuracy of the conceptualization processes outlined above is likely to make more capacity available at the formulation level. If we think of repetition as enabling a second 'draft', then task repetition involves targeting improvement not just of the draft (i.e., the language produced) but of the actual drafting process. That is, task repetition can have an impact on the processing, and not just on the product. (Bygate & Samuda, 2005, p. 45)

However, given that Iwashita et al. (2001) and Wigglesworth and Elder (2010) found no effect for strategic planning in testing situations, Bygate and Samuda's (2005) reasoning may not hold in the present case. I predict that, rather than affecting complexity and fluency, test task repetition will lead to improvement of only accuracy. As Iwashita et al. (2001) suggested, the testing situation itself may alter the focus of the learners:

> In a test, where tasks are carried out alone in a computer-mediated environment and hence lack an interactive dimension, the cognitive focus may be on display, and this may alter the relation between task characteristics and language output. For example, *a focus on accuracy may be paramount in the testing situation* regardless of the conditions under which the task is performed, and this in turn may affect the fluency and complexity of candidates' speech. Delivery may be halting whether the task is easy or difficult, because the candidates are focusing primarily on correctness. The lack of complexity in candidates' production may likewise be due to their anxiety about how their speech is being evaluated, making them reluctant to venture beyond what they know how to say properly even when the task conditions allow for this. (Iwashita et al., 2001, p. 431; emphasis added)

Thus, given that learners are likely to be focused on accuracy in a testing situation, the trade-off hypothesis (e.g., Foster & Skehan, 1996), which was developed to account for differences in CAF under different task conditions, predicts that the learners will have fewer attentional resources to devote to complexity and fluency. When learners repeat the test tasks in the current study, I predict that accuracy will improve, but not complexity and fluency.

Method

Participants

Thirty-nine English-language learners enrolled at the Michigan State University English Language Center (ELC) participated in this study outside of their normal class hours. Nineteen of the learners were in Level 3 classes and 14 were in Level 4 classes in the intensive English program, 3 were in English for academic purposes courses (already enrolled in the university), and the levels of 3 were unknown. The learners were invited to participate in the study by their teachers, at the request of the researcher. They were told that they would practice for the independent speaking portion of the iBT TOEFL test, and they received extra credit in their classes for participating.

Materials

I used TOEFL iBT Test Independent Speaking prompts to elicit speech from the participants. For this analysis, only responses to one prompt will be considered:

> Some college students choose to take courses in a variety of subject areas in order to get a broad education. Others choose to focus on a single subject area in order to have a deeper understanding of that area. Which approach to course selection do you think is better for students and why? (ETS, 2006, p. 230)

This prompt was originally published by ETS in a TOEFL preparation book. The rubric that was used to score the recordings was published on the Internet by ETS; it is the same one used in the iBT TOEFL. An opinion and background questionnaire was also administered to the learners; see the Appendix.

Procedure

The learners came to a computer lab outside of their normal class time. The researcher explained the study and demonstrated the technology that was used. Then, the learners recorded three audio and three video speech samples in response to TOEFL iBT Test Independent Speaking prompts using Audio and Video Dropboxes (created by the Center for Language Education and Research at Michigan State University, http://clear.msu.edu/clear/index.php), with the order of the prompts and technology counterbalanced. The first sample of each mode (one audio and one video) was used for the learners to practice using the technology and was not analyzed. The learners read the prompts on the computer screen, then made as many recording attempts as they liked. They were given no time restrictions on pretask planning, note taking, or rehearsal attempts, and they were also given no guidance. All rehearsal attempts were recorded. The time for the response was limited to 45 seconds, as in the TOEFL test, which served to limit their online planning time. After completing the recordings, the learners filled out a questionnaire on their opinions of the two modes of recording and demographic information. All of the learners received extra credit in their courses for participating in the research.

Analysis

Only the responses to one of the prompts in the audio mode were considered. This prompt (see above) was chosen because of the convenient numbers of participants who made one or multiple recordings: Nine of the participants made two or more audio recordings for this prompt, and 9 of the participants made only one.

Ten trained raters rated the recordings holistically, using the TOEFL rubric. Two raters, including the researcher, evaluated CAF. Grammatical complexity was evaluated using the amount of subordination, operationalized as the total number of clauses divided by the total number of AS units. Lexical complexity was measured using the number of words outside the most frequent 1000 English words divided by the total number of words in the response. Accuracy was evaluated using the general measure of error-free clauses and the specific measure of target-like use of finite verb phrases. Fluency was evaluated by dividing the number of syllables in the pruned speech by the total time allotted (45 seconds).

Discussion and Conclusion

Limitations

The learners decided themselves whether they would make multiple recording attempts or just one, which means that they, in effect, self-selected whether they were in the experimental group or control group. This makes the division into the groups nonrandom and limits the conclusions that can be drawn from this study. In addition, the learners self-selected the amount of time that they spent on pretask planning, making that variable uncontrolled. It is also possible that the learners rehearsed their performances before they actually began recording, although I did not see evidence of this.

Another limitation is that the students were not asked how they oriented to the task. That is, although the task is viewed as a test by the researcher, the learners themselves may not have treated it that way. Given that previous results showed a difference in learners' performance based on the task being a test or not, the results need to be interpreted with caution.

Testing Implications

If the results are as anticipated, the learners in this study will improve on accuracy but not on holistic ratings, fluency, or complexity when they repeat the task. The learners who repeat the task will also have higher accuracy than learners who do not repeat the task. On the other hand, if results are not as anticipated, an alternative explanation is as follows: A learner focus on accuracy when performing a test task may not lead to gains in accuracy when the task is repeated simply because of the focus on accuracy from the beginning. Instead, it is possible that in subsequent performances, the learners will switch their attention to other aspects of production.

If learners were allowed to rehearse on the real TOEFL speaking test, they might similarly increase their accuracy (or alternatively, other aspects of production), which would change the outcome of the test by biasing for the best (Fox, 2004; Swain, 1983). The test could still be regarded as "fair" in the sense that all test takers would be given the same opportunity to rehearse. I regard this as a positive change because, as Swain (1983) claimed,

> [I]f the testee does well, then it can be said with some confidence that the learner can do what is expected of him or her when given the opportunity. However, if the testee does not do well, then it is not clear whether this occurs because the testee cannot do what is expected, or is prevented from doing it because of other distracting factors, or whatever. (p. 141)

The chief advantage to allowing this is to reduce the stressfulness of a very high-stakes exam, which may allow test takers to better show their full capabilities.

References

Bygate, M., & Samuda, V. (2005). Integrative planning through the use of task-repetition. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 45–82). Amsterdam, The Netherlands: John Benjamins.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11*(4), 367–383.

Educational Testing Service. (2006). TOEFL iBT Speaking. *Official Guide to the New TOEFL iBT with CD-ROM* (pp. 207–248). New York, NY: McGraw-Hill.

Educational Testing Service. (2008). TOEFL iBT Tips: How to prepare for the TOEFL iBT.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.

Ellis, R. (2005). Planning and task-based performance: Theory and research. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 3–34). Philadelphia, PA: John Benjamins.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second langauge performance. *Social Research*, *18*, 299–323.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*(3), 215–247.

Fox, J. (2004). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly, 1*(4), 235–251.

Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, *49*(4), 549–581.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401–436.

Levelt, W. (1989). *Speaking: From intention to articulation*. Boston, MA: MIT Press.

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research, 4*(3), 221–250. doi:10.1177/136216880000400303

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83–108.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*(1), 109–148.

Swain, M. (1983). Large-scale communicative language testing: A case study. *Language Learning and Communication, 2*(2), 133–147.

Wendel, J. N. (1998). *Planning and second-language narrative production*. Philadelphia, PA: Temple University.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*, 37–41. doi:10.1080/15434300903031779

Appendix

Learner Background Questionnaires

Learner background questionnaire

1. Age in years:
2. Gender
   Male Female Other
3. What is your first language?
   Mandarin (Chinese) Cantonese (Chinese) Korean Arabic
   Japanese Spanish Other ____
4. How many years have you been studying English?
5. How long have you been living in the US or another English-speaking country?
6. Have you taken the TOEFL before? If so, which version did you take most recently?
   iBT PBT CBT
7. What was your most recent TOEFL score?