Proficiency effects in L2 processing of English number agreement across structurally complex material
Bronwyn G. WILSON and Edson T. MIYAMOTO

Introduction

Models of second language acquisition and their predictions of ultimate attainment

Models of second language acquisition focus on different mechanisms to explain differences between the performance of L1 (first language) and L2 (second language) speakers. Knowledge deficit models posit that when L2 speakers do not display L1-like behaviour, it is because L2 speakers do not have the level of grammatical knowledge that L1 speakers do. Parsing deficit models posit that differences are due to a difference between the L1 and L2 parser, that is, the sentence comprehension algorithm. These models can further be divided into those that predict L2 and L1 behaviour will never fully converge, and those that predict they eventually can.

An example of the knowledge deficit type of model is the Full Access / Full Transfer model (Schwartz & Sprouse, 1996). This model claims that all of the L2 grammar can, though not necessarily will, be ultimately acquired, the assumption being that this knowledge is sufficient to display L1-like performance. Differences between L1 and L2 speakers are predicted not to be persistent, but to decrease with increased L2 grammatical knowledge.

A model that predicts persistent deficiencies in grammatical knowledge is the Failed Functional Features hypothesis (Hawkins & Yuet-Hung, 1997). It posits that some types of grammatical knowledge cannot be acquired by L2 learners, these being functional features that are different in the L2 from the learner’s L1. For these features, L2 speakers will never display L1-like behaviour, because of their imperfect grammatical knowledge.

In contrast, parsing deficit models state that the L2 parser is different. The Shallow Structure Hypothesis (Clahsen & Felser, 2006), states that differences in parsing result from L2 speakers’ exclusive reliance on good enough syntactic representations (Ferreira, Baily & Ferraro, 2002), that lack the depth and detail required to show L1-like performance. These are structures computed to deal with the demands of processing, and do not represent a deficit in grammatical knowledge.

Previous results regarding Japanese speakers' sensitivity to number agreement

One construction that has been used to test these models is number agreement for Japanese-speaking learners of English. English verbs agree with their subject in number in some environments, while Japanese verbs are not marked for number. Being a form of functional morphology, number agreement is a good candidate to test the predictions of persistent deficits in grammatical knowledge. Further, it has been specifically stated that subject-verb agreement with an intervening prepositional phrase is too syntactically complex to be fully calculated by L2 speakers (Keating, 2009; Wen et al., 2010), while they can acquire within-NP agreement. This makes it a good candidate to test the Shallow Structures Hypothesis. Hence, in the current study, we use sentences with both number agreement and intervening prepositional phrases.

Many previous studies have examined the processing of agreement by looking for sensitivity to violations of grammatical agreement. The logic is that if Japanese learners of English are slower to read ungrammatical versus grammatical agreement, this indicates that they are sensitive to agreement.

Jiang et al. (2011) found that Japanese-speaking participants did not display sensitivity to violations of number agreement within noun phrases, while Russian-speaking learners of English did. They took this to support a persistent deficit of grammatical knowledge model of acquisition, as Russian has number agreement, while Japanese does not.

Wen et al. (2010), in contrast, found that Japanese learners of English were sensitive to agreement violations within NPs. They use these results to argue for the Full Access / Full Transfer model and against the Failed Functional Features Hypothesis. However, if the structure of a NP
could be considered simple, as it has been (Keating, 2009), these results are also in line with the Shallow Structure Hypothesis.

Yamada and Hirose (2012) also found sensitivity to agreement errors in a construction that involved a complex-NP-containing relative clause, though the elements involved in agreement were adjacent. These results are difficult to interpret, however, as there was only a significant effect of grammaticality when a singular subject was followed by a plural verb, and not vice-versa. This experiment is partial evidence against the Failed Functional Features Hypothesis, as some grammatical knowledge of agreement was shown, but it does not provide evidence related to the Shallow Structure Hypothesis, as L1-like performance in agreement between adjacent elements is possible according to that model.

**Importance of proficiency in previous results**

Both the studies that have found sensitivity to violations of number agreement in Japanese learners of English have found that sensitivity is contingent on participants’ proficiency (Wen et al., 2010; Yamada & Hirose, 2012). The measures used to look at proficiency in this kind of research have varied, with some using all or parts of the TOEFL (Jiang, 2004; Yamada & Hirose, 2012), some self-assessed proficiency (Hoshino et al., 2010), some referring to the amount of exposure to the L2 (Keating, 2009), and still others using a C-test (Wen et al., 2010), or cloze test (Jiang et al., 2011)

In the current study we suggest the C-test as an effective and practical measure, and compare it to the TOEFL-ITP.

**Experiment: Sensitivity to ungrammaticality**

**Participants**

Twenty-six Japanese-L1 learners of English were paid to participate. All reported that their formal English education started at or after the age of 11 years.

**Materials**

Sixteen pairs of sentences as in (3) were seen by the participants.

(3) a. The cakes with the cream were baked for forty minutes (grammatical)

b. *The cake with the cream were baked for forty minutes (ungrammatical)

The sentence pairs were exactly the same except for the first noun, which in the ungrammatical condition was singular. The 16 pairs of sentences were distributed according to a Latin Square design such that each participant saw eight grammatical and eight ungrammatical sentences, interspersed with 80 grammatical filler sentences.

**Procedure**

Sentences were presented on a computer monitor using a moving window non-cumulative self-paced reading procedure. Button presses in lieu of reading times were recorded for each word. Each sentence was followed by a true / false comprehension question, which never related to number agreement, so as not to draw participants’ attention to the aim of the experiment.

**Data treatment and analysis**

Reading times shorter than 100ms and longer than 10s were discarded, and trimming was also undertaken based on model residuals (Baayen, 2008). This process affected 1.6% of the data. Reading times were analysed using mixed-effects models in the R version 3.2.1 package LME4 1.1-8 (Bates et al., 2015), constructed using backward selection (Baayen, Davidson & Bates, 2008). The independent variables were grammaticality and C-test score, as well as word length in all regions except those containing only the and were. C-test scores were centred and scaled before being entered into the analyses. P-values were calculated using the anova function of the lmerTest package 2.0-25 (Kuznetsova, 2015), and are based on F-values and Satterthwaite-corrected degrees of freedom. Only the items where the comprehension question was answered correctly were included in the reading time analysis.
**The C-test**

The C-test is a fill-in-the-blanks style test, where the second half of every second word has to be completed. The test was taken from Babaii and Shahri (2010) and consisted of five texts with a total of 100 blanks. The test was limited to 15 minutes based on pilot testing.

**Results**

**Comprehension questions**

Average test item comprehension question accuracy was 88% (range=77%-90%), and no participants were eliminated on this basis. A mixed-effects logit model revealed that comprehension question accuracy did not differ across conditions, (ps>.2).

**Reading times**

At the first noun (cake(s)) there was a main effect of grammaticality. This was the difference between plural and singular nouns, singular nouns being read faster (β=-390, p<.001). There was also a main effect of C-test, as participants with higher C-test scores read faster (β=-10.1, p=.007).

The critical region was the verb (were), where grammaticality first became apparent. This region had no significant differences (see Wagers et al., 2009, for a similar spillover pattern with native speakers). In the following spillover region (baked), there was a main effect of C-test (β= -9.2, p=.002), as higher C-test scorers read faster. There was also a significant interaction between grammaticality and C-test (β=6.6, p=.018). To examine this interaction, C-test, which was a continuous variable in the main analysis, was split into two levels of 13 participants each around the median score, and pair-wise analyses were conducted. When C-test was low there was no significant effect of grammaticality (β= -.83.02, p>.1), while when the C-test was high ungrammatical sentences were read slower than grammatical sentences (β=173.19, p=.007).

Other regions showed no significant differences.

**Discussion**

As in previous research, sensitivity to violations of agreement was found, but only in learners of higher proficiency. This sensitivity was found in a construction that involved an intervening prepositional phrase, which has been suggested to be too complex a construction for L2 learners to calculate agreement (Keating, 2009; Wen et al., 2010). This suggests that the higher proficiency participants have gained enough grammatical knowledge to recognise violations of agreement, and can parse detailed enough structure to show sensitivity to this ungrammaticality on-line, while reading for content, contra models that predict persistent parsing or grammatical deficits.

**C-test comparison with TOEFL-ITP**

**Participants**

Twenty of the participants in the reading time experiment submitted their TOEFL-ITP scores, sat within a month of the reading experiment.

**The TOEFL-ITP**

The TOEFL-ITP (Test Of English as a Foreign Language – Institutional Testing Program) is a paper-based test for institutions. The version used took 115 minutes and has three sections: reading comprehension (55 mins), structure and written expression (25 mins), and listening comprehension (35 mins).

**Method**

C-test and TOEFL-ITP scores were compared using Pearson’s product-moment correlations, the TOELF-ITP both in total and as sections.

The values were also compared by constructing models with each proficiency measure as the independent factor and reading time as the dependent factor. The models were then compared using the anova function of the R stats package (R Core Team, 2015). The reading times included in these analyses were from the filler sentences only, as the test items contained ungrammaticalities.
Results

Twenty participants

Correlations. The correlations between the C-test and the TOEFL-ITP were high and highly significant; see Table 1 and Figure 1. Further, the C-test correlated strongly with all the sections of the TOEFL-ITP, over and above the correlations between sections, showing that the C-test reflected elements of proficiency from all these areas, beyond the element of proficiency they shared.

Table 1: Correlations between the C-test and TOEFL-ITP

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>TOEFL total</th>
<th>reading</th>
<th>structure</th>
<th>listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL total</td>
<td>0.842***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reading</td>
<td>0.716***</td>
<td>0.85***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>structure</td>
<td>0.6***</td>
<td>0.76***</td>
<td>0.46***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>listening</td>
<td>0.6***</td>
<td>0.76***</td>
<td>0.46***</td>
<td>0.2***</td>
<td></td>
</tr>
</tbody>
</table>

***p<.0001

Figure 1: Relationship between C-test and TOEFL total

Model comparisons. The model comparisons, summarised in Table 2, showed that the TOEFL total score was the best predictor of the reading times, followed by the reading section and the C-test.

Table 2: Model Comparisons: P-values indicate if the model using measure in the row was a significant improvement on the model using the measure in the column

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>TOEFL total</th>
<th>reading</th>
<th>structure</th>
<th>listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test</td>
<td>p=1</td>
<td></td>
<td>p=1</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>TOEFL total</td>
<td>p&lt;.001</td>
<td>p=1</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td>p=1</td>
</tr>
<tr>
<td>reading</td>
<td>p&lt;.001</td>
<td>p=1</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td>p=1</td>
</tr>
<tr>
<td>structure</td>
<td>p=1</td>
<td>p=1</td>
<td>p&lt;.001</td>
<td>p=1</td>
<td>p&lt;.001</td>
</tr>
<tr>
<td>listening</td>
<td>p=1</td>
<td>p=1</td>
<td>p=1</td>
<td>p&lt;.001</td>
<td>p=1</td>
</tr>
</tbody>
</table>

Nine participant group

Correlations. Inspection of Figure 1 showed there were nine participants with similar TOEFL-ITP scores (around 480) but different C-test scores. A correlation analysis including only these participants revealed that measures had weak or negative correlations, as summarised in Table 3.

Table 3: Correlations with 9 participants

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>TOEFL total</th>
<th>reading</th>
<th>structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL total</td>
<td>-0.07***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reading</td>
<td>0.13***</td>
<td>0.16***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>structure</td>
<td>-0.17***</td>
<td>-0.18***</td>
<td>-0.62***</td>
<td></td>
</tr>
<tr>
<td>listening</td>
<td>0.06***</td>
<td>0.41***</td>
<td>-0.11***</td>
<td>-0.67***</td>
</tr>
</tbody>
</table>

***p<.0001

Model Comparisons. None of the proficiency measures predicted this group’s reading times (ps>.1).
**Reading time analysis with the nine-participant group**

To look at the relationship between the inconsistency in the proficiency measures and sensitivity to ungrammaticality, the reading time analysis of the test items was conducted with only these nine participants.

There was an effect of grammaticality at the first noun ($\beta=-223.8$, $p=.009$). Similar to the analysis that included 26 participants, the plural noun was read more slowly than the singular. There was also an effect of C-test, as participants with higher C-test scores read faster ($\beta=-16.4$, $p=.009$).

In the region containing the verb, the critical region, the interaction between C-test and grammaticality was significant ($\beta=-6.55$, $p=.036$), showing the same numerical pattern as the 26 participants did in the spillover region. Higher proficiency participants were numerically slower in the ungrammatical condition (n=6, $\beta=92.4$) while lower proficiency participants showed the opposite trend (n=3, $\beta=-98.6$), though these contrasts were not significant in the pairwise analyses ($ps>.3$).

In the following region, there was a marginal effect of grammaticality ($\beta=84.8$, $p=.092$), such that ungrammatical sentences were read more slowly, and a main effect of C-test ($\beta=-8.73$, $p=.036$), as higher proficiency participants read more quickly. However, the interaction was not significant here ($\beta=1.1$, $p=.793$).

**Discussion**

The comparison of the TOEFL-ITP and the C-test showed that, overall, the TOEFL-ITP is better at predicting the learners’ reading times. The reading times were also better predicted by the reading section of the TOEFL-ITP. Correlations showed that there was a high level of similarity between the C-test and the TOEFL-ITP, replicating the high correlation found by Babaii and Shahri (2010) between this test and the TOEFL, apart from nine participants who scored in the mid-range in the TOEFL-ITP but had varying C-test scores. Analyses involving these participants showed that, while none of the proficiency measures could predict these participants’ reading times overall, the C-test was still correlated with the distinction between grammatical and ungrammatical conditions. While it is difficult to draw conclusions from such a small set of participants, it is worthwhile to note that identical analyses using the TOEFL revealed no significant effects of grammaticality in this group.

**General Discussion**

The current paper has emphasised two points. First, that number agreement can be acquired by Japanese learners of English, as suggested by the high proficiency participants who did not show persistent deficiencies in grammatical knowledge or parsing in this area. Second, that the C-test is a reasonable test of L2 proficiency for the purposes of psycholinguistic research.

The display of sensitivity to ungrammatical agreement requires both knowledge of morphology and the ability to apply that knowledge to the structure, and thus the current results argue against persistent deficit models (Hawkins & Yuet-Hung, 1997; Clahsen & Felser, 2006).

The sensitivity was found to be modulated by proficiency as measured by a C-test. These results, therefore, are compatible with models such as the Full Transfer / Full Access model (Schwartz & Sprouse, 1996) which suggest that the difference between L1 and L2 speakers is the level of grammatical knowledge, which can be gained by L2 speakers over time.

The current study used a C-test to measure proficiency, and this was found to be a useful measure in that is was sufficient to distinguish between those who were and were not sensitive to violations of number agreement. The TOEFL-ITP predicted reading times better than the C-test. However, the C-test showed a high level of correlation with the TOEFL-ITP and its subsections. This means it had a large amount of overlap with proficiency that is being measured in the more varied TOEFL-ITP.

Moreover, for the group of nine participants who had similar TOEFL-ITP scores but varying C-test scores, the C-test still reflected reading time differences. This distinction between participants reflected in the C-test scores but not reflected the TOEFL-ITP scores was a predictor of
which readers were more likely to show sensitivity to ungrammatical agreement, as well as which readers read faster overall, though the sample size was very small.

Finally, the C-test is easier to administer than the TOEFL-ITP. The TOEFL-ITP takes 115 minutes to complete (or 55 minutes for the reading section). In contrast, the C-test used in the current study required only 15 minutes, allowing it to be administered in the same session as the experiment (see Wen et al., 2010, for concerns about the applicability of old test scores).

Conclusion
Japanese-speaking learners of English are able to show sensitivity to violations of English number agreement, though it is not present in their L1. This sensitivity was displayed by higher-proficiency learners, and the C-test is a useful measure of proficiency for distinguishing such learners.

References
Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-8