Modeling systematicity and individuality in nonlinear second language development

Murakami, Akira

DOI:
10.1111/lang.12166

Document Version
Peer reviewed version

Citation for published version (Harvard):
Murakami, A 2016, 'Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes', *Language Learning*, vol. 66, no. 4, pp. 834-871. https://doi.org/10.1111/lang.12166

Supporting Information for: Murakami, A. Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. Article accepted in *Language Learning* on 4 September 2015.

Appendix S1: General Issues in Regression Modeling

Some General Issues in Regression Modeling

The modeling techniques introduced in the paper are all variants of regression models. In this document, some general issues in regression modeling are briefly discussed. They are all relevant to the models discussed in the paper, and the present document lays the foundation for more complex models introduced in the main text of the paper.

Regression Modeling with Continuous Independent Variables

A basic form of regression models estimates the relationship between one dependent variable and one or more independent variables (or predictors). Independent variables can be continuous or categorical. Figure 1A shows a simple linear regression in which a dependent variable (L2 proficiency) is modeled by a continuous independent variable (number of years learning L2). Each observation, denoted by a circle, represents one learner, and the same follows for the rest of the panels. The regression line is drawn to minimize squared *residuals*. Residuals are squared differences between the observed (i.e., circles) and the predicted (the regression line) values and are indicated by dashed lines in the figure.

This regression model has two *parameters*, or values to be estimated from the data; the intercept and the slope of the regression line. The intercept (0.950 in this case) denotes the value of the dependent variable when the value of the independent variable is zero. In the present case, it shows the value of L2 proficiency when the number of years learning L2 is zero. The slope of the regression line (0.731 in this case) represents the size of the predicted change of the

dependent variable per unit change of the independent variable. In the present context, L2 proficiency increases by 0.731 units in a year of learning the language.

We build regression models that explain maximum variance. Explaining variance means decreasing the squared residuals. Figure 1B shows the predicted values of the regression model that only has one parameter, the intercept. The model has a poor fit because the residuals are large, particularly at the beginning and the end of the horizontal axis. This so-called intercept-only model constitutes the model against which explained variance is calculated in the other regression models fitted to the same data. A common measure of explained variance, R^2, is defined as follows: $1 - \frac{\text{sum of squared residuals in the target model}}{\text{sum of squared residuals in the intercept-only model}}$. What matters is the ratio between the sum of the squared residuals in the target regression model (e.g., the regression model in Figure 1A) and that in the intercept-only model. If the former is close to zero (i.e., if there is little difference between observed and predicted values) and the variance of the dependent variable is large (i.e., the intercept-only model fits poorly), the value will be close to 1. If, however, the target model fits poorly and the residuals are nearly as large as those in the intercept-only model, the R^2 will be small. It is, therefore, the difference between the target regression model and the intercept-only model that is important in evaluating the fit.

There can be more than one predictor in regression models. In Figure 1C, the test score is modeled by two continuous independent variables; L2 experience (operationalized as the number of years learning L2) and the results of an aptitude test. Note that this time we have a regression surface (or regression plane) rather than a regression line because we used two independent variables. Similar to the regression line, the surface is drawn such that the squared difference between the observed values (i.e., small spheres) and the predicted values (i.e., the surface) is minimized. Here, the surface goes up as the value of L2 experience increases, thus suggesting

that increased L2 experience leads to increased test scores. This model has three parameters; the intercept and two slopes, one for each independent variable. The intercept is the predicted value when both L2 experience and aptitude are zeros, while the slopes, as before, represent the change of the predicted value per unit change of the independent variables.

Regression Modeling with Categorical Independent Variables

In regression modeling, categorical independent variables are expressed with dummy variables. If a factor has two levels (e.g., male and female), we need one dummy variable whose values are zero for one level (e.g., male) and one for the other (e.g., female). In Figure 1D, the test score is modeled by a factor called group with two levels, control and experimental. Herein, the control group was assigned zeros, and is called the *reference level* or *baseline level* because all the other levels (in this example, the experimental group) are compared against it. Because the independent variable is categorical, it can only take two values and there is no data point between the two factor levels. The regression line is drawn in the same way as in Figure 1B. The intercept of the line (3.41) is the test score of the reference-level group (i.e., control group, whose value of the group variable is zero), and the slope (or *contrast*) corresponds to the difference in the mean test score between the control and the experimental groups because they are a unit apart in the dummy variable. Note that what is called treatment contrast is assumed in the above coding of the dummy variable, and using other contrasts (e.g., zero-sum contrast that assigns -1 and 1 rather than 0 and 1 to two factor levels) would require a slightly different interpretation. When independent variables only include categorical variables, as in the example just given, the analysis is essentially the same as a t-test or ANOVA. In other words, a t-test or ANOVA is just a special case of a regression model.

Figure 1E illustrates a case where there is one independent variable of a factor with three levels (L1 Japanese, L1 Russian, and L1 Spanish). The factor is represented by two dummy variables, one denoting whether the value of the factor of an observation is L1 Russian and the other denoting whether it is L1 Spanish. The reference-level group, L1 Japanese, is assigned zeros for both variables. Following the notation of (L1 Russian, L1 Spanish), the values of the two dummy variables are (0, 0) for L1 Japanese, (1, 0) for L1 Russian, and (0, 1) for L1 Spanish. This is quite similar to having two independent variables. More generally, if we have a factor with *k* levels, the factor is expressed by *k - 1* dummy variables. The resulting figure is naturally similar to the previous figure with two independent variables. The intercept represents the mean value of the dependent variable by the reference-level group (i.e., L1 Japanese). The slope of L1 Spanish is equal to the difference in the mean test score between the reference-level group and L1 Spanish, while that of L1 Russian indicates the difference between the reference-level group and L1 Russian.

Interaction, Centering, and Standardization

Regression models are much more flexible and expressive when they include *interactions*. The presence of an interaction indicates that the effect of one variable depends on the value of another. Figure 1F illustrates this point. Herein, the test score is regressed against L2 experience (operationalized as the number of years learning L2) and a factor L1 that has three levels - L1 Japanese as the reference level, L1 Russian, and L1 Spanish. A factor with three levels is expressed by two dummy variables, and together with the continuous variable of L2 experience, the model produces a three-dimensional regression plane (i.e., one dimension higher than the two-dimensional regression surface in Panel C and E in Figure 1), which cannot be readily drawn as a figure. The figure, therefore, draws separate regression lines for the three levels of the factor.

In the figure, an interaction between L1 and L2 experience is evidenced in that the effect of the latter (i.e., the slope of the number of years learning L2) depends on the values (i.e., levels) of L1. Including the interaction terms, there are six parameters in the model; an intercept, two coefficients for the dummy variables representing L1 Russian and L1 Spanish, one coefficient for L2 experience, one coefficient for the interaction between L1 Russian and L2 experience, and one for the interaction between L1 Spanish and L2 experience. The coefficients of the interaction terms represent the adjustments to the main effects. In the present case, the main effect of L2 experience is 0.329, which shows the amount of change per year of learning L2 for L1 Japanese learners (the reference group). The interaction coefficient between L1 Russian and L2 experience is 0.739. This means that the slope of L1 Russian is calculated by adding this value to the main effect (0.329), thus resulting in 1.068.

An oft-used technique in regression modeling is *centering* and *standardization*. In the model just discussed, the main effect of L1 examines whether there are differences in the test score between L1 Japanese and the other two L1 groups. However, in the current form of the model, it does so where L2 experience = 0. In other words, and more generally, the main effect indicates the magnitude of the effect when the other coefficients are zeros. This is not a problem if the model does not include interaction terms, as the differences between L1 groups do not vary across L2 experience. However, with the interaction between L1 and L2 experience, it may be more reasonable to compare L1 groups when the other variables take their average values. This is what centering does. In centering, the mean value of the variable is subtracted from all of the values of the variable. In this way, the value of the centered variable is zero when its original value equals the mean, and the main effect of the other variable indicates the effect when the other variables take their mean value. Standardization not only centers the variable but also

divides each value by the standard deviation, thus allowing different variables to have the same scale (i.e., a mean of zero and a standard deviation of one) and to be comparable. Figure 1G demonstrates the case where L2 experience is standardized. Notice that the figure is nearly identical to the one before, the only difference being the values of the horizontal axis. The main effect of L1 is significant for both L1 groups in Figure 1F because both L1 Russian and L1 Spanish groups mark moderately different scores from the L1 Japanese group when L2 experience = 0. However, when L2 experience is standardized in Figure 1G, the main effect is non-significant for both L1 groups because between-L1 differences are small at the average point of L2 experience (i.e., standardized L2 experience = 0). It is often sensible to do this as it makes the coefficients more meaningful.

Generalized Linear Models

In the regression models discussed thus far, the dependent variable and the independent variables were linearly related in the scale of the dependent variable. This, however, is at times inconvenient. Suppose that we want to model the accuracy of a linguistic feature as a function of L2 experience and that the accuracy is measured as the percentage of correctness in obligatory contexts. If we build a regression model similar to the one discussed previously, the predicted accuracy may exceed 100% or go below 0% at large and small values of L2 experience. As percentage can only fall between 0 and 100, this would not be an appropriate model. To avoid this issue, we build a *generalized linear model* (*GLM*) with a non-identity link function and non-normal error distribution. A GLM is a regression model whose dependent variable is transformed by what is called a link function, and it assumes error distribution in any of the exponential family such as normal, binomial, or Poisson distribution (Hoffmann, 2004). When a GLM employs the identity link function and assumes normal error distribution, it is identical to the

type of regression models discussed earlier. In other words, the regression model we have discussed thus far is a special case of a GLM. A GLM covers a variety of models, and in modeling accuracy, we can use the logit link function and binomial error distribution. The logit link function transforms probability (e.g., accuracy) such that the transformed value can take any value between $+\infty$ and $-\infty$ while maintaining a monotonic relationship. Or perhaps more commonly, an inverse-logit function can be applied to convert predicted logit values into probability so that large or small values in the predictors still fall between 0 and 1 of the dependent variable. Binomial error distribution is assumed because probability dependent variables often follow binomial distribution. This model has also been known as the logistic regression model.

This point is illustrated in Figure 1H, where accuracy is modeled by L2 experience. If a normal regression model is employed, the regression line (dashed line) falls below 0% and exceeds 100%. If a GLM is employed, however, the increment or decrement of accuracy levels off in the probability scale as the value of the independent variable becomes either large or small. The logistic regression line is linear in the logit scale where the independent variable linearly exerts influence, but it becomes nonlinear when the value is back-transformed into the probability scale (i.e., the original scale) by applying the inverse-logit function. In this way, independent variables retain the same form as before but the dependent variable only takes the value between 0 and 1.

Another feature of logistic regression modeling is that it weighs each observation according to its data size. For instance, suppose that a learner was supposed to use a linguistic feature 100 times and correctly supplied it 50 times, while another learner was supposed to use the feature four times and correctly used it twice. Although the accuracy of the feature is 50% for

both learners, the former case is much more reliable. In other words, in the former, the true ability of the learner is likely to be within the small region around 50%, while the region is much larger for the latter learner. Logistic regression takes this into account as it weighs each observation by the number of attempts. Logistic regression thus weighs the former case much more than the latter when estimating the parameter in the model.

Unlike simple and multiple regression models discussed earlier, where parameters were estimated by minimizing squared residuals, GLMs estimate parameters through what is called *maximum-likelihood* estimation (Myung, 2003). Maximum likelihood is an iterative process that maximizes the likelihood that observed data are obtained given the coefficients. It gradually shifts the values of the coefficients, computes how likely the observed data are obtained given the new set of coefficients, compares the likelihood with the likelihood previously obtained, and shifts the coefficient values into the direction that is likely to increase the likelihood of observing the data given the new coefficients. Through this process, parameter values converge on the optimal values that are most likely to have generated the observed data.

Model Comparison

A common way to tell whether an independent variable influences the dependent variable is by comparing multiple models and testing whether the best model includes the variable of interest (Long, 2012). Two ways have often been used to compare models; likelihood ratio tests (LRTs) and information-theoretic measures such as Akaike Information Criterion (AIC). The LRT examines whether a model has a significantly better fit to the observed data than another model. The test, however, is limited in two ways. First, it is unclear in what sense the model chosen by LRTs is “better” because LRTs do not directly relate to the inference one can draw from models, such as predictive accuracy (Burnham & Anderson, 2002). Second, LRTs can generally only be

used in *nested* models. A model is nested if it is a subset of another model. For instance, a regression model with x_1 as the sole independent variable ($\hat{y} = \beta_0 + \beta_1 \times x_1$, where β_0 and β_1 are estimated from the data) is nested within a regression model with x_1 and x_2 as the independent variables ($\hat{y} = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2$) because the second model would be identical to the first if $\beta_2 = 0$.

AIC overcomes both of these weaknesses. It has been mathematically shown that AIC is a measure of how close the model is to the true model that generated the sample data on which the built model is based (Harrell, 2001; Long, 2012). Put another way, AIC is an index of predictive accuracy, or how well the model generalizes to new data. In general, the more parameters a model has, the better fit it has to the data. The fit never worsens when the number of independent variables increases. Having independent variables that are unrelated to the dependent variable, however, deteriorates prediction accuracy, as the model with irrelevant predictors may overfit the observed data and capture the randomness or noise that does not generalize to new data (Myung, 2000; Pitt & Myung, 2002; Venables & Dichmont, 2004). What AIC does is to balance the complexity (i.e., the number of parameters) and the goodness of fit of the model. The model with lower AIC values (often referred to as a more *plausible* model) is more likely to make better predictions regarding new data. Moreover, AIC can also be used to compare non-nested models.

Mixed-Effects Models

The regression models discussed thus far assume independent observations. However, in longitudinal studies, data are correlated within learners. This paper addresses the dependency of data through *mixed-effects models*. As mentioned earlier, the purpose of regression modeling in general is to explain the variance in the dependent variable using a set of predictors. The mixed-

effects model partitions the variance into multiple levels (Baayen, 2008; Baayen, Davidson, & Bates, 2008; Cunnings, 2012; Dingemanse & Dochtermann, 2013; Hox, 2002). In this paper, the total accuracy variance between writings is divided into three levels; (i) L1-level variance (i.e., certain L1 groups are more accurate in morpheme use than other groups in general), (ii) learner-level variance (i.e., some learners in an L1 group are more accurate than others in the same L1 group), and (iii) writing-level and morpheme-level variance (i.e., accuracy changes within individual learners as they develop and also across morphemes). Why is it necessary to divide the total variance into separate levels? An important point here is that the data are correlated within L1 groups and within learners. A highly proficient learner, for example, is likely to achieve high accuracy throughout, while a low proficiency learner is likely to show a reverse pattern. This means that observations are not independent from each other because accuracy can be more or less predicted based on which learner composed the writings the accuracy is calculated in or which L1 group the learner belongs to. Ignoring the assumption of independent observations leads to unjustifiably small standard errors, which in turn invites spurious "significant" results (Hox, 2002). To account for data dependency, we need to capture the between-L1, between-learner, and between-writing (and between-morpheme) accuracy differences separately. This alone explains accuracy variance to a certain extent.

For the sake of simplicity, the following example assumes a two-level model that divides accuracy variance into learner-level and morpheme-level variance. The mixed-effects model takes into account individual variation by allowing the intercept and the slope of the regression line to vary across learners. The intercept here represents the accuracy of the reference-level morpheme, or articles, while in the context of this paper there are two slopes or contrasts representing the accuracy difference between articles and the past tense *-ed* and the accuracy

difference between articles and the plural *-s*. Allowing the intercept to vary across learners is called *random intercepts* and allowing the slope to vary is called *random slope* or *random contrast*. By adding these random effects, it is possible to model the relationship between morpheme and its accuracy when the variance at the learner level is accounted for.

Predictors can explain both learner-level and morpheme-level variance. To explain variance at the level of learners, learners' L1 might be a good predictor to account for intercept differences (i.e., learners with a particular L1 background outperform those with another L1 in the reference-level morpheme). Notice that the value of L1 is unchanged across morphemes in the same learner. This is why it explains between-learner variance and not within-learner variance. To explain variance within learners, the morpheme is a good predictor because its value changes within learners. In this example, the learner is called a *random-effects* variable, and L1 and morpheme are called *fixed-effects* variables. The term mixed-effects stems from the feature of the model that the two effects are put into a model simultaneously.

It is not always easy to decide whether a variable is a fixed effects or random effects. The basic idea is that in random effects, we assume that the levels (e.g., learners) are randomly drawn from a normally distributed large population, and while they differ in many ways, we are not necessarily certain of or interested in why and how they differ (Crawley, 2007). Whereas we know that individual learners vary, we are not necessarily interested in how each of them performs. Rather, we often want to make general inferences that are not dependent on the particular group of learners. It is appropriate, then, to have learners as a random-effects variable and not a fixed-effects variable (cf. Pinheiro & Bates, 2000).

Figure 2 visualizes the point of random contrasts and how predictors reduce variance based on hypothetical data. The vertical axis represents the TLU score, and the horizontal axis

represents morphemes (articles and plural *-s*). Here, let us suppose that Figure 2A represents the accuracy of articles and plural *-s* in a number of learners, each represented by one line. In this case, the accuracy difference between the two morphemes (i.e., contrast) is constant across learners. What differs is the absolute accuracy of each learner. A learner marks the TLU score of 0.4 in articles, while another marks 0.8, and the rest in between. These differences in the absolute accuracy between learners should be taken into account in modeling, which is what random intercepts do. Random intercepts make adjustments to the mean accuracy on an individual basis and allow learners to be of different overall accuracy. The accuracy of the learners at the intercept (where morpheme = articles) is {0.80, 0.76, 0.73 . . . 0.44, 0.40}, and the variance is 0.017. This is the variance between learners at the intercept, and one question we can ask is how much of it can be explained by the predictors. Let us say that the learners, in fact, had two different L1s, L1 Spanish and L1 Japanese, and in Figure 2B L1 Spanish learners are represented by dashed lines and L1 Japanese learners by solid lines. Here, L1 explained some portion of the variance in the random intercept. Now between-learner variance at the intercept should be computed within each L1 group, and the value (0.005 for both groups) is much smaller than the original variance (0.017). The reduction is achieved by taking L1 into account.

The lower two panels illustrate an example of random contrasts. Let us suppose here that the accuracy of articles was the same across learners, but the accuracy of plural *-s* varied. As a result, the accuracy difference between the two morphemes ranges from 0.350 to -0.350 depending on learners, and its variance is 0.053. This is called b*y-morpheme random contrasts* because the morpheme contrast (i.e., the accuracy difference between morphemes) varies across learners. Introducing random contrast takes the difference into account in modeling the data: It makes adjustments to the accuracy difference between morphemes on an individual basis. Again,

this difference can be explained by L1. However, this time, it is not L1 that explains the accuracy difference between morphemes but the interaction between L1 and morpheme. L1 as a predictor only allows the overall accuracy to vary depending on learners' L1s while other variables are held constant. This was fine for explaining random intercepts because random intercepts only take care of the overall accuracy and are not related to between-morpheme accuracy difference. However, varying accuracy difference across learners means a varying effect of morpheme depending on learners' L1s. This type of effect can only be captured by *cross-level interactions* of predictors, which are the interactions between learner-level and morpheme-level predictors (Hox, 2002). Introducing the L1-morpheme interaction allows the contrast to vary depending on learners' L1s, which is exactly what we want in order to capture random contrast variance. When the interaction is introduced into the model, the variance of the accuracy difference between the two morphemes reduces to 0.014, indicating that the between-morpheme accuracy difference is partially explained by L1.

Although random effects models make adjustments to the mean, the adjustments (called *conditional mode*; Bates, 2010) are not the parameters of the model. The model instead estimates the variance of the adjustments based on the assumption that they are normally distributed. In addition to the variance parameters, mixed-effects models often estimate correlations between random effects within individual learners (Baayen, 2008; Kliegl, Masson, & Richter, 2010). When both random intercept and random contrast are simultaneously entered in a model, the correlation tells us whether those with higher accuracy in articles tend to have higher or lower between-morpheme contrast values.

Figure 2 is, of course, a highly idealized scenario and real data are much messier. Hopefully, however, the point is clear as to what random effects mean and how they can be explained. GLMMs are an extension of mixed-effects models. In the same way that simple and multiple regression models are extended to GLMs as discussed earlier, GLMMs can have non-identity link functions and handle non-normal errors (Barr, 2008; Bolker et al., 2009; Dixon, 2008; Gelman & Hill, 2007; Hox, 2002; Jaeger, 2008; Quené & van den Bergh, 2008).

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474. doi: 10.1016/j.jml.2007.09.002

Bates, D. M. (2010). *Lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–35. doi: 10.1016/j.tree.2008.10.008

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach (Second edition)*. New York: Springer-Verlag New York.

Crawley, M. J. (2007). *The R book*. West Sussex: John Wiley & Sons.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369–382. doi: 10.1177/0267658312443651

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *The Journal of Animal Ecology, 82*(1), 39–54. doi: 10.1111/1365-2656.12013

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. doi: 10.1016/j.jml.2007.11.004

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

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hoffmann, J. P. (2004). *Generalized linear models: An applied approach*. Boston: Pearson Education.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. New York: Lawrence Erlbaum Associates.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. doi: 10.1016/j.jml.2007.11.007

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655–681. doi: 10.1080/13506280902986058

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Thousand Oaks, CA: Sage Publications.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*(1), 190–204. doi: 10.1006/jmps.1999.1283

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(1), 90–100. doi: 10.1016/S0022-2496(02)00028-7

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer New York.

Pitt, M. a., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*(10), 421–425. doi: 10.1016/S1364-6613(02)01964-2

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413–425. doi: 10.1016/j.jml.2008.02.002

Venables, W., & Dichmont, C. (2004). GLMs, GAMs and GLMMs: An overview of theory for applications in fisheries research. *Fisheries Research, 70*(2-3), 319–337. doi: 10.1016/j.fishres.2004.08.011

Appendix S2: Accuracy of the R Scripts Used to Identify Errors

The accuracy of R scripts used to identify the errors of the target morphemes was manually verified against error annotation as the gold standard. That is, I checked accuracy on the assumption that the error tags are exhaustive and accurate. A hundred errors were manually identified in each morpheme such that the number of identified errors in each Englishtown level is proportional to the total number of words of all of writings submitted at that level. Errors were identified in a different set of writings from those used to tune the script. Table 1 shows the results. Precision refers to the percentage of correct hits, while recall refers to the degree to which the script captures what it is intended to capture. For example, if a script to count article errors identified 70 instances of errors and 60 out of the 70 included errors, the precision rate is 86% (60/70). If, however, there are 100 instances of article errors, the recall rate is 60% (60/100) because only 60 out of the 100 cases that should have been captured were indeed captured. F1 is the harmonic mean of precision and recall, and represents the total accuracy of the script. It is calculated by

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall}$$

Overall, script accuracy is fairly high in all of the three morphemes; thus results based on these scripts should be generally reliable.

Table 1 Accuracy of the scripts used to retrieve errors

Morpheme	Precision	Recall	F_1
Articles	90%	98%	0.94
Past tense *–ed*	76%	85%	0.80
Plural *–s*	75%	88%	0.81

Appendix S3: Correlation Parameter and Shrinkage in Mixed-Effects Models

In mixed-effects models, we can gain insights into systematic individuality by looking into within-learner correlations between random effects. Based on the final GLMM (Model 8) constructed in the main text, Table 2 shows the within-learner correlation between conditional modes (i.e., adjustments to individual intercepts and slopes). The negative correlation between the random intercept and the random contrast of past tense *-ed* (-0.397) means that when fixed-effects variables are accounted for learners with higher accuracy in articles (represented by random intercepts) tend to have a more negative contrast between article accuracy and the accuracy of past tense *-ed* (represented by the random contrast), or, put more simply, lower accuracy in past tense *-ed*. Because the overall accuracy of past tense *-ed* is higher than that of articles as indicated by fixed effects (cf. Table 4 in the main text), this means that the accuracy difference between articles and past tense *-ed* tends to be smaller if a learner marks relatively high accuracy in articles. This is natural because past tense *-ed* is closer to the ceiling and higher article accuracy cannot always be accompanied by higher past tense *-ed* accuracy. The correlational structure can inform us of systematic individual differences in this manner. A parametric bootstrap indicated that based on 1,000 samples this was the only significant correlation parameter at $p < 0.05$, and thus I will not interpret the other correlations in the table.

Shrinkage in Mixed-Effects Models

Although mixed-effects models make adjustments to intercepts and slopes at the individual level, the adjustments are not made to minimize the difference between observed and predicted values due to a notable feature of mixed-effects models called *shrinkage* (Baayen, 2008; Gelman & Hill, 2007; Kliegl, Masson, & Richter, 2010). The idea is that the data points of individual learners

may be unreliable owing to their small sample size. Therefore, when mixed-effects models make predictions, the values shrink toward the population mean because it is presumably more reliable. The degree of shrinkage is larger in the learners when (i) their values are extreme, (ii) their number of observations is small, and (iii) their variance is large (Kliegl et al., 2010).

Figure 1 illustrates shrinkage. The figure demonstrates the longitudinal development of article accuracy in L1 Chinese learners. Each panel shows the longitudinal development of one learner. Each bubble represents the TLU score of a writing, and its size corresponds to the reliability of the value indicated by the number of obligatory contexts and overgeneralization errors. The three lines are the predicted values based on the GLMM (solid line, based on Model 8 in the main text), the GLM on all the data (dotted line, based on Model 7 of the GLM discussed in Online Supporting Document 4), and the GLMs on individual learners' data (dashed line). For the individual GLMs, I built for each learner a logistic regression model including (standardized) proficiency and (standardized) writing number as the predictors. Each individual GLM only targeted the data in one learner, and shrinkage is not in effect, as the GLM does not know the population mean. The overall GLM was constructed without taking into account individual variation, and thus shrinkage is not in effect in this model, either, as the GLM does not know data dependency within individual learners.

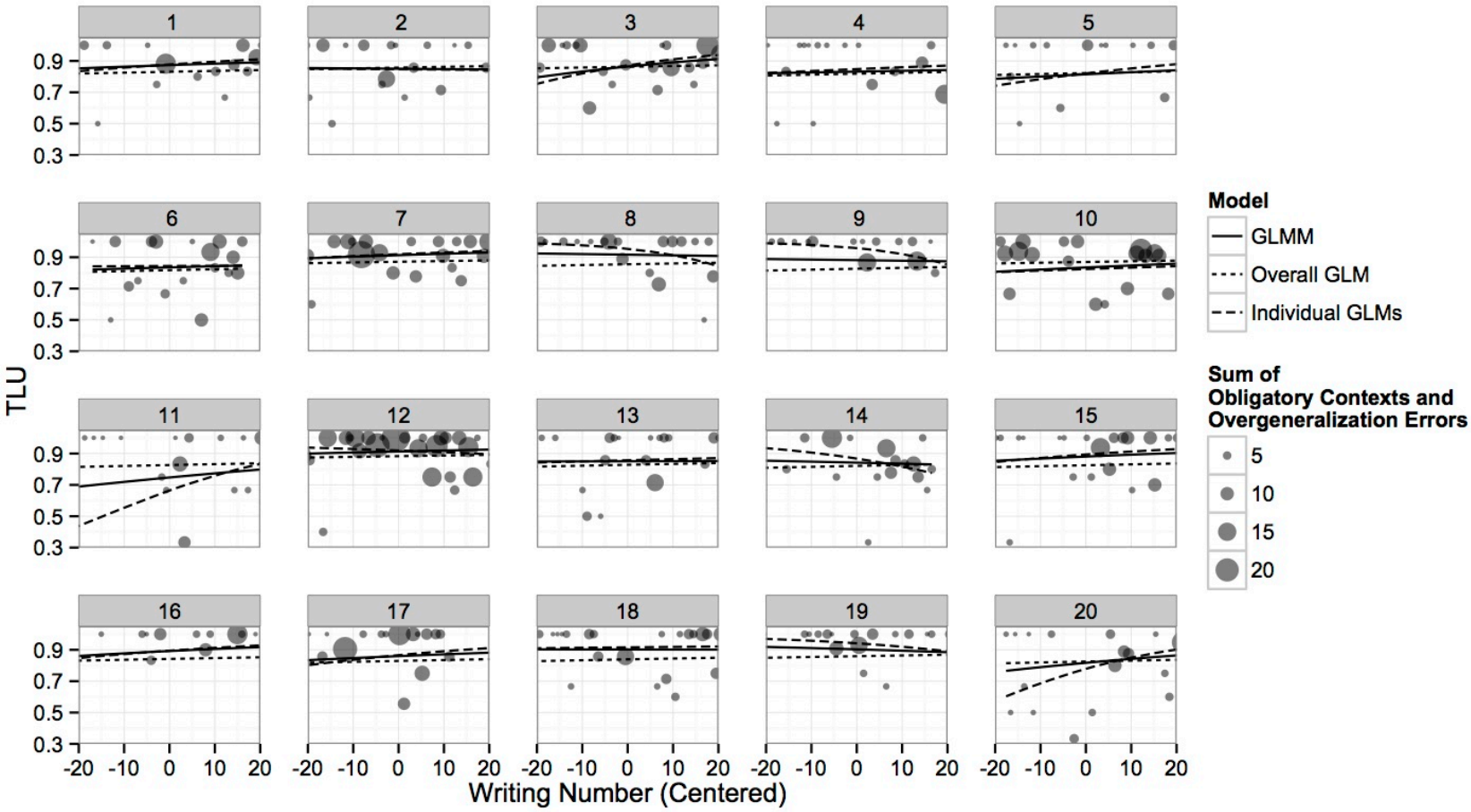

Figure 1 GLMM versus GLM in the longitudinal development of article accuracy in L1 Chinese learners

We can see that while the three approaches make similar predictions in many learners, we can also observe some prominent differences where the individual GLMs make the predictions that are closer to the observed values but are more different from the overall pattern than the GLMM, which in turn is more flexible than the overall GLM because it makes adjustments for individual learners while the GLM does not. In Panel 20, for example, because this learner's observed accuracy tends to be low at the beginning, the individual GLM predicts relatively low accuracy at the beginning. The GLMM, however, draws on the overall mean and predicts higher scores. It believes that the observed low accuracy occurred by chance because it differs considerably from the population mean (i.e., mean of all of the learners) and that therefore the true ability of the learner is likely to be higher than the observed performance. It, however, still predicts that his/her true accuracy is lower than the average represented by the overall GLM. By making individual adjustments, the GLMM strikes a balance in this manner between what can be inferred from the average and observed data points. Similar is the case in Panel 11. In Panels 8 and 9, the reverse is true. While the observed accuracy is on a decreasing trend in these learners, the GLMM's predicted values show less extreme patterns. This, again, occurs because the overall longitudinal developmental pattern in the population is accuracy increase, and the adjustments by the GLMM are made toward the overall pattern.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655–681. doi: 10.1080/13506280902986058

Table 2 Correlation between random effects

	Morpheme (past tense *–ed*)	Morpheme (plural *–s*)	Writingnum (standardized)
Learner			
Intercept	–0.397	–0.264	0.028
Morpheme (past tense *–ed*)		0.408	–0.062
Morpheme (plural *–s*)			0.022

Appendix S4: Generalized Linear Models and Generalized Additive Models

Generalized Linear Models (GLMs)

Model Specification and Model Selection

The models assumed binomial error distribution and used a logit link function. In other words, logistic regression models were constructed. As in the GLMM/GAMM, the dependent variable was accuracy in the form of odds. The potential independent variables were L1 type, morphemes, standardized proficiency, standardized writing number (writingnum), and their interactions. As in the GLMM, I employed maximum likelihood for estimation and the AIC-based forward-selection procedure for model selection: A variable was entered into the model only when it reduced AIC.

Table 1 shows the model selection procedure.

1. Model 1 is the intercept-only model without any predictors.
2. Model 2 added morpheme and this improved the model. It indicates that different morphemes are of different accuracy.
3. Model 3 additionally included proficiency, leading to further model improvement.
4. Model 4 likewise added L1type.
5. Model 5 entered the morpheme-proficiency interaction, indicating that the accuracy difference between morphemes varies across proficiency levels and that cross-sectional developmental patterns vary across morphemes.
6. Model 6 further included writingnum. This means that accuracy changes as learners develop.
7. Model 7 added the proficiency-L1type interaction, which suggests that cross-sectional developmental patterns differ between the ABSENT and the PRESENT learners.

While adding the proficiency-writingnum interaction to Model 7 very marginally decreased AIC (ΔAIC = -0.16), the parameter was not included in the final model as the reduction was too small. A likelihood ratio test did not support the inclusion, either (χ^2 (1) = 2.16, p = 0.142). Thus, I consider Model 7 as the final model, and the model included morpheme, proficiency, L1type, writingnum, the morpheme-proficiency interaction, and the proficiency-L1type interaction.

Interpretation of the Model

Table 2 shows the summary of Model 7. The main effect of morpheme (Row 2-4) indicates that both past tense *-ed* and plural *-s* are more accurate than articles. The morpheme-proficiency interaction (Row 9-11), however, suggest that this is only the case at the mean proficiency level, and the difference shrinks as learners' proficiency rises. The main effect of proficiency (Row 5) suggests that article accuracy increases as learners' proficiency gores up in the ABSENT group. However, the morpheme-proficiency interaction again shows that accuracy increase in the other two morphemes is much smaller. In fact, on average, the cross-sectional development of plural *-s* is better characterized by very marginal accuracy decrease rather than accuracy increase (0.203 - 0.217 = -0.014). Proficiency further interacts with L1type (Row 12-13). The interaction suggests that the rate of accuracy increase is higher in the PRESENT group than in the ABSENT group. The main effect of L1type (Row 6-7) indicates that the PRESENT group outperforms the ABSENT group. Its interaction with proficiency shows that the accuracy difference between the two groups is larger at higher proficiency levels. Finally, the main effect of writingnum (Row 8) supports accuracy increase as learners develop.

Figure 1 visualises the predicted cross-sectional development based on Model 7 across the morphemes and across the L1 types at writingnum = 0. The shaded area represents the 95% confidence interval. It can be seen that, as Table 2 suggests, cross-sectional developmental

patterns differ depending on morphemes. Accuracy increase is the steepest in articles, less steep in past tense *-ed*, and least steep in plural *-s* possibly because plural *-s* is the most accurate morpheme and learners have reached the ceiling. The figure also demonstrates that the accuracy difference between the ABSENT and the PRESENT group increases as the proficiency goes up. Given the relatively high accuracy in all the three morphemes, the accuracy increase is slower in the ABSENT group possibly because they have reached the ceiling that is difficult to surpass without the assistance of L1 (Jiang, Novokshanova, Masuda, & Wang, 2011).

Generalized Additive Models (GAMs)

Model Specification and Model Selection

The models assumed binomial error distribution and used a logit link function. The dependent variable was accuracy in odds. The potential independent variables were L1 type, morphemes, standardized proficiency, standardized writing number (writingnum), and their interactions. For proficiency and writingnum, both linear and nonlinear terms were considered. Once a variable was entered as a smooth, the interaction terms that include the variable were also turned into smooths. I will explain this in more detail in the model selection part.

All of the nonlinear terms were first entered with thin plate regression splines. However, the final model turned out to have a proficiency-writingnum nonlinear interaction as a tensor product smooths. When smooths are nested, as in the case where a model includes both the proficiency-writingnum interaction and the main effect of proficiency as smooths terms, it is better to use the same bases for smooths (Wood, 2010). The present analysis thus employs tensor product smooth throughout the process. A separate smooth was constructed for each factor level in the interaction between a factor and a smooth.

As in the GLMMs, maximum likelihood estimation was generally employed for parameter estimation. It is, however, more desirable to use restricted maximum likelihood (REML), a variant of maximum likelihood, to compare two models with the same parametric terms but different smooth terms (Wieling, 2015). The present analysis followed this practice: When testing whether to add a smooth term to the model, the analysis used REML to build both models that are compared.

Table 3 shows the constructed models and the results of the comparison between them. AIC (ML) shows the AIC of the models based on maximum likelihood estimation, while AIC (REML) shows that of the models based on REML. The results of likelihood ratio tests are not presented in the table due to space limitations, but they agree with the AIC-based comparison with $p = 0.05$ as the significance level. The point of the comparison is whether the most plausible model includes the L1type-proficiency interaction or the L1type-writingnum interaction. If it does, it shows that L1 type affects cross-sectional and/or longitudinal development, and we can explore the model to analyze how L1 influences changes throughout development.

1. Models 1 through 5 are GLMs without any smooth terms. Model 1 is the intercept-only model that does not include any predictors. In Models 2 through 5, morpheme, proficiency, L1type, and morpheme-proficiency interaction were sequentially added to Model 1, and improvement of the model was observed at each step.
2. Model 6 additionally included nonlinear writingnum smooth, leading to a further improvement of the model. The nonlinear effect of writingnum improves the model marginally more than the linear effect (-11.8 vs -11.0 ΔAIC).
3. Model 7 further added a proficiency-writingnum wiggly surface. Introducing this nonlinear interaction automatically allows the main effect of proficiency to be nonlinear as well. To

be consistent with it, the morpheme-proficiency interaction was also replaced with separate by-morpheme wiggly proficiency curves.

4. By-L1type separate wiggly writingnum curves were entered in Model 8. Comparison with Model 7 supported the difference in the longitudinal developmental pattern between the ABSENT and PRESENT group.

Although further adding by-L1type separate wiggly proficiency curves marginally reduced AIC (ΔAIC (REML) = -2.4), the model was not considered more plausible than Model 8 due to the small size of ΔAIC. We thus take Model 8 as the most plausible model. This model includes morpheme and L1 type as parametric terms, and as smooths terms separate wiggly proficiency curves for each morpheme, separate writingnum curves across L1 types, and a proficiency-writingnum wiggly surface.

Interpretation of the Model

Interpreting Parametric Terms

Table 4 shows the parametric terms of Model 8. When the nonlinear effects of proficiency and writingnum and their interaction are controlled for, both past tense *–ed* and plural *-s* are significantly more accurate than articles. The L1 type parameter indicates that the PRESENT group outperforms the ABSENT group when nonlinear effects are taken care of.

Interpreting Smooth Terms

Table 5 shows estimated degrees of freedom (EDF), reference degrees of freedom (Ref.df), χ^2, and *p*-values for the splines. The presence of the L1type-writingnum (Row 6-8) interaction indicates that the general, morpheme-independent longitudinal developmental pattern varies across L1 types.

Smooth terms are generally not easy to interpret, especially when they participate in multiple interactional terms as in the present case. An appropriate way to explore the model in GLM AND GAM 7 such a case is through the visualization of fitted (or predicted) values. Figure 2 illustrates the fitted values with wiggly surfaces. It visualizes the nonlinear cross-sectional and longitudinal development across the three morphemes and two L1 types. In each panel, the horizontal axis represents the overall proficiency of learners (i.e., cross-sectional development), and the vertical axis represents centered writing number (i.e., longitudinal development). Each writing is represented by a small dot, and the part of the graph with denser dots is likely to be more reliable. Although a few learners produced more than 62 writings (= mean $\pm$2 SDs), the figure only shows the fitted value of the development over 62 writings, which captures 94.5% of the data. Shade indicates accuracy. Darker gray corresponds to lower accuracy and lighter gray represents higher accuracy. A contour line is drawn by 0.025 TLU score. In other words, the accuracy between two lines differs by 0.025. The figure does not display the part that is far from the regions where predictors lie. In the article PRESENT panel, the color tends to become lighter from left to right, which indicates that as learners' overall proficiency goes up, so does the accuracy of articles. If we look at the same panel from the bottom to top, the shade changes from dark to light at lower proficiency levels. This indicates that as learners produce more writings, the accuracy of articles increases.

We can make three observations about the figure. First, both cross-sectional and longitudinal developmental patterns are nonlinear in the probability scale. For example, at lower proficiency (e.g., proficiency = 4) in the article PRESENT panel, contour lines are not drawn equidistantly. There are more lines towards the bottom, which indicates that accuracy increase slows down as learners produce more writings, just like power-law development. A similar

pattern is observed in the article ABSENT panel as well. Furthermore, we can also observe nonlinear cross-sectional development. If we look horizontally at the centered writing number of approximately -20 in the past tense *-ed* PRESENT panel, we can again see that there are more contour lines at early stages of development (e.g., up to proficiency seven) than at later stages (e.g., proficiency of 10). Second, the nonlinear developmental pattern interacts with overall proficiency. In the article PRESENT panel, accuracy at lower proficiency levels tends to increase as learners produce more writings. However, at higher proficiency levels, accuracy remains relatively unchanged. This indicates that the developmental pattern differs across the overall proficiency of learners. Third, the two nonlinear effects further interact with morpheme. Accuracy tends to be more stable in plural *-s* than in articles and past tense *-ed* both cross-sectionally and longitudinally, perhaps due to the ceiling effect. This is particularly the case at lower proficiency levels.

Although the fitted figure as a contour plot is comprehensive and informative, it can be cognitively demanding to determine precise accuracy transition with it. To complement the figure, Figure 3 illustrates the fitted cross-sectional and longitudinal development across the three morphemes and the two L1 types at the mean writing number (upper panels) and the fitted longitudinal development at Level 4 Unit 1 (lower panels). In the upper panels, the horizontal axis represents learners' proficiency, and the vertical axis represents fitted TLU scores. The curves in each panel are the predicted TLU score for each L1 type, and the shaded area corresponds to the 95% confidence interval. Each tick mark at the bottom of the panels represents one learner (upper panels) or one writing (lower panels). Regions with denser marks are where the fitted value is likely to be more reliable. The lower panels are similar to the upper panels except that the horizontal axis represents centered writing number.

We can make a few observations here as well. First, the cross-sectional developmental pattern is relatively linear, while longitudinal development is nonlinear, at least for the PRESENT group. Although the GAM supports nonlinearity in logit TLU in both proficiency and writingnum effects, the nonlinearity in proficiency effect does not look very strong. On the other hand, relatively clear nonlinearity is observed in longitudinal development at a low proficiency level, especially in the PRESENT group. It is difficult to draw a straight line from left to right without going outside of the shaded region, indicating nonlinear developmental patterns. Here, as was suggested in the contour plot presented earlier, we can see that accuracy increase slows down as learners progress. Second, the longitudinal developmental pattern differs across L1 types but the cross-sectional developmental pattern does not (cf. Table 3). This means that the strength of L1 influence does not change much across proficiency levels. Longitudinal developmental patterns, however, clearly differ between the ABSENT and PRESENT group. The PRESENT group exhibits wigglier learning curves than the ABSENT group. Third, as the GAM suggests, we can observe differences in the cross-sectional developmental patterns across the morphemes. Accuracy increase is more rapid in articles and in past tense *-ed* than in plural *-s*, whose accuracy is relatively unchanged throughout the development. Fourth, although the longitudinal development of the PRESENT group is somewhat reminiscent of power-law development, the developmental pattern in Figure 3 generally does not exhibit typical U-shaped or power-law learning curves.

References

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morpheme. *Language Learning, 61*(3), 940–967. doi: 10.1111/j.1467-9922.2010.00627.x

Wieling, M. (2015). *Analyzing EEG data using GAMs: Lecture 4 of advanced regression for linguists*. Retrieved 16 January, 2015, from http://martijnwieling.nl/statscourse/lecture4/presentation.pdf

Wood, S. (2010). *A toolbox of smooths*. Retrieved 18 June, 2014, from http://people.bath.ac.uk/sw283/mgcv/tampere/smooth-toolbox.pdf

Table 1 Comparison of GLMs

				Likelihood ratio test against the previous model	
Model	Parameter	AIC	ΔAIC	Statistic	*p* value
Model 1	None	14454.6			
Model 2	Morpheme	14023.9	-430.7	χ^2 (2) = 434.73	< 0.001
Model 3	Model 2 + proficiency (standardized)	13906.0	-117.9	χ^2 (1) = 119.89	< 0.001
Model 4	Model 3 + L1type	13852.5	-53.5	χ^2 (1) = 55.49	< 0.001
Model 5	Model 4 + morpheme-proficiency interaction	13835.1	-17.4	χ^2 (2) = 21.40	< 0.001
Model 6	Model 5 + writingnum (standardized)	13824.1	-11.0	χ^2 (1) = 12.98	< 0.001
Model 7	Model 6 + proficiency-L1type interaction	13819.7	-4.4	χ^2 (1) = 6.41	0.011

Table 2 Summary of GLM Model 7

Parameter		B		SE
Intercept	(Intercept)	1.769	***	0.029
Morpheme				
	Past tense -*ed*	0.157	*	0.069
	Plural -*s*	0.786	***	0.042
Proficiency (standardized)		0.203	***	0.027
L1type				
	PRESENT	0.269	***	0.036
Writingnum (standardized)		0.062	***	0.018
Morpheme : Proficiency (standardized)				
	Past tense -*ed* : Proficiency	-0.111		0.069
	Plural -*s* : Proficiency	-0.217	***	0.042
Proficiency (standardized) : L1type				
	Proficiency : PRESENT	0.091	*	0.036

Table 3 Comparison of GAMs

	Model description		AIC			
Model	Parametric terms	Smooths	AIC (ML)	ΔAIC (ML)	AIC (REML)	ΔAIC (REML)
Model 1	Intercept-only model	None	14454.6		14454.6	
Model 2	Model 1 + morpheme	None	14023.9	-430.7	14023.9	-430.7
Model 3	Model 2 + proficiency (standardized)	None	13906.0	-117.9	13906.0	-117.9
Model 4	Model 3 + L1type	None	13852.5	-53.5	13852.5	-53.5
Model 5	Model 4 + morpheme-proficiency interaction	None	13835.1	-17.4	13835.1	-17.4
Model 6	Same as Model 5	writingnum (standardized)	13823.7	-11.4	13823.3	-11.8
Model 7	Model 6 - proficiency - morpheme-proficiency interaction	Model 6 + proficiency-writingnum interaction + proficiency for each morpheme	13812.6	-11.1	13808.4	-14.9
Model 8	Same as Model 7	Model 7 - essaynum + writingnum for each L1type	13803.3	-9.3	13799.7	-8.8

Table 4 Parametric terms of GAM Model 8

Parameter		B		SE
Intercept		1.777	***	0.028
Morpheme				
	Past tense *-ed*	0.141	*	0.069
	Plural *-s*	0.788	***	0.042
L1type				
	PRESENT	0.278	***	0.036

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.10$

Table 5 Smooths terms of GAM Model 8

Parameter	EDF	Ref.df	χ^2	*p* value	
Proficiency (standardized) : Writingnum (standardized)	10.507	13.294	25.434	0.023	*
Proficiency (standardized) : Morpheme			13.751	< 0.001	**
Proficiency (standardized) : Articles	1.000	1.000	1		*
Proficiency (standardized) : Past tense - *ed*	1.662	2.053	8.403	0.016	*
Proficiency (standardized) : Plural -*s*	1.009	1.017	3.036	0.083	.
Writingnum (standardized) : L1type					
Writingnum (standardized) : ABSENT	1.002	1.003	0.717	0.398	*
Writingnum (standardized) : PRESENT	3.151	3.570	12.232	0.012	

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.10$

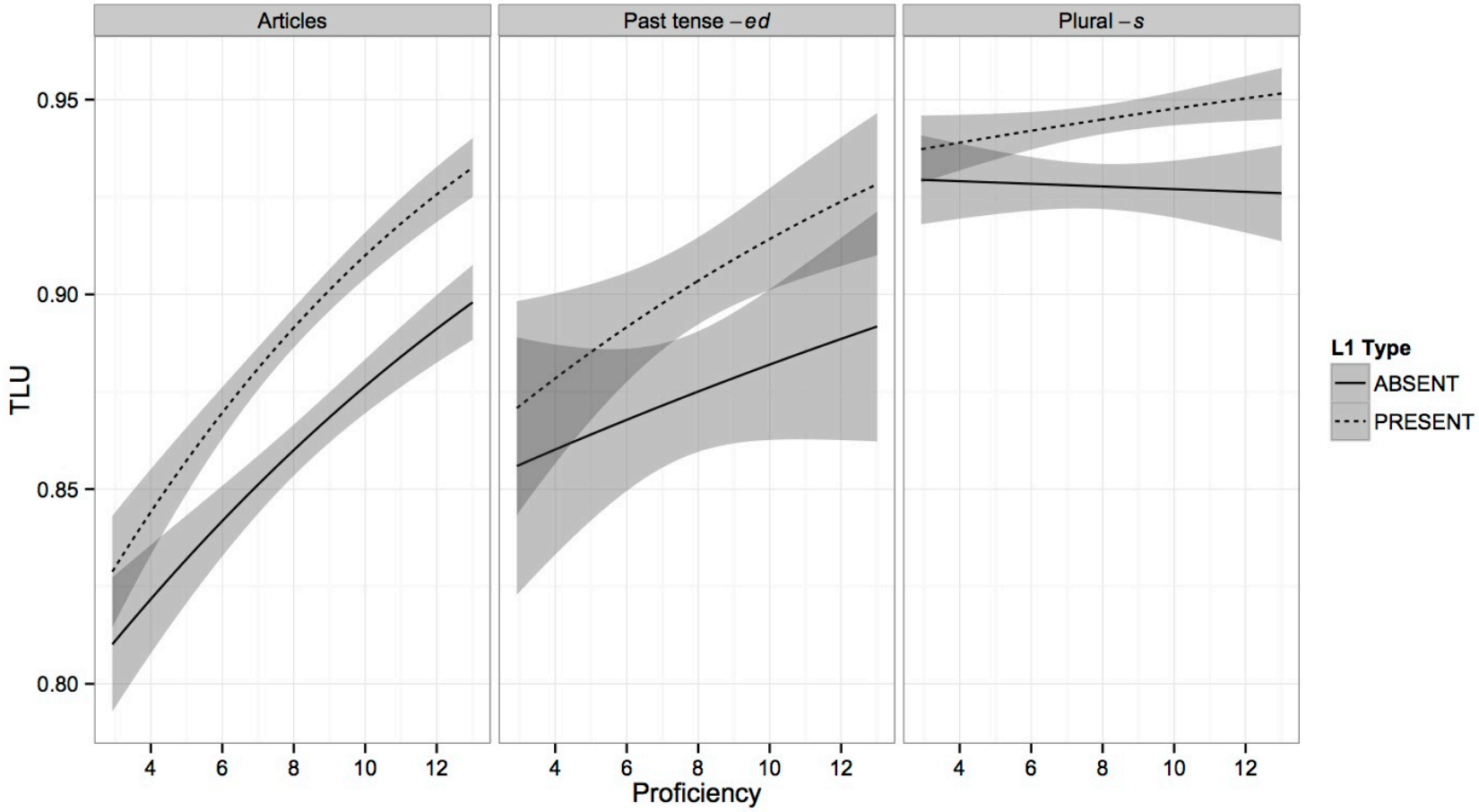

Figure 1 Fitted values of GLM Model 7.

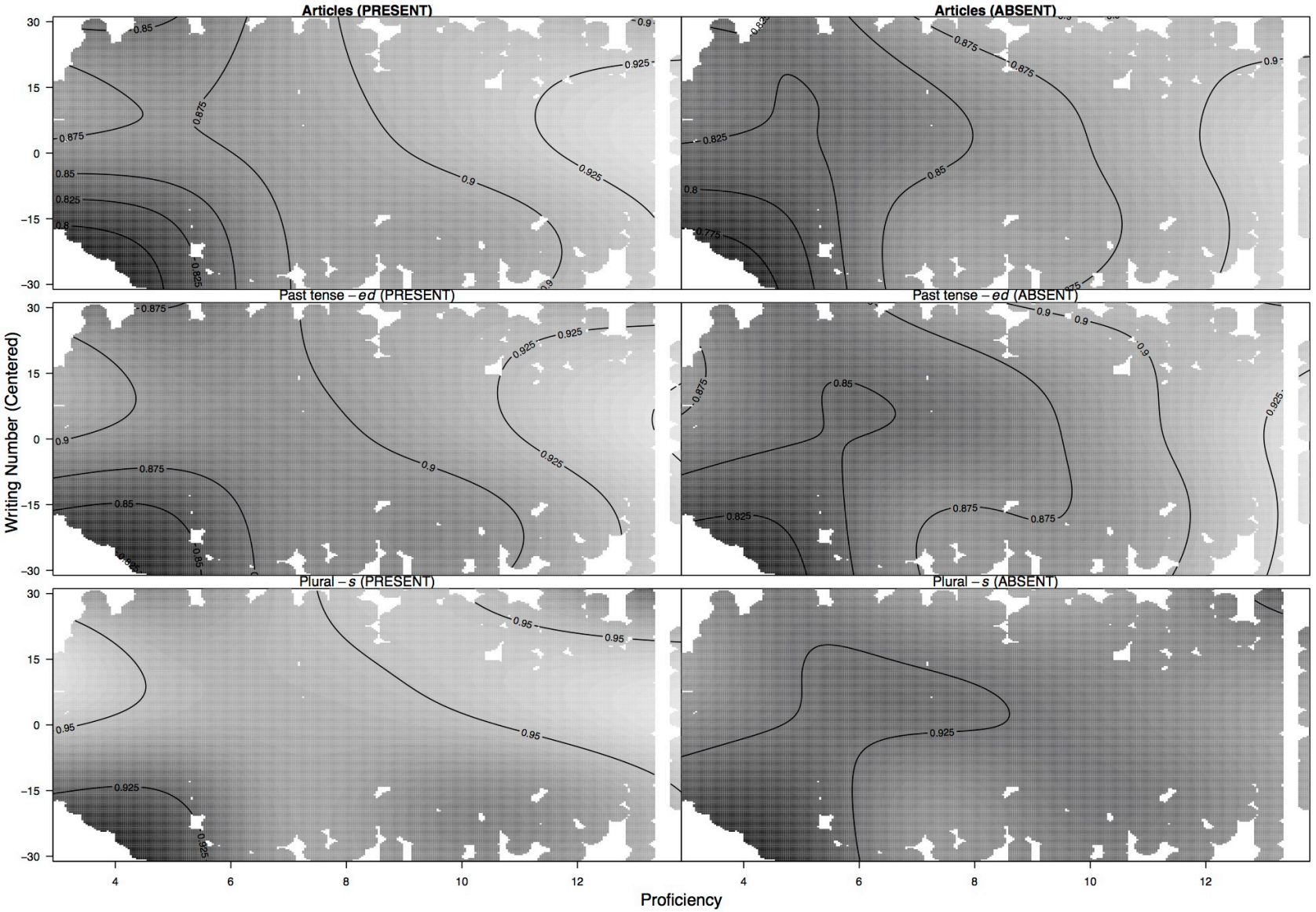

Figure 2 Fitted values of GAM Model 8

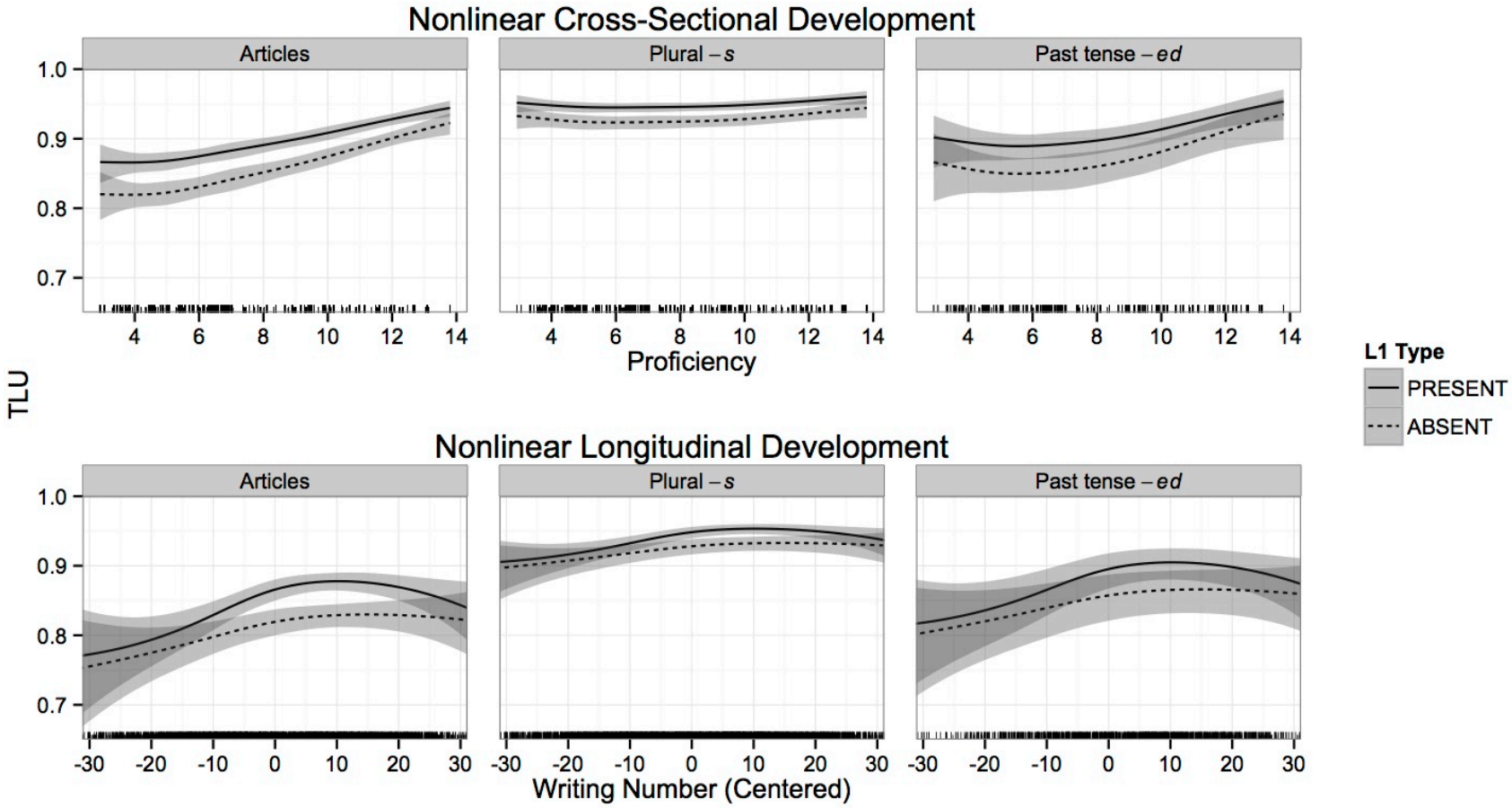

Figure 3 Nonlinear cross-sectional and longitudinal accuracy development