**Online appendices for “Learning second language morphosyntax in dialogue under explicit and implicit conditions: An experimental study with advanced adult learners of German”**

**Overview:**

* **Appendix S1 – Task instruction in the explicit condition:** English translation (from Dutch) of the extra instruction page that informed the participants in the explicit condition about the true purpose of the study.
* **Appendix S2 – The unaware participant group:** A description of the unaware subgroup (*n* = 6) of the implicit condition, containing information about the participants’ language background, descriptive statistics of their performance data, as well as an interpretation thereof.
* **Appendix S3 – The 15-minute delayed posttest:** A description of the brief delayed posttest that was administered 15 minutes after the main experimental task, including an explanation of why this task was discarded from the study, a presentation of descriptive statistics, as well as a (speculative) discussion of the results.
* **Appendix S4 – Model comparisons:** A detailed description of the model comparisons that were used to determine the structure of the final mixed-effects model, modelling the learning task data of the critical items.
* **Appendix S5 – Goodness of model fit:** An evaluation of the goodness of model fit of the critical items mixed-effects model.
* **Appendix S6 – Interpretation of model estimates:** A detailed explanation of how to interpret the estimates (logits, probabilities, odds) provided in the model output, and of how to understand the odds ratios as an indicator of effect size.

**Appendix S1. Task instruction in the explicit condition**

English translation (from Dutch) of the extra instruction page that informed the participants in the explicit condition about the true purpose of the study:

“The verbs play a central role in this experiment. In German, there are several verbs that change their stem vowel in the third person of the present tense (he/she/it). A verb like *graben* (“to dig”) becomes *er gräbt* (“he digs”); *gelten* (“to be valid”) becomes *gilt*… In general, this aspect of the German language is rather difficult to acquire for native speakers of Dutch.

It now is your task to try to conjugate the verbs accurately and to apply the vowel change whenever it is required. Furthermore, you will get the possibility to learn the correct forms from the verbs produced by the experimenter because she’s supposed to form sentences using the same verbs. So listen attentively.”

**Appendix S2. The unaware participant group**

Six participants of the implicit condition neither had awareness of the target structure [-TS], nor of the task’s learning purpose [-LP] during the learning task (i.e., the dialogue game). We therefore called these participants ‘unaware’. Due to the small sample size, this subgroup was excluded from all statistical analyses. Table S2.1 provides a description of the unaware participants’ average learner profile. Three participants in our sample reported no prior experience with the stem-vowel change; they were all part of the unaware subgroup.

**Table S2.1**

*Descriptive statistics of variables related to the unaware participants’ language background in L2 German (n = 6)*

|  |  |  |
| --- | --- | --- |
| Variable | Median | IQR |
| Age | 19.00 | 6.50 |
| Years of instruction (school) | 3.00 | 0.00 |
| Years of instruction (university) | 0.00 | 0.06 |
| Number of other L2s | 2.50 | 1.00 |
| Frequency of usage\* | 1.00 | 0.00 |
| Pre-measure critical items | 0.00 | 0.00 |
| Pre-measure control items | 100.00 | 0.00 |
|  | Mean | *SD* |
| Vocabulary size(LexTALE) | 66.46 | 0.87 |
| Proficiency\* | 2.54 | 0.87 |

*Note*. IQR = interquartile range. Asterisks mark variables based on self-ratings on a 1-5 Likert scale (1 = *very low*, 5 = *very high*). Pre-measures and LexTALE scores are percentages.

***Description of the results***

We present the data of the unaware participants by means of descriptive statistics (Table S2.2). In line with the research questions, we aimed to detect any trends that might reflect learning in this group.

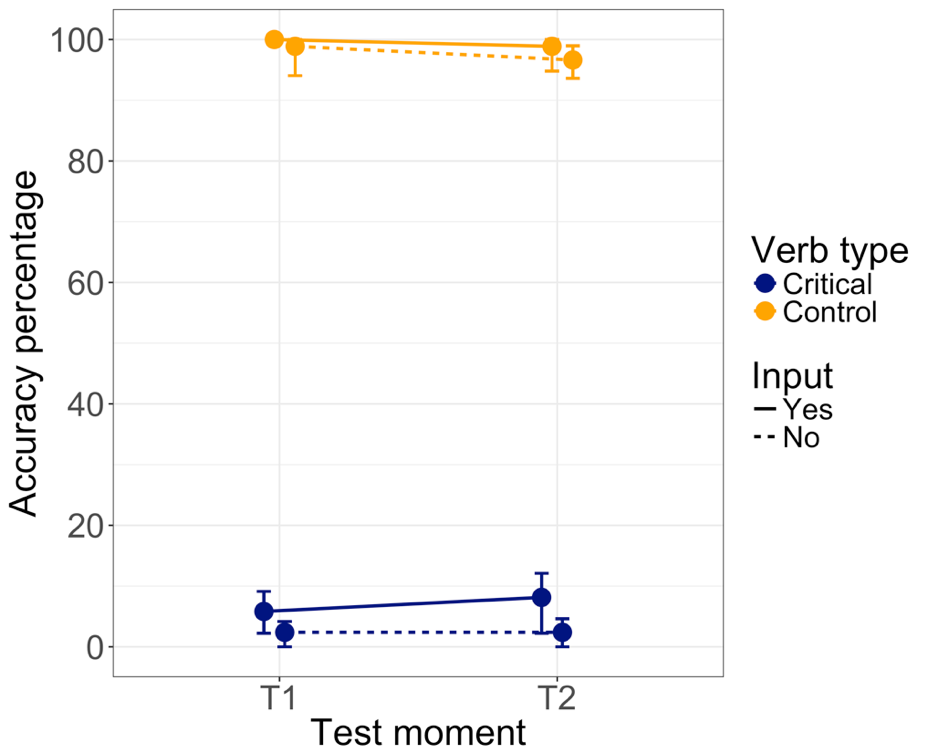
Figure S2.1 visualizes the test scores of the unaware subgroup. These participants obtained extremely low scores on critical items throughout the task (*M* = 4.46, *SD* = 3.93, averaged over Input and Test moment). Neither input nor no-input items changed from T1 to T2, suggesting there was no learning. The opposite picture emerges for the control items: throughout T1 and T2, accuracy scores were very close to ceiling (*M* = 98.61, *SD* = 1.67, averaged over Input and Test moment). Thus, as a default, the unaware group appears to have applied hardly any stem-vowel changes to the verbs, and the provision of input seems not to have altered this behavior.

**Table S2.2**

*Descriptive statistics of the percentage of correctly produced stem vowels in the unaware group*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Critical items | | | |  | Control items | | | |
|  | T1 | | T2 | |  | T1 | | T2 | |
|  | Input | No input | Input | No input |  | Input | No input | Input | No input |
| Mean | 5.71 | 2.08 | 7.80 | 2.23 |  | 100.00 | 98.81 | 98.96 | 96.66 |
| *LL* | 2.22 | 0.00 | 2.22 | 0.00 |  | 100.00 | 94.05 | 94.79 | 93.61 |
| *UL* | 9.13 | 4.17 | 12.10 | 4.61 |  | 100.00 | 100.00 | 100.00 | 98.96 |
| *SD* | 5.07 | 5.10 | 6.60 | 3.47 |  | 0.00 | 2.91 | 2.55 | 3.67 |
| *Mdn* | 6.91 | 0.00 | 9.59 | 0.00 |  | 100.00 | 100.00 | 100.00 | 96.875 |
| Min. | 0.00 | 0.00 | 0.00 | 0.00 |  | 100.00 | 92.86 | 93.75 | 92.86 |
| Max.m | 13.33 | 12.5 | 14.29 | 7.14 |  | 100.00 | 100.00 | 100.00 | 100.00 |

*Note*. *LL* = lower limit of the 95% confidence interval (10,000 samples BCa bootstrapping); *UL* = upper limit. T1 = Test moment 1; T2 = Test moment 2. *Mdn* = median; min. = minimum value; max. = maximum value.



*Figure S2.1*. Mean test scores on the different test moments for the unaware group (*n* = 6). Error bars represent 95% BCa confidence intervals.

***Discussion: No learning without awareness***

The six participants of the unaware group reported not having noticed the vowel-changing strong verbs at all, and their data did not reveal any trends suggesting learning. Throughout the learning task, these participants almost never applied stem vowel changes to critical items; their average accuracy percentage was only 4.46%. Thus, the beneficial role of output production for learning did not hold for all participants. These findings point towards an important role of awareness in our study: only those participants who were or became aware of the target structure also showed learning (see Andringa, 2020). Notwithstanding, implicit learning in the form of an increased sensitivity to the target forms might have been present, but without being strong enough to show in the oral production data.

It should be noted that our speculations concerning the unaware group are based on non-significant trends in the data, implicating that no valid claims can be made. Moreover, the only three participants in our study with no prior knowledge of the target structure were all unaware participants. Although this may suggest that a lack of prior knowledge decreases the likelihood of noticing a morphosyntactic feature as subtle and non-salient as the stem-vowel change in strong verbs, we cannot know for certain whether it was the lack of awareness or the lack of prior knowledge that explains the absence of learning.

***References***

Andringa, S. (2020). The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research*, *36*, 335–357. https://doi.org/10.1177/0267658320915502

**Appendix S3. The 15-minute delayed posttest**

We administered an unannounced delayed posttest at the end of the experimental session. The test took place approximately 15 minutes after the end of the learning task and had a duration of three minutes. All critical and control verbs from the learning task were presented in their written infinitive forms on the screen, one by one and in a random order. The participants had to orally produce the verbs, in isolation, in the third person of the singular in the present indicative tense.

Originally, this test was included to measure the short-term-retention of what had been learned during the learning task: how much of what had been learned during the learning task would still be recalled after a time lapse of 15 minutes, and would this retention rate be comparable in both groups? However, during the early stages of testing it already became clear that the test outcomes could not be used as a pure measure of the effect of time on the learning outcomes, and that the test’s internal validity could thus not be guaranteed.

The first reason is that the difference in test format between the learning task and the delayed posttest was likely to affect the results, both in the explicit and implicit conditions. This issue is related to the fact that different types of tasks may elicit the use of different types of knowledge, and therefore may lead to differences in performance on task (e.g., Andringa, de Glopper, & Hacquebord, 2011). The learning task had a rather implicit task format because of its meaning-based, dialogic nature, and required the participants to produce the verbs embedded in full sentences. This meant that in addition to the strong verbs, the learners also had to pay attention to sentence meaning and case marking of the noun phrases. The learning task was thus more cognitively demanding than the delayed posttest, which simply required the participants to produce the verbs in isolation and was thus likely to trigger the use of a much more explicit kind of knowledge.

The second reason is that we administered the delayed posttest at a moment where all participants of the implicit condition had become aware of the actual purpose of the study, due to the information they had received during the awareness interview. This is likely to have considerably altered their test performance: while their productions during the learning task were rather spontaneous, the productions at the delayed posttest involved controlled attention and a maximal reliance on conscious knowledge.

Unfortunately, it was methodologically impossible to question the participants about their beliefs and thoughts regarding the learning task without raising their awareness of the study’s true purpose. Conducting the awareness interview after the delayed posttest was not considered as a suitable alternative solution for two reasons. Firstly, as the posttest explicitly and exclusively focused on strong verb conjugation, this would have undoubtedly raised the participants’ awareness of the target structure, thereby biasing the outcomes of the awareness interview. Secondly, conducting the interviews immediately after the learning task was indispensable, as delaying it might have caused the participants to start forgetting about how they experienced the learning task and becoming unable to recall their experiences and thoughts.

Despite the delayed posttest’s suboptimal validity to measure short-term retention, we believe that the test results may bring some complementary insights at the level of the type of knowledge that the two groups had been relying on during the learning task. However, the analysis was post-hoc, and the results and our interpretation thereof need to be seen as speculative.

***Post-hoc research question***

How do the participants in our study perform on strong verb conjugation at the delayed posttest fifteen minutes after the learning task? Does the outcome reflect a) forgetting, as an effect of time; and/or b) an effect of the difference in task formats between learning task and the delayed posttest; and/or c) an effect of awareness in the incidental group, as caused by the awareness interview?

***Analysis***

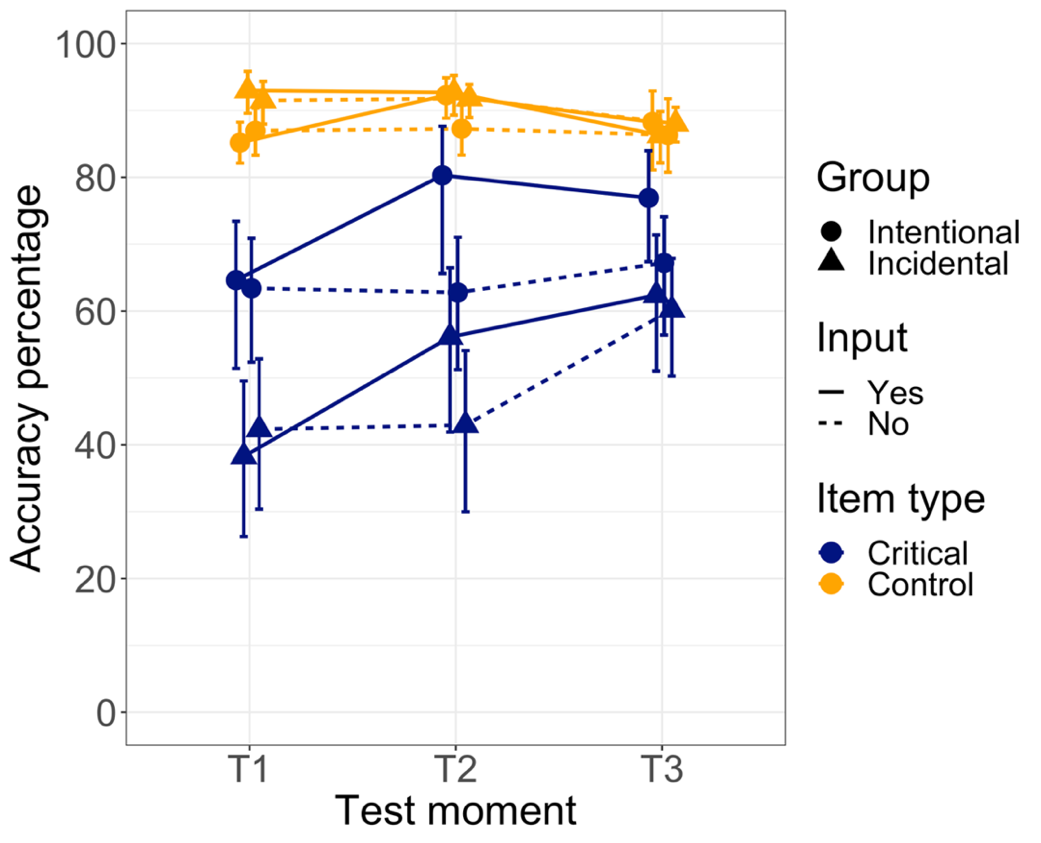
The scoring of the delayed posttest happened identically to the scoring of the learning task. Immediately after hearing each participant production, the experimenter determined the accuracy of the produced verb form. This was later recoded in a binary way, with 0 for vowel errors and 1 for correctly produced stem vowels.

The outcomes of the delayed posttest will be presented by means of descriptive statistics. To find out in what way performance during the learning task was related to performance fifteen minutes later at the posttest, we planned to inspect the interaction between Test moment (T2 of the learning task, delayed posttest) and Input (input, no input) for critical items. We were also interested in a modulation of this interaction by Group (intentional, incidental).

The different effects that we might observe were operationalized as follows. An effect of time – that is, forgetting what just has been learned as time goes by – should become visible as a decrease for input items while the scores on no-input items remain steady. This holds for both participant groups. An effect of the difference in task formats, however, could become visible as an increase for no-input items in the intentional group. If present, this task effect could then be assumed to affect the input items in the same way, and to also affect the incidental group. If we observed a decrease for input and an increase for no-input items in the intentional group, this may point towards a combined effect of time and task. The third type of effect that might occur is an effect of the awareness that only the incidental learners experienced during the awareness interview between T2 and the posttest. We operationalized this as an increase for no-input items in the incidental group, minus the increase for no-input items in the explicit condition (i.e., the task effect). If present, this effect could be assumed to also affect the input items to some extent.

***Results***

The descriptive statistics of the delayed posttest are provided in Table S3.1 and graphically represented in Figure S3.1 (intentional and incidental groups) and Figure S3.2 (unaware group). Although the learning task and the delayed posttest (T3 in the figures) share the same test items and are presented within the same graphs, we remind the reader that they represent different experimental tasks with very different formats. Presenting both tasks in the same graphs was simply done to make a direct visual comparison of the test scores easier.



*Figure S3.1*.Mean test scores on the different test moments of the learning task (T1, T2) and on the delayed posttest (T3) for the intentional and incidental groups, and for critical and control items. Error bars represent 95% BCa confidence intervals.

**Table S3.1**

*Descriptive statistics of the percentage of correctly produced stem vowels at the delayed posttest*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Critical items | | Control items | |
|  | Input | No input | Input | No input |
| Explicit group (*n* = 21) |  |  |  |  |
| Mean | 76.15 | 66.03 | 88.40 | 86.31 |
| 95% BCa CI (*LL*) | 67.38 | 56.40 | 84.81 | 80.58 |
| 95% BCa CI (*UL*) | 83.08 | 74.09 | 91.80 | 90.33 |
| *SD* | 18.66 | 21.12 | 8.40 | 11.45 |
| Median | 81.25 | 66.67 | 87.50 | 87.50 |
| Minimum score | 36.36 | 20.00 | 73.33 | 56.25 |
| Maximum score | 100.00 | 100.00 | 100.00 | 100.00 |
| Incidental group (*n* = 21) |  |  |  |  |
| Mean | 61.44 | 59.28 | 86.27 | 87.91 |
| 95% BCa CI (*LL*) | 51.05 | 50.26 | 82.17 | 85.28 |
| 95% BCa CI (*UL*) | 71.38 | 67.86 | 89.84 | 90.48 |
| *SD* | 24.31 | 21.52 | 9.12 | 6.29 |
| Median | 60.00 | 50.00 | 87.50 | 87.50 |
| Minimum score | 20.00 | 7.14 | 68.75 | 76.92 |
| Maximum score | 93.75 | 100.00 | 100.00 | 100.00 |
| Unaware group (*n* = 6) |  |  |  |  |
| Mean | 26.37 | 21.30 | 89.58 | 88.88 |
| 95% BCa CI (*LL*) | 10.97 | 5.56 | 63.54 | 71.10 |
| 95% BCa CI (*UL*) | 45.72 | 46.95 | 96.88 | 95.55 |
| *SD* | 24.22 | 26.84 | 17.97 | 14.40 |
| Median | 17.38 | 11.67 | 100 | 93.33 |
| Minimum score | 0.00 | 0.00 | 56.25 | 60.00 |
| Maximum score | 64.29 | 68.75 | 100.00 | 100.00 |

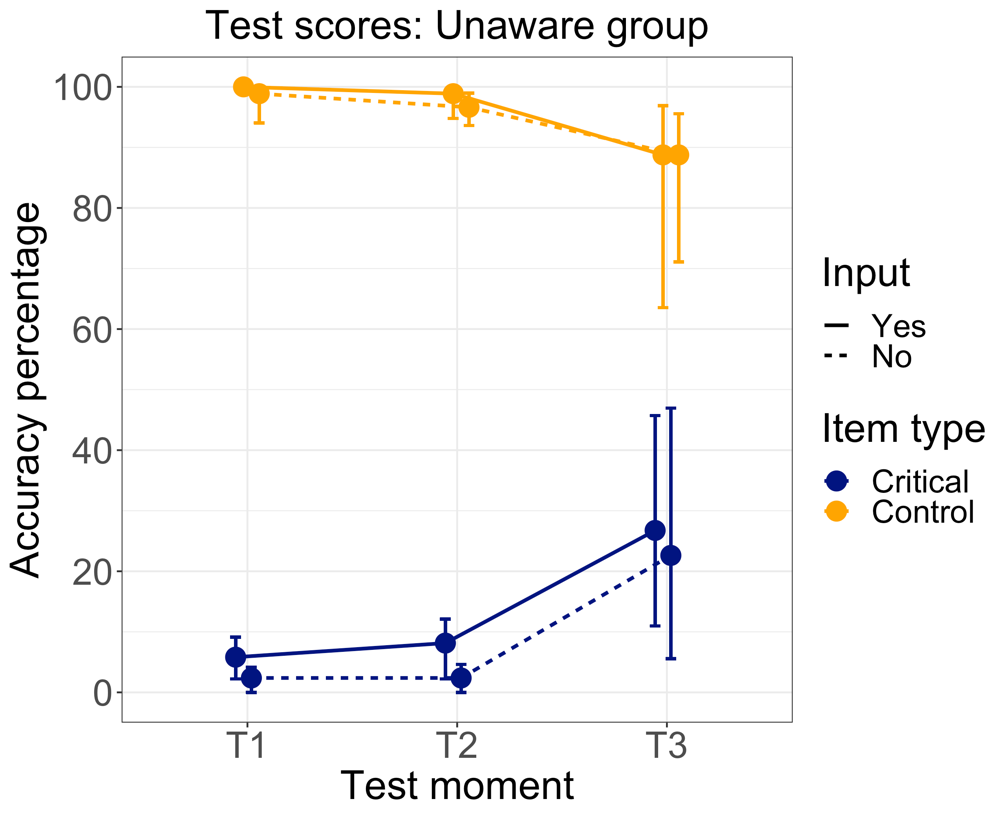
*Note.* 95% BCa CI = 95% confidence intervals with a 10,000 samples BCa bootstrap. *LL* = lower limit; *UL* = upper limit.

The following trends could be observed: For critical items, there appeared to be an interaction of Test moment and Input, which looks different in the intentional and incidental groups. The intentional learners appeared to have obtained slightly lower scores at the posttest as compared to T2 for input items, possibly reflecting a weak effect of time (that is, they were forgetting what they just had learned). The same group had slightly higher scores for no-input items at the delayed posttest as compared to T2, which might be due to the difference in task formats.

In the incidental group, we can observe an increase from T2 to the posttest both for input and no-input critical items. The increase for no-input items was strongest, leading to similar test scores for input and no-input items at the posttest. These outcomes point towards an effect of the awareness that this group experienced during the interview.

At the delayed posttest, the overall test scores of the incidental group did not seem to differ from the intentional group’s performance on no-input items. This could be interpreted as a confirmation that the two groups did not differ with respect to their pre-experimental explicit knowledge of the vowel change. Only the scores for input items remained higher in the intentional group – as compared to the incidental group, and as compared to no-input items in the intentional group. This might suggest that the exposure to the target forms during the learning task still had a beneficial effect on performance at the delayed posttest, but only in the explicit condition.

The results of the unaware group are visualized in Figure S3.2. From T2 to the posttest, there appeared to be a rather strong increase for both input and no-input critical items, probably reflecting an effect of the different task formats and of the newly gained awareness about the target structure. However, the graphs representing performance on control items look almost mirrored to those of the critical items, suggesting that awareness of the target structure caused the unaware participants to overgeneralize the vowel change to weak verbs. Moreover, awareness of the target structure appeared to have entailed considerable variance at the level of performance.



*Figure S3.2*. Mean test scores on the different test moments of the learning task (T1, T2) and on the delayed posttest (T3) for the unaware group. Error bars represent 95% BCa confidence intervals.

***Discussion***

The results suggest that the delayed posttest indeed could not be used to measure knowledge development as a pure function of time. Rather, the findings point towards a combined effect of time, of the difference in task formats, and of the awareness experienced by the incidental group.

The interpretation of the results may become more straightforward if we only consider the performance on the test items for which no input was provided, as this rules out any effects of time (i.e., there was no learning on these items, so there could not be any forgetting either). By doing so, the posttest could be read as a separate explicit knowledge test, and no longer as a ‘posttest’. The picture that emerges then is the following: There does not seem to be a difference between the groups, which confirms the absence of group differences other than caused by our experimental manipulation. Both groups performed better at the delayed posttest than during the learning task, suggesting that they could benefit from being able to direct their full attention to the conjugation of strong verbs and fully access their explicit knowledge. This task-related advantage was less pronounced in the intentional group, possibly because these participants could already rely on a more conscious, explicit type of knowledge during the learning task – which was also why overall, they outperformed the incidental group during this task. For the incidental group, the advantage was more pronounced, possibly because these participants had been relying on a more implicit, spontaneous kind of knowledge during the learning task.

***References***

Andringa, S., de Glopper, K., & Hacquebord, H. (2011). Effect of explicit and implicit instruction on free written response task performance. *Language Learning*, *61*, 868–903. https://doi.org/10.1111/j.1467-9922.2010.00623.x

**Appendix S4. Model comparisons**

We modelled the binary accuracy scores of the productions of critical items on T1 and T2, starting with the following model, which we call ‘start model’: (Stem-vowel accuracy) ~ 1 + Input\*Test moment\*Group + (1|Item) + (1|Participant). The first term of the notation represents the dependent variable. The terms at the right of the tilde character are the model terms. The asterisks mark a three-way interaction. The random-effects terms are those which include the bar symbol (|); 1 represents an intercept.

To begin with, we investigated whether random intercepts over items and participants significantly increased model fit by, in turn, removing one of them and checking whether this decreased model fit. We first compared the start model to that same model but without random intercepts over items. The ANOVA showed that the inclusion of random intercepts over items significantly increased model fit (χ2 = 240.35, *df* = 1, *p* < .001). Similarly, we compared the start model to that same model without random intercepts over participants. Again, the start model fit the data significantly better (χ2 = 626.13, *df* = 1, *p* < .001).

Following Bates, Kliegl, Vasishth and Baayen (2015), we explored how further random effects might increase model fit. To this end, we explored adding our variables in the model’s random structure one by one, first as main effects, later as interaction effects. We started by including random slopes of Group over items, but this did not significantly increase model fit (χ2 = 1.94, *df* = 2, *p* = .38). We thus again removed this random effect from the model. We then added random slopes of Test moment over participants, but again this did not lead to an improvement (χ2 = 0.95, *df* = 2, *p* = .62), so we discarded the effect from the model. We added random slopes of Test moment over items, but no model improvement was achieved (χ2 = 2.78, *df* = 2, *p* = .25) and we removed the effect again. Next, we added random slopes of Input over participants but this did not improve model fit (χ2 = 4.42, *df* = 2, *p* = .11), and the effect was removed from the model. Adding random slopes of Input over items, however, significantly increased model fit (χ2 = 7.00, *df* = 2, *p* = .03). It should be noted that the decrease of AIC was minimal, from 2559.1 to 2556.1. Following the principal component analysis presented by Bates and colleagues (2015), we found that the model could support the current number of parameters (this could be seen from the observation that each parameter explained some of the variance and no parameter was at zero), so we decided to keep the random slopes of Input over items in the model.

We then started to add the interaction effects. First, we added the random slopes of Input x Test moment over participants; however, this model failed to converge and we removed the effect again. Adding random slopes of Input x Test moment over items led to a small but significant improvement (χ2 = 15.06, *df* = 7, *p* = .04), but the principal components analysis showed that the model now contained parameters that could not be supported by the data, so we removed the effect. We added random slopes of Group x Test moment over items, but the model failed to converge and we discarded the effect. Adding random slopes of Group x Input over items did not significantly improve the model (χ2 = 11.92, *df* = 7, *p* = .10), nor did adding random slopes for the three-way interaction Input x Group x Test moment over items (χ2 = 31.97, *df* = 33, *p* = .52), so none of these effects were kept in the model.

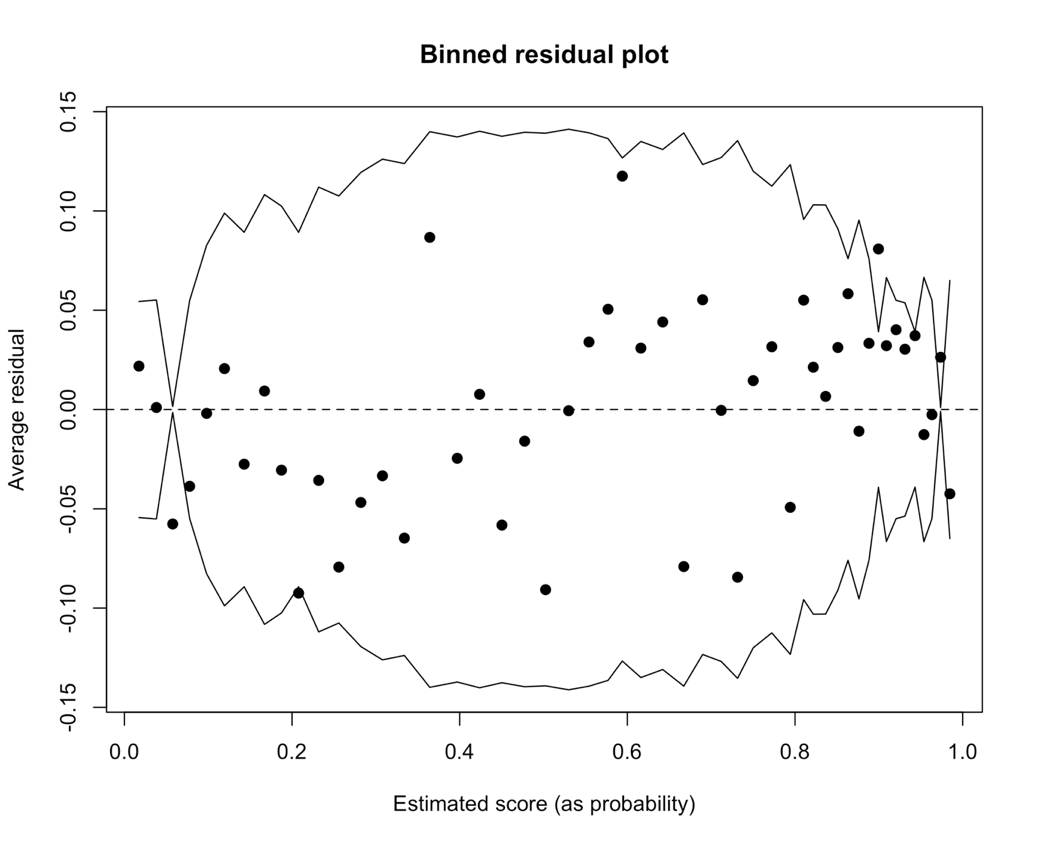
Thus, the best model we could identify took the following form: (Stem-vowel accuracy) ~ 1 + Input\*Test moment\*Group + (1 + Input|Item) + (1|Participant). This model only differs from the start model by its inclusion of the random slopes of Input over items. All inferential statistics were computed with this final model. Marginal *R*2 was 0.11; conditional *R*2 was 0.55.

***References***

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint, arXiv:1506.04967.

**Appendix S5. Goodness of model fit**

In order to evaluate model fit, we created a binned residual plot (Figure S5.1) by using the *arm* package in R (version 1.10-1, Gelman & Su, 2018). Instead of plotting raw residuals against fitted values – which would not be very informative, since we deal with binary (0 or 1) data – this plot splits up the data into distinct groups (or ‘bins’) based on the fitted values, and then plots, per bin, the corresponding average residual (y-axis) against the average fitted value (x-axis) (Gelman & Hill, 2007, p. 97). The fitted values are the scores that the model predicts, here expressed as probabilities. The residuals are the model error; they represent the difference between the observed and the predicted values. There are 51 bins (the square root of number of observations, 2560), which are the dots on the plot. The zigzagged lines around the bins represent ±2 standard-error bounds. If the model fits the real data well, we would expect 95% of the bins to fall between these bounds.



*Figure S5.1*. Binned residual plot for the logistic mixed-effects model.

Figure S5.1 shows that five bins fall outside of the bounds, which corresponds to 9.8%. However, most of them fall on or very close to the line. Further, the bins look rather evenly distributed, pointing towards homoscedastic data. This means that the variance of residuals does not seem to depend on the fitted values, a good sign. Yet, there appears to be a trend of low fitted values having negative residuals and high fitted values having positive residuals. In other words, our model has a tendency of predicting low values slightly lower than the true values, and high scores slightly higher than the true values. However, this trend does not seem to be very strong, so we expect it to not interfere much with the conclusions we have drawn from our model.

***References***

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gelman, A. & Su, Y. (2018). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models* (R package version 1.10-1.) [Computer software]. https://CRAN.R-project.org/package=arm.

**Appendix S6. Interpretation of model estimates**

The model’s parameter estimates presented in Table 5 (see main text) are on the logit scale. The logit values are the logarithm of the odds and can also be transformed again into probabilities by means of the formula ex / (1+ex), where x is the logit and e a mathematical constant, approximately 2.72. Thus, although they use different scales and ranges, logits, odds and probabilities all express the same thing from a different perspective – which is in this case the probability of the strong verbs’ stem vowels being produced either correctly or incorrectly, given certain circumstances (i.e. specific factor-level combinations). For a more detailed explanation, see Gries (2013, p. 299). We decided to also report the estimated probabilities, since they seemed most intuitive and easy to interpret to us.

The intercept of the model represents a specific combination of factor levels. In Table 5, it represents the intentional group, tested at T1, on critical items for which they would not receive any input. The logit is 0.69 and corresponds to a probability of 0.67, which means that our model predicts a 0.67 probability for intentional learners at T1 to produce a correct stem vowel for no-input items. Or, phrased in terms of percentages, our model predicts an average accuracy of 67% for intentional learners at T1.

All simple effects and interaction effects need to be interpreted against the intercept. We report these effects along with the factor level that is to be compared with the intercept level of that factor. For instance, the line ‘Input = yes’ refers to the effect of Input and tells us that at T1 in the intentional group, items with input were not significantly different (*p* = .624) from items without input, the latter being the reference level. To obtain the logit for this effect, we calculate 0.69 (the intercept logit) + 0.10 (the Input = yes logit) = 0.79, which corresponds to a 0.69 probability; i.e., 2.720.79/ (1+2.720.79) = 0.69. To obtain the estimated logit of an interaction, we have to sum up the estimates of the interaction, of the simple effects included in the interaction, and of the intercept.

We used odds ratios (ORs) as an effect size. They tell us the change in the odds of producing a correct stem vowel, associated with the change from the reference factor level to another factor level. Based on a brief literature search, de Vos, Schriefers, Nivard and Lemhöfer (2018) suggested that ORs around 1.5 may be interpreted as small, between 2.5–3.5 as medium, and above 4–9 as large. For ORs under 1, a conversion of 1 divided by the OR must be applied before interpreting its size. For instance, the estimated OR for a change from Group = intentional to Group = incidental was 0.27, meaning that the odds to produce a correct stem vowel are 0.27 times higher in the incidental group than in the intentional group (at T1, for no-input items). However, this can be turned around by saying that the odds to produce a correct vowel are 3.70 (i.e., 1/0.27) times higher in the intentional group than in the incidental group, which would be considered a medium-to-large sized effect.

***References***

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction*. Berlin/Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110307474

de Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta‐analysis and meta‐regression of incidental second language word learning from spoken input. *Language Learning, 68,* 906–941. https://doi.org/10.1111/lang.12296