Output tasks and vocabulary gains

Keywords
pushed output, vocabulary, motivation

Folse (2004) argued for the importance of vocabulary instruction and the effectiveness of list learning, while Laufer and Girsai (2008) found mechanical output tasks using contrastive analysis and translation effective for vocabulary learning, following Swain and Lapkin's (1995) advocacy of pushed output using creative tasks. Vocabulary gains over one semester were compared from a treatment group of 37 learners taught vocabulary using mechanical tasks with a control group of 67 learners assigned creative output tasks in a quasi-experimental design. Rasch measurement was used to provide equated scores from vocabulary pre-tests and post-tests. Both groups showed substantive gains in vocabulary knowledge but the control group showed larger vocabulary gains than the treatment group, contrary to expectations. These results suggest that mechanical tasks alone may not lead to optimal gains in vocabulary knowledge.

Swain and Lapkin (1995)は創造的タスクを使う強制的アウトプットを提唱し、Folse (2004)は語彙指導の重要性とリスト学習の有効性を主張した。一方、Laufer and Girsai (2008)は対照分析と翻訳を用いる機械的なアウトプットを促すタスクが語彙学習に有効であると指摘した。本論では、1学期間での語彙習得度を準実験的形式で、機械的なタスクを使い語彙指導を受けた37名の実験群と、創造的アウトプットタスクの指導を受けた67名の統制群を比較した。語彙テストの事前・事後のスコアを等価するために、ラッシュ分析を用いた。両グループの事後テストにおいて実質的な語彙習得が認められたが、予測に反して実験群よりも統制群における語彙習得の方が大きいという結果になった。これは、強制的アウトプットが言語習得に効果的な手段とする主張を支持する結果であり、長期的な語彙習得には機械的タスクのみでは不十分であることを示唆する。

Trevor A. Holster
Darcy F. de Lint
Kyushu Sangyo University

Vocabulary is undoubtedly crucial to language, but "most vocabulary research in applied linguistics is based on a narrow linguistic agenda that was to a large extent defined by the concerns of the vocabulary control movement in the 1920s" (Meara, 2002, p. 393), an agenda Meara termed the "vocabulary manifesto". Folse (2004), advocating this agenda, argued for the effectiveness of list learning, despite being "dull", and claimed that:

> Unfortunately, traditionally vocabulary has received less attention in second language (L2) pedagogy than any of these other aspects, particularly grammar. Arguably, vocabulary is perhaps *the* most important component in L2 ability (p. 22).

In contrast, Swain and Lapkin (1995), following Schmidt's (1990) argument for conscious "noticing", argued that output tasks can lead to noticing of linguistic shortcomings, "pushing" learners to modify output. Laufer and Girsai (2008) compared contrastive analysis and translation (CAT) tasks with meaning-focused instruction (MFI) and form-focused instruction (FFI), finding superior results from the CAT task on vocabulary post-tests conducted one week later. This was attributed to pushed output, on the claim that translation tasks force learners to confront problematic language, unlike open-ended tasks, which allow avoidance. Laufer and Girsai's (2008) pushed output thus refers to highly constrained mechanical output (MO) tasks, whereas Swain and Lapkin (1995) investigated creative output (CO) writing of original compositions. Although Laufer and Girsai (2008) found superior results from the CAT task, Rott, Williams, and Cameron (2002) found that, while multiple-choice glosses led to greater immediate learning compared with text reconstruction, "a significant receptive word gain was retained for five weeks only for the combined treatment condition" (p.

183), highlighting the importance of longitudinal studies to investigate whether experimental treatments translate into improved long-term proficiency.

Schumann and Wood (2004, p. 23) described Sustained Deep Learning (SDL) as underlying long-term language proficiency gains, but in experimental studies, "participants typically learn material unrelated to their goals and are tested on it after relatively short periods of time". The SDL model sees learning as an evolutionary adaptation employing neural systems originally used for foraging for food. Opportunities for both feeding and learning require positive goal appraisals and environmental engagement, each situation evaluated on novelty, pleasantness, goal relevance, coping ability, and self/social image compatibility. Biological value thus underlies preferences and enables choices, with positive rewards affecting future preferences and choices, making positive assessment of learning experiences crucial for future motivation. This raises questions about the motivational effect and opportunity cost of dull mechanical vocabulary tasks relative to the other tasks that must be dropped to make time for vocabulary instruction. These questions require long-term comparisons under classroom conditions.

This is consistent with Hattie's (2009) review of educational meta-analysis, emphasizing comparison of classroom interventions to identify those that are most effective in promoting long-term gains. Although both Folse (2004) and Hattie (2009) argued that pedagogy should be guided by research, Folse assumed that isolated experimental studies generalize to classrooms, whereas Hattie emphasized comparison of the effect sizes of different interventions under classroom conditions over extended periods. A further benefit of Hattie's approach is that even pilot studies with small sample sizes or null findings can contribute useful data, providing a richer perspective than if only large-scale experimental studies with statistically significant findings are considered. The importance of considering effect sizes, indicating substantive significance, was addressed by Thompson (1999) in a scathing critique of statistical significance tests, which claims that large sample sizes can lead to results that are both statistically significant but substantively meaningless.

Background and research hypothesis

In 2009, a private Japanese university in southwestern Japan introduced a vocabulary curriculum in an attempt to improve scores on the reading section of the TOEIC Bridge test (ETS, 2008), following disappointment at modest gains in previous years. Students at this institution take two compulsory 90-minute English lessons per week, a "Communication" class with a native speaker of English (NST), and an "English" class with a Japanese teacher of English (JTE). NSTs and JTEs are respectively held responsible for improving listening and reading scores on the TOEIC Bridge test. Despite anecdotal evidence strongly pointing to the fact that most, if not all JTEs were already teaching vocabulary, and an institution-wide compulsory online vocabulary homework curriculum for low-level students was in place, a faction of NSTs were insistent that this was inadequate and that explicit prescriptive vocabulary instruction should be introduced into the English Communication classes, based largely on Folse's (2004) endorsement of vocabulary lists.

The *Longman English-Japanese Dictionary* (LEJ) (2006) was adopted as a mandatory supplementary text for all first-year students. An expedient wordlist for instruction was seen in the approximately 550 overlapping words listed in the LEJ as appearing in both the most frequent 1000 spoken and written wordlists. Lists containing three meanings for each target word and bilingual example sentences were distributed to teachers in September of 2008 for instruction and testing in 2009. The availability of bilingual example sentences raised the possibility of contrastive analysis of usage between English and Japanese without the need for bilingual teachers. This allowed the framing of a research hypothesis:

Mechanical output (MO) tasks based on bilingual example sentences provide greater long-term vocabulary gains than creative output (CO) tasks requiring creation of original meaning.

Task design

The vocabulary tasks, influenced by Laufer and Girsai's (2008) use of contrastive analysis, aimed

to draw attention to target word forms and meanings and involved the following steps:

- Copy target words from projector to work sheets
- Compare gapped English example sentence with ungapped Japanese translation and choose target word to complete the gap
- Take a multiple-choice spelling test of target words
- Take a multiple-choice gap completion test

Another multiple-choice gap completion test was administered at the beginning of the next class as a review. Preliminary analysis in 2009 focused on classes taught by one teacher. Control (CO) and treatment (MO) groups were assigned identical homework, the vocabulary homework, and a diary writing assignment. The vocabulary homework comprised word-search, word-scramble, gapped sentence completion, and crossword tasks. The diary task required students to write as much as possible about five interesting events from the previous week. Although the vocabulary homework contributed 10% to the semester grade for the MO group, it was not collected from the CO group. Instead, explicit vocabulary instruction was replaced by a diary review and discussion task.

The commonsensical expectation was that the MO group would show greater vocabulary gains, the question being whether these would be large enough to justify spending such a large proportion of class time on mechanical tasks. Surprisingly, among the target group of low-level students, the CO group showed slightly better vocabulary gains over the first semester (Holster & DeLint, 2010), although the differences overall were not statistically significant ($t(120) = -1.112$, $p>.05$). Given that the vocabulary treatment targeted very high frequency words for low-level learners, the empirical results did not support the research hypothesis. Additionally, the vocabulary tasks imposed a considerable workload on teachers, and the two teachers using this material had impressions of poor engagement from the MO groups, buttressed by low attendance and high attrition. Thus, although quantitative and qualitative evidence supported discontinuation of the MO tasks, the 2009 CO group had been given vocabulary homework and weekly review tests, making the effectiveness of the CO tasks alone unclear. Therefore, in 2010 the vocabulary homework was discontinued, allowing comparison between the 2009 MO group and a 2010 CO group without exposure to the vocabulary materials. In order to provide a larger sample size and greater generalizability, students taught by a second teacher were included in the current study. This teacher used the MO tasks in 2009 but not in 2010, instead assigning short personalized compositions based on coursebook speaking practice activities for homework, later used in class for small group presentations and transcription or note-taking tasks.

Research instrument and methodology

As TOEIC Bridge post-test results were not available until the end of the second semester, 50-item vocabulary tests were administered as pre-tests and post-tests at the beginning and end of the first semester. A clustered word deletion format was chosen to match the format of the weekly review tests, using example sentences from the LEJ (2006), as shown in Figure 1. Only sentences where all words except the tested word came from the first 1000 in the General Service List (West, 1953) were used to minimize the effect of non-target vocabulary on item difficulty.

1)	What's your _____?	A)	disassociate
2)	The country has serious _____ problems.	B)	fresh
3)	The teacher divided us into _____ of five.	C)	groups
4)	The red light _____ "stop."	D)	means
5)	We _____ about $100 a week on food.	E)	name
		F)	settles
		G)	social
		H)	spend

Figure 1. Semester test example item cluster

The two test forms used each comprised 50 items in 10 clusters of five items each, with eight multiple-choice answer options per cluster.

Analysis of the vocabulary pre-tests and post-tests was conducted using the Winsteps software

package for Rasch analysis (Linacre, 2010), providing detailed analysis of test performance and the interval level measurement required for statistical comparisons of the results (Bond & Fox, 2007). Winsteps provides outputs in a probabilistic unit called the "logit", or log-odds unit, so outputs were specified on a scale of 1 logit = 10, with mean item difficulty specified as 50, providing a user-friendly score range. Measurement based on odds-ratios provides very practical measures of effect sizes (Field, 2009, pp. 699-700), and in probabilistic terms, a person with ability of 50 would have a 50% expectation of success on an item of mean difficulty, increasing to 73% for an item of difficulty of 40, and 27% on an item of difficulty 60. Engelhard (2009) reports a threshold of .30 logits as commonly considered a substantively meaningful effect size, equal to 3.0 on the score scale used here.

Table 1 gives summary statistics from the anchoring analysis used to measure the difficulty of the items in order to anchor them at specified values. Anchoring the items in this way allows person ability to be directly compared between pre-test and post-test scores, showing relative gains in vocabulary knowledge. The separation index of 2.94 means that the ratio of measurement error to the range of person ability is small enough that this test can separate the persons in the anchoring sample into at least two distinct bands. The sample of persons in the anchoring analysis had a much larger range of ability than the research sample, so the reported separation index and person reliability of .90 must be considered an upper limit for this test. The separation index and person reliability are sample dependent (Bond & Fox, 2007), so limiting the research sample to low-level learners drastically constrains the range of person ability, leading to lower reported reliability and separation when this sample is analyzed in isolation.

The research sample was limited to first-year students with TOEIC Bridge scores below 100, the target group for the MO tasks, giving a convenience sample of three classes from each year. Attendance and attrition are often problematic with these low-band classes, but this proved especially so of the MO group, as shown in Table 2. Of the 189 Japanese students assigned to the six classes, five students with less than eight correct responses were eliminated from the pre-test as the expected score from random guessing with this test format is six. Following pilot administrations, students were allowed 25 minutes to complete the test, but some did not attempt to answer difficult items while others spent large amounts of time on difficult questions, resulting in incomplete answer sheets. Missed responses were coded as incorrect, following assumed practice in *TOEIC Bridge* tests, but items with both a correct and incorrect response were coded as missing data. With 25 items printed on each side of the question sheet, students who did not attempt the final 20 items were assumed to have been plodding or sleeping, eliminating four students, all from the MO group. Of the 92 MO group students, 68 satisfactorily completed the pre-test, compared with 80 of the 97 eligible CO group students. However, only 47 MO group students completed the post-test, compared with 72 CO group students. Ultimately, 37 MO students completed both tests, compared with

Table 1. Vocabulary test anchoring administration performance

	Total Score	Count	Measure	Model Error	Infit MS	Infit Z-Std	Outfit MS	Outfit Z-Std
Mean	26.2	49.6	51.80	3.61	1.00	.0	1.08	.1
SD	9.7	4.0	11.69	.34	.20	1.1	.56	1.2
Max.	76.0	100.0	78.94	5.08	1.75	4.2	4.74	5.4
Min.	8.0	30.0	27.26	2.35	.52	-3.4	.28	-2.4

Real RMSE 3.77	True *SD* 11.06	Separation 2.94	Person reliability .90
Model RMSE 3.62	True *SD* 11.11	Separation 3.07	Person reliability .90

SE of person mean = .24

Note. n = 2325, Scale of 1 logit = 10.00, Mean item difficulty = 50.00, Person raw score-to-measure correlation = .97 (approximate due to missing data), KR-20 person raw score reliability = .80 (approximate due to missing data)

67 CO students, attrition rates of 60% and 31% respectively, leaving a sample of 104 of the 189 eligible students, an overall attrition rate of 45%.

Table 2. Summary statistics for vocabulary and output groups

	Group	*n*	Mean	*SD*
Pre-Test	Vocabulary	37	41.42	6.84
	Output	67	42.92	8.15
Post-Test	Vocabulary	37	44.94	7.33
	Output	67	47.94	7.09
Gain	Vocabulary	37	3.53	7.24
	Output	67	5.02	6.31

Results

The pre-test and post-test scores are summarized in Table 3, and effect sizes are shown in Table 4. Logit gains greater than .30 can be considered substantively meaningful, while Hattie (2009, pp. 7-10) favors Cohen's *d* as an effect size measure, with .40 argued as a guideline for useful interventions, indicating a gain equivalent to 40% of the pooled standard deviation. The difference between pre-test mean scores of 1.50 scaled points (.15 logits) was substantively small, and an independent-samples *t*-test did not find statistical significance ($t(102) = -.950$, $p>.05$, $r = .09$, $d = -.18$), so the two groups were of similar ability prior to instruction. Both groups showed substantive gains in vocabulary knowledge, 3.53 scaled points (.35 logits) for the MO group ($d = .50$) and 5.02 scaled points (.50 logits) for the CO group ($d = .60$), as shown in Table 4. Gains of these magnitudes mean that a person having a 50% expectation of success on an item in the pre-test would have respectively a 59% and a 62% expectation of success on an item of the same difficulty in the post-test. The .15 logit smaller gain of the MO group compared with the CO group was neither statistically nor substantively significant ($t(102) = 1.098$, $p>.05$, $r = .11$, $d = -.22$),

implying an expectation of success falling from 50% to 46%, or a lag of 22% of the pooled standard deviation. Thus, the MO tasks did not result in vocabulary gains substantively or statistically significantly greater than the CO tasks, justifying rejection of the research hypothesis.

Table 3. Summary statistics for MO and CO groups

	Group	*n*	Mean	*SD*
Pre-Test	MO	37	41.42	6.84
	CO	67	42.92	8.15
Post-Test	MO	37	44.94	7.33
	CO	67	47.94	7.09
Gain	MO	37	3.53	7.24
	CO	67	5.02	6.31

Table 4. Effect sizes of score gains same as table 3 format

Group	*n*	Logit	Odds Ratio	*r*	*d*
Combined	104	.45*	61/50	.29	.60*
MO	37	.35*	59/50	.24	.50*
CO	67	.50*	62/50	.31	.66*
Difference (MO-CO)		-.15	46/50	.11	-.22

* Indicates substantively significant effect size

Discussion and conclusions

The hypothesis that mechanical output (MO) tasks provide greater long-term vocabulary gains than creative output (CO) tasks was not supported. Both the treatment (MO) and control (CO) groups showed substantively significant gains in vocabulary knowledge. Although the MO group showed smaller gains than the CO group, the difference between them was neither substantively nor statistically significant. However, preparing and administering the MO tasks placed a heavy workload on teachers, and both teachers' impressions were that students found them dull, consistent with Folse (2004). The attrition rate of 60% for the MO group versus 31% for the CO group was of great concern, raising the possibility that the dull nature of MO tasks led to differential attrition of higher aptitude learners from the MO group. However, for an equal attrition rate between the groups and for the MO group to better the CO group's gains by a substantively significant .30 logits, an extra 29% of the MO students with mean gains of approximately 1.25 logits would have been needed to be retained. An effect size of 1.25 logits means that an expectation of success of 50% on the pre-test would rise to 78% on the post-test, an implausibly large reversal. The evidence from this study thus justifies a conclusion that this treatment was not effective for students of this level at this institution.

However, a number of concerns would need to be addressed before wider generalizability was warranted. These students had previous exposure to English at high school, took compulsory online vocabulary homework, were probably taught vocabulary by JTEs, and had incidental exposure to vocabulary from the coursebooks used by JTEs and NSTs, making discussion of specific mechanisms of acquisition highly speculative. It is plausible that CO served as a mechanism to consolidate acquired learned knowledge, but no claim is justified that such incidental exposure will be an efficient mechanism for learning previously unknown low-frequency words, so one important future research direction will be to compare CO and MO tasks for lower frequency vocabulary that students are less likely to encounter incidentally.

The causes of the high attrition rate could not be investigated for this report, so qualitative investigations of this should be undertaken in future studies. It is possible that the CO tasks led to the positive goal appraisals theorized to underlie sustained deep learning (Schumann & Wood, 2004), while MO tasks were perceived as dull busy-work by students, leading to demotivation and high attrition. However, many other factors may have contributed to the differential attrition rate, including social effects leading to a small number of individuals disproportionately affecting the behavior of the group. If this did occur, which the authors consider plausible, the chance assignment of a few exceptionally motivated or unmotivated students who influenced others to drop out or continue attending class may have contributed to the differential attrition. Resolving such questions would require qualitative research far beyond the practical scope of this

investigation, but essential if the achievement or lack of achievement of low-level learners such as these is to be understood.

This study also highlights important considerations for teachers seeking to develop classroom tasks based on experimental research findings. One is awareness of the problem of publication bias, where positive findings supporting the research hypothesis are emphasized over studies with null results. An intervention found to be successful in a small number of experimental studies may have failed on numerous other occasions not considered worthy of publication, so multiple replications are needed before the relative effectiveness of interventions can be judged. Secondly, findings from experimental studies cannot be automatically assumed to generalize to classroom contexts, nor can classroom studies conducted in one context be assumed to generalize to other contexts. The results of the current investigation support the view that new interventions should be carefully piloted to gather quantitative and qualitative evidence of effectiveness under local conditions before large-scale adoption, regardless of previous research findings.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum Associates.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602. doi: 10.1177/0013164408323240

ETS. (2008). *TOEIC Bridge user guide.* Retrieved from <ets.org/Media/Tests/TOEIC_Bridge/pdf/TOEIC_Bridge_User_Guide.pdf>.

Field, A. P. (2009). *Discovering statistics with SPSS* (3rd ed.). London: Sage.

Folse, K. S. (2004). *Vocabulary myths*. Ann Arbor: The University of Michigan Press.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York: Routledge.

Holster, T. A., & DeLint, D. F. (2010). *Pushed output and vocabulary gains*. Paper presented at the JALT Pan-SIG 2010, Osaka Gakuin University, Osaka, Japan.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716. doi: 10.1093/applin/amn018

Linacre, J. M. (2010). *A user's guide to Winsteps* 3.70.02. Retrieved from <winsteps.com/winman/index.htm?copyright.htm>.

Longman eiwajiten: English-Japanese dictionary. (2006). Harlow: Pearson Longman.

Meara, P. M. (2002). The rediscovery of vocabulary. *Second Language Research, 18*(4), 393-407. doi: 10.1191/0267658302sr211xx

Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research, 6*(3), 183-222. doi: 10.1191/1362168802lr108oa

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158. doi: 10.1093/applin/11.2.129

Schumann, J. H., & Wood, L. A. (2004). The neurobiology of motivation. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert, & L. A. Wood (Eds.), *The neurobiology of learning.* (pp. 23-42). London: Lawrence Erlbaum Associates.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*(3), 371-391. doi: 10.1093/applin/16.3.371

Thompson, B. (1999). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191-196. doi: 10.1177/095935439992007

West, M. P. (1953). *A general service list of English words*. London: Longman, Green & Co.

Trevor Holster has taught English in Japan for over 15 years. His research interests include vocabulary acquisition and peer assessment.

Darcy de Lint has been teaching English in Japan for 20 years. He has research interests in the areas of pushed output, communication strategies and peer assessment