Abstract

Instructors of English in Japanese universities are first teachers, second researchers, and sometimes, a distant third, readers of literature. Keeping abreast of the latest studies involving all the relevant fields of second language acquisition (SLA) is almost a full time job in and of itself. This paper is my attempt to ease the burden of my fellow instructors somewhat by providing a synthesis of relevant and crucial research that has been conducted in speaking a second language over the last two decades. It is my hope, through highlighting important theoretical works to lead my fellow instructors and researchers in grounding themselves firmly in the wide body of theory that supports and encourages SLA research on language learning output.

Keywords: SLA, speaking/output, synthesis, theoretical frameworks

In de Bot’s (1996) article evaluating and expanding upon Swain’s output hypothesis, he uses Levelt’s L1 speaking production model (which introduces the concept of the lexicon, conceptualizer, formulator, and articulator), focusing specifically on the role of the lexicon (p. 537), and Anderson’s skill-acquisition theory, which pertains to declarative and procedural stages of knowledge development or simply the development from controlled to automatic processing (p. 546), as theoretical underpinnings for his explanation of the critical nature of output in second language acquisition. In effect, the language learner’s speaking is influenced through four cognitive language learning activities: noticing,
hypothesis testing, metalinguistic function, and enhancing fluency. In noticing, learners become aware of gaps in their knowledge, which can lead to greater attention to the gaps or problems, and an incentive to correct or fill the gap. “Output serves to test hypotheses” in that whether using internal or external speech, the “phonetic plan generated by the formulator is fed back into the speech-comprehension system to monitor the internal speech” (de Bot, 1996, p. 551). Learners use internal speech to test speech via their understanding of language discourse rules, before using in external speech with the goal of eventually improving patterns of language use. The third function of output is metalinguistic in nature: learners of a similar level can talk about the gaps in their knowledge to problem-solve. Last, enhancing fluency through practice plays a “direct role in enhancing fluency by turning declarative knowledge into procedural knowledge” (p. 553): In affect, creating automaticity for the learner so that attentional resources can be devoted elsewhere. These four functions of output, within Levelt’s speaking model framework, provide a basis for examining the second language acquisition literature on speaking.

Muranoi (2007) took a closer look at Swain’s output hypothesis, examining empirical studies in a meta-analysis of the subject. In his treatment of the output hypothesis’ effect on L2 communicative competence development, he also addressed Levelt’s speaking production model and Anderson’s skill-acquisition theory. Muranoi’s findings boil down to the simple statement that “output practice (i.e. any activity designed to provide L2 learners with opportunities to produce output) is effective for developing L2 learners’ well-balanced communicative competence” (p. 76). In the following synthesis three broad categories of the output hypothesis are used to categorize the studies reviewed: enhancing fluency, metalinguistic function, and noticing.
Enhancing Fluency

What language skill variables best predict speaking proficiency? According to De Jong, Steinel, Florijn, Schoonen, and Hulstijn’s (2012) study examining skill variables, such as vocabulary, grammar, linguistic processing (as measured by reaction time), and pronunciation (such as speech sounds, word stress, and intonation), vocabulary (Levelt’s “lexicon”) and intonation were the strongest predictors of speaking proficiency (p. 29). These two variables, according to De Jong et al. explained nearly 76% of the variance in their participants’ communicative speaking success (p. 32). As the authors argued that speaking proficiency is “componential in nature” (p. 10), it would certainly be worthwhile from a teaching standpoint to know which skills are most critical to what is understood as speaking proficiency. However, one possible qualifier to the usefulness of this study could very well be the lack of pragmatic knowledge assessment. The inclusion of a pragmatic knowledge assessment would strengthen De Jong et al.’s findings and further the field’s understanding of the nature of speaking development.

In a similar study to De Jong et al., Iwashita, Brown, McNamara, and O’Hagan (2008) also investigated the nature of English speaking proficiency while developing a rating scale for the TOEFL-iBT. Iwashita et al. used five measures to represent speaking proficiency: grammatical accuracy, complexity, vocabulary, pronunciation, and fluency. And much like De Jong et al. found that vocabulary and intonation were key facets of speaking proficiency, Iwashita et al. determined that vocabulary and fluency represented the strongest factors of speaking proficiency, though all five measures had an impact on participants speaking proficiency as determined by their spoken test scores. Interestingly, the authors’ results conflicted with earlier studies which were focused on grammatical accuracy as the most important feature of oral proficiency scores. They suggest that “an exaggerated emphasis on grammatical accuracy, reflected in attitudes and
behavior of many learners and teachers, is misplaced” (p. 47). To be sure, the authors did not conclude that grammatical accuracy be eliminated as a feature of oral proficiency evaluation, rather they simply advised that production features such as fluency and vocabulary knowledge not be ignored by teachers and testers alike. By this time, the importance of vocabulary to SLA should no doubt be obvious, thought to some extent it appears that within the context of the Japanese English-language classroom the old paradigm of “exaggerated emphasis on grammatical accuracy” still holds strong.

Skehan (2009) stressed the importance of adding vocabulary to the complexity-accuracy-fluency (CAF) paradigm. In this study he argues that by adding lexis to the CAF, new generalizations as organized by Levelt’s model can be made between accuracy and fluency — Skehan measures fluency by looking at pausing, repairs, speed of speech, and length-of-run much like Iwashita et al. (2008). Three generalizations that Skehan proposes are “the need to organize and integrate more demanding information,” the “need to retrieve and make available more difficult lexis,” and the “capacity to draw on clear macrostructures” (p. 519) — much as de Bot (1996, p. 553) argues for the need to automatize lexis and grammar so as to redirect attentional resources elsewhere. Skehan also discusses the trade-off hypothesis, which simply refers to the rarity of “(s)imultaneously advantaging all three (CAF) performance areas” (p. 512), where one of the three performance areas usually suffers for the other two areas to gain. Like De Jong et al. (2012), Skehan’s study is also missing a pragmatic component.

Another investigation of CAF is Revesz, Eikiert, and Torgensen’s (2016) examining the extent to which complexity, accuracy, and fluency predict adequacy, and whether or not proficiency and task type moderate these relationships. Here adequacy is whether or not a learner successfully achieved a task’s goals sufficiently. Using a Rasch analysis they found a large variation in communicative adequacy. Interestingly, breakdown frequency — as a measure of fluency — was
one of the strongest predictors of adequacy providing nearly 15% of variance. In advanced proficiency speakers, false starts were also found to be a predictor of communicative adequacy. All three CAF dimensions were, to a limited degree, predictors of adequacy and task type was not found as an influence between the three linguistic features measured and adequacy. This last feature of the study, the lack of influence of task type between CAF and adequacy, might have been due to the tasks not being sufficiently different.

This next study was not focused on CAF per se, but de Jong and Perfetti (2011) looked at fluency development, thus placing their study within the CAF framework. The focus of their study was on speech repetition, through the experimental use of the 4/3/2 timed speech task (Ellis, 2009 similarly looked at the effect of tasks on fluency, using planning instead of repetition), and its impact on fluency development. The authors discussed the need to limit the cognitive processing demands of working memory by means of proceduralizing production rules (p. 537), much as discussed by de Bot (1996). Evidence supports fluency gains from students who repeated their speeches, and these gains were transferred to new topics and persisted over the course of the study. Thus, the authors found evidence that fluency gains can be transferrable and persistent through use of the repetition task used in this study. The study would have been more convincing if the data were shared and qualitative details included, but the possible impact of chunking on fluency development is possibly an important and underutilized aspect of language learning and the teaching that goes with it.

Du (2013) also examined the proceduralization of linguistic knowledge, specifically in observing students’ fluency development in the context of studying abroad time-on-task. Du found that those students who devoted the most time-on-task hours while studying abroad in China had dramatic increases in their fluency through the first period of the study, but initial gains plateaued as the study continued. The author suggested that time-on-task does promote fluency gains,
though the small population size of his study might not allow his findings to be
generalizalbe. Like de Jong and Perfetti (2011) before him, Du used the length of
utterance as a measure of fluency though mainly because “research does not as yet
have a well-established criterion of hesitation indicators for Chinese” (p. 141).
The qualitative aspects of this study did help to pinpoint an interesting anomaly
related to individual differences, that the wide variability among students in
groups and between groups could very well be due to personality traits of the
participants.

Baker-Smemoe, Dewey, Bown, and Martinsen (2014) studied fluency by
looking for relationships between L2 proficiency and utterance fluency. They used
excerpts from the ACTFL Oral Proficiency Interviews (OPI) to examine various
measures of fluency: speech rate, number of hesitations, number and length of
pauses, number of length of runs, and the number of false starts. They examined
native speakers of English who spoke a variety of L2s and their research suggests
that the measures of speed rate, meant length of run, and number of pauses can
predict broad differences in L2 proficiency, despite not correlating to OPI ratings.
They also found that L2 utterance fluency differs based on the L2 studied,
suggesting that fluency generalizations across languages might not be possible.
Somewhat in contradiction to other studies on fluency, Baker-Smemoe et al.
suggested that:

instructors need not be concerned by aspects of fluency such as false starts
(stopping mid-word or mid-sentence) or hesitations (self-correction, self-
repetition, using natural fillers such as “um” and “uh,” etc.), since these occur
regularly in native speech and do not appear to relate closely to L2
proficiency. (p. 723)

Obviously, this assertion, if it holds merit, impacts the validity of multiple studies
that use false starts and hesitations as a marker of fluency.
Metalinguistic Knowledge

Unlike previous studies on CAF discussed in this paper (Revesz et al., 2016; Skehan, 2009), Ellis’ (2009) meta-analysis studied strategic planning and its effect on complexity, accuracy, and fluency in variety of studies. Generally, in all but two of the studies examined, strategic planning was found to have a positive effect on fluency. The results for complexity and accuracy were more mixed, though evidence does support the assertion that strategic planning does allow language learners to produce more complex language. Accuracy was a more difficult feature to predict as nearly a third of the studies looking at strategic planning found that there was no effect on grammatical accuracy when students were given strategic planning time. Some confounding factors in accuracy might be learner proficiency, attitudes towards tasks and the nature of the tasks themselves. These findings appear to support the trade-off hypothesis as discussed in Skehan (2009) — that is learners focus on one aspect of performance at the expense of others (Ellis, p. 498) — and might also direct attention to the ease with which fluency is influenced in terms of learner gains, while accuracy appears to be more difficult to move. Obviously, lexis and pragmatics are not represented in the studies examined by Ellis, so their impact on the speaking proficiency continues to go under researched.

Norris and Ortega (2009) provided another meta-analysis investigating CAF, but narrowing the focus to complexity and how and why researchers measure it. They argued that in:

the measurement of syntactic complexity, there are some metrics that are redundant if used together, because they tap the same measurable dimension of the construct and, conversely, there are other measures that are distinct and complementary and thus can be best used and interpreted together, because they tap different dimensions of complexity. (p. 562)

In addition, the authors introduced a measurement process (p. 557) that is very
similar to Levelt’s model (as introduced in de Bot, 1996, p. 536). They found, somewhat like Revesz et al. (2016) did, that CAF variables have different effects on proficiency levels. Moreover, as Spoelman and Verspoor (2010) contend in their study, L2 development when examined through the CAF framework is not linear and, additionally, measurements used to assess language development should “provide learner learner-, task-, and L2 form-sensitive accounts of the local SLA ecology, given the ways in which these factors moderate the observations we might be making about CAF” (p. 574).

This next study, also involving CAF, is like a duck out of water in regards to the topic of this paper — a speaking synthesis. Instead, what we find in Spoelman and Verspoor’s (2010) longitudinal study is a case study looking at writing samples and analyzing those samples within the CAF paradigm. Given the nature of writing and generous online-planning time associated with such tasks, this study marginally fits within a metalinguistic category (if it indeed fits within this synthesis at all). Briefly, the authors found evidence in writing to support the generalizations that L2 development is not linear and that complexity should not be viewed as a single construct, instead as a complex variable made up of many different components (p. 551). No relationship was found between accuracy and complexity over time, and any correlations between the two were up-and-down over the length of the study (p. 550).

Kormos and Trebits’ (2012) study researching the role of task complexity, modality, and aptitude in narrative task performance is related to several of our previous studies; They look at task type and planning effects on oral production as did Ellis (2009), the benefits of task feature manipulation when examining individual differences a la Norris and Ortega (2009), and they use Levelt’s L1 speech production as discussed in de Bot (1996), specifically the concepts of conceptualizer and formulator. Perhaps unsurprisingly, in examining a written modality versus an oral modality, the authors found a higher proportion of error-
free clauses and more varied vocabulary in writing tasks. Overall, writing samples had more accurate sentence structures and varied vocabulary, but the oral performances were similar in relation to syntactic complexity. Thus, similar to Ellis’ findings, planning time as related to writing can and does provide greater accuracy, but as was usually the case in the previous studies of CAF tasks related to speaking the trade-off hypothesis factors in to gains — that is students improved on two of three factors, but showed losses in the third.

The final study in the metalinguistic category, Hernández’s (2010) examination of American students studying abroad for one semester in Spain, will only receive brief attention due to its statistical flaws. However, despite issues with the quantitative data presented in the study, it is the only study to look at an important attentive factor in speaking development: integrative and instrumental motivation. The participants in Hernández’s study showed no instrumental motivation to interact with Spanish culture; instead it was integrative motivation that appeared to be a driving factor for students to actively interact with the culture around them. As the sole study looking at motivation’s effect on second language speaking acquisition it is important to keep in mind that though underrepresented in the literature of this synthesis paper, it is still one of many crucial factors that should be considered when discussing speaking proficiency acquisition.

Noticing

An interesting longitudinal study that looks at a single speaker through the lenses of CAF over a year is found in Polat and Kim’s 2014 study. They looked at the naturalistic language learner, or untutored development, in a Turkish immigrant, Alex. The participant’s language learning was compared with three native English speakers and the authors found that Alex’s vocabulary showed the clearest improvement over the length of the study, while his syntactic complexity
did not show any real gain and his accuracy also did not appear to improve (pp. 197-198). As discussed in de Bot’s (1996) treatment of the output hypothesis, language learners require a chance to notice gaps in their knowledge, process the lack, and then attempt correction. In the case of naturalistic learners, David Beglar has often mentioned (personal correspondence, 2015), their cognitive processing capacity is overloaded with attempting to understand incoming input, leaving little capacity for noticing gaps and a general early fossilization or plateauing of the L2 speaking ability. The language learner might be communicatively competent within the domains that they most frequently need to use their L2, but this does not translate into more than minor language gains over time.

So how can teachers help their students with noticing the gaps in their language use? According to Stafford, Bowden, and Sanz (2012; this study could equally be categorized under metalinguistic) explicit feedback is a good place to start. Stafford et al. used the theoretical framework of the competition model, whereby the “strength of cues in linguistic input governs the mapping between form and function” (p. 748), to examine the effects of pre-practice grammar explanation (in some ways like the planning task introduced in Ellis, 2009) and feedback on participants’ learning of Latin morphosyntax. Pre-practice grammar explanation alone was not enough to provide a benefit to learners, but explicit metalinguistic feedback coupled with practice led to gains in productive abilities: Answering the research question of whether or not explicit corrective feedback during practice influences initial learning. In the modern “communicative” language classroom context, the fear of explicit corrective feedback might very well be a detriment to students’ second language acquisition.

Lyster and Saito (2010) provided a final meta-analysis on the most effective ways in which to help students’ noticing. They looked at 15 studies covering recasts, prompts, and explicit correction to determine how effective corrective feedback (CF) is on L2 language development in the classroom and, if effective,
how does it vary according to: types of CF, types and timing of outcome measures, instructional setting (second language vs. foreign language), length of treatment, and learner’s age. Overall they found that CF does have a significant impact on performance (d = 0.74) (p. 281). All the types of CF had a significant effect on performance, though in the study the “relative effects of explicit correction remained indistinguishable from those of recasts and prompts” (p. 290) which might very well be due to the difficulty in a meta-analysis of distinguishing whether or not instructors continuously managed to clearly use each CF type without mixing or overlapping. In any event, the results, much like Stafford et al.’s (2012) indicate that corrective feedback is beneficial to SLA and what is more they seem to be effective no matter the setting.

**Conclusion**

Swain’s output hypothesis is only one cog in the complex language learning machine that is the human brain. Mackey (2007) in a meta-analysis of the various theories surrounding interaction, attempts to tie together many of the disparate ideas of social theories on learning, such as Krashen’s input theory, Swain’s output hypothesis, and Long’s interactional hypothesis. As any researcher who has read about Wes or naturalistic learner outcomes can attest, input and output are not the only components of second language acquisition. Factors such as willingness to communicate, integrative and instrumental motivation, speaking self-efficacy, working memory, salience, explicit instruction and corrective feedback, just to name a few, are all likely important to the process of proficient language development. As language professionals and classroom teachers, it is important that input, output, interaction, and all the myriad of components be considered when we examine speaking development proficiency in the field of SLA.
References


second language: Perspectives from applied linguistics and cognitive psychology (pp. 85-110). Cambridge, UK: Cambridge University Press.


