Supplementary Information:

Supplementary Method.

Early non-immersion learners. Subjects reporting learning in a non-immersion environment beginning at 1, 2, or 3 years of age exhibited strange results (Fig. S1). As noted in the main text, these were excluded.

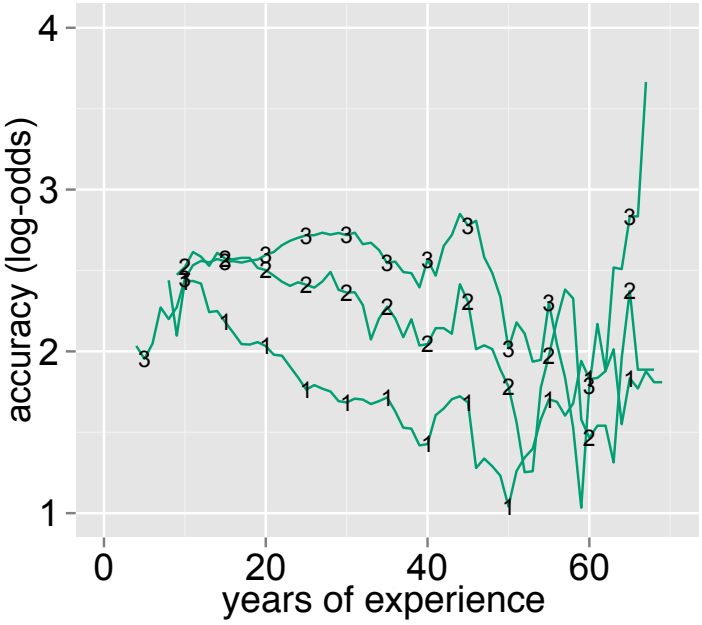


Figure S1. Performance curves for non-immersion learners with ages of first exposure of one, two, or three years (indicated by numbers overlaid on the lines).

Models of changes in the learning rate

Data preparation was identical to that of the permutation analyses. In order to find parameter values that minimized R^2, we employed Differential Evolution following a local-to-best strategy, with 500 iterations and a population size of 10x the number of parameters (Mullen et al., 2011).

Five different models were considered. The Exponential Learning with Sigmoidal Decay (ELSD) model is presented in the main text. Performance curves are derived by combining the two equations in the main text and integrating:

$$g(t) = \begin{cases} \left(1 - e^{-Er_0(t-t_e)}\right)a + b & , t_e \le t_c,\ t \le t_c \\ \left(\left(1 - e^{-Er_0(t_c-t_e)}\right) - 1\right)e^{-Er_0\left((t-t_c)+\frac{1}{\alpha}\ln\left(\frac{1+e^{-\alpha\delta}}{1+e^{\alpha(t-t_c-\delta)}}\right)\right)}a + b & , t > t_c, t_e \le t_c \\ \left(1 - e^{-Er_0\left((t-t_e)+\frac{1}{\alpha}\ln\left(\frac{1+e^{\alpha(t_e-t_c-\delta)}}{1+e^{\alpha(t-t_c-\delta)}}\right)\right)}\right)a + b & , t \ge t_e > t_c \end{cases}$$

with the additional of scale parameters *a* and *b*, which were set to 2.0 and 1.5, respectively, in order for the results to span the empirical range in log-odds. The ELSD model is capable of capturing a wide range of possibilities in terms of how learning ability changes over the lifespan (Figs. 1, 4E, S2).

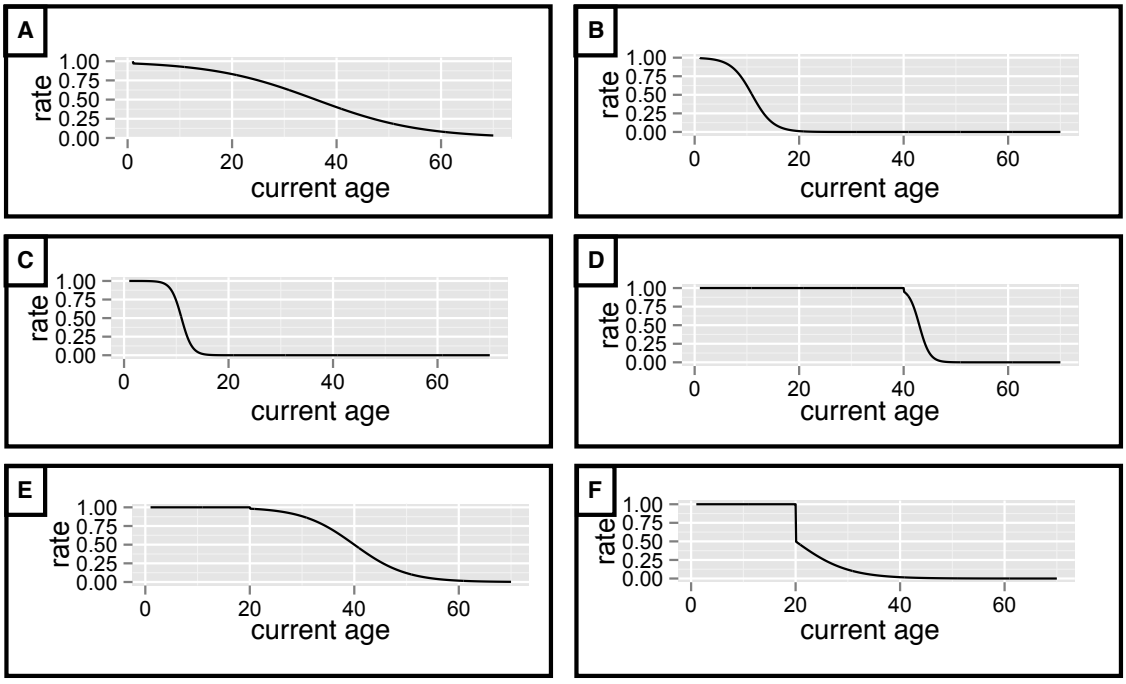


Figure S2. The ELSD model can consider learning rate declines that are slow (A), rapid (B),

or discontinuous (F), and which occur at any age between 1 and 40 (compare B with C with E). In these examples, the initial learning rate is set to 1.0 in order to better highlight the different shape possibilities; however, initial learning rate is also a parameter that must be fit. **A**: t_c = 1, r_0 = 1, α = 0.1, δ = 34. **B**: t_c = 1, r_0 = 1, α = 0.5, δ = 10. **C**: t_c = 1, r_0 = 1, α = 1, δ = 35. **D**: t_c = 40, r_0 = 1, α = 1, δ = 3. **E**: t_c = 40, r_0 = 1, α = 1, δ = -20. **F**: t_c = 20, r_0 = 1, α = 0.1, δ = 0.

We also considered modified versions of ELSD: a *discontinuous rate change* model, namely a simple step function in which the learning rate changed from r_0 to r_1 at age t_c, and a *flat rate* model, where the learning rate remained constant. We also included variants of the ELSD and the discontinuous rate change model in which the learning rates changed as a function of years of experience rather than age. In all cases, we used the same values for the scale parameters *a* and *b* (2.0 and 1.5, respectively). The best fitting parameters for these models are given in Figs. S4-S8. For ease of comparison, the empirical data from Figs. 4 and 7 have been combined in Fig. S3.

Note that Differential Evolution requires defining a range of possible values for each parameter. For all models the learning rate was constrained to be between 0 and 1. The age at which the learning rate began to change in the ELDS and discontinuous model was constrained to be between 1 and 40 years of age or experience, as appropriate. In the ELSD models, α could range from 0 to 1 and δ could range from -50 to +50. The experience discount factor *E* was set to 1 for monolinguals and could range between 0 and 1 for simultaneous bilinguals, later immersion learners, and non-immersion learners.

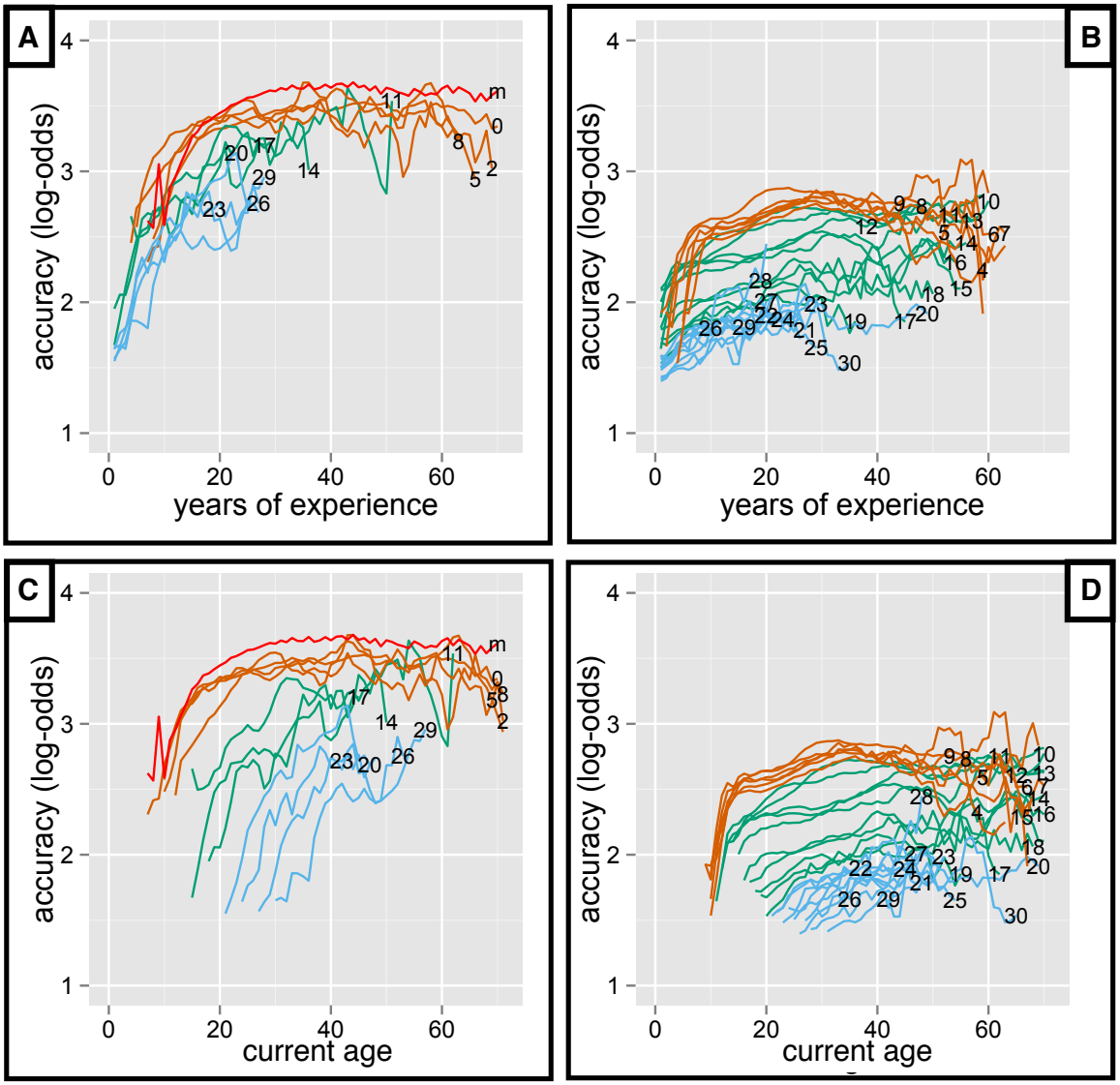


Figure S3. Empirically measured learning-curves. **A-B:** Performance as a function of years of experience for monolinguals and immersion learners (A) and non-immersion learners (B). **C-D**: Performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). Age of exposure is indicated on each curve.

In order to compare the fits of different models, we conducted ten runs of Monte Carlo split-half cross-validation, splitting the raw data, not the averaged data. In pairwise t-tests, the resulting R^2s were significantly higher for the ELDS model than for any other model. The reason is visible in the graphs: the flat-rate model and the models with rate

changes based on experience rather than age could not capture the differences in ultimate attainment across exposure ages. The discontinuous rate-change model fits nearly as well as ELSD, but cannot distinguish the slopes of the performance curves among participants who began learning at varying amounts of time after the learning rate changed (that is, in adulthood).

Note that because our interest was in changes in learning as a function of age of first exposure and experience, the model was fit to performance curves, not to the aggregate data as a whole: fitting to the raw data directly would have overweighted the monolinguals, who contributed nearly half of the data, and underweighted the later learners, vitiating the goals of the analysis. Thus, for instance, we calculated the squared difference between the predicted value for immersion learners who began at 5 years and had 10 years of experience against the mean empirical performance at that point.

Note that we smoothed the performance curves with five-year floating windows in order to dampen noise. We confirmed that smoothing the data did not affect the pattern of results: R^2s based on non-smoothed data were lower, as expected, but the ELSD model still fit significantly better than all other models ($R^2 = .66$), and provided a similar estimate for when the underlying learning rate began to change (18.2 years old).

We also ran the models using percent correct as the measure of accuracy rather than log-odds, both with and without smoothed performance curves. In both cases, the ELSD model fit significantly better than the others and produced similar estimates for the age at which learning rate begins to decline (17.9 years and 18.1 years, respectively).

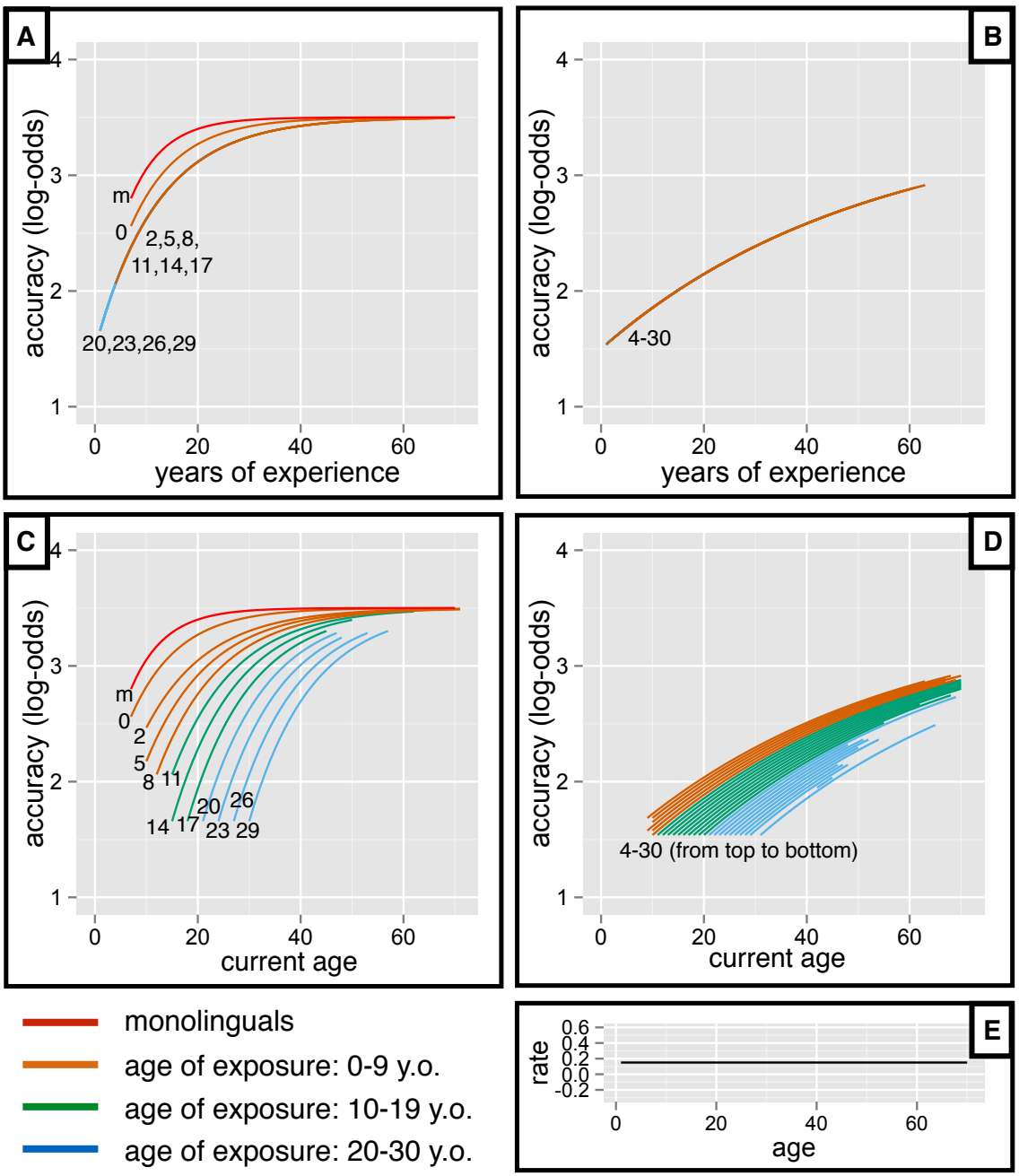


Figure S4. The best-fitting flat-rate model (a, with r = .15; E = 1.00, .72, .55, .13 for monolinguals, simultaneous bilinguals, later immersion learners, and non-immersion learners, respectively), R^2 = .66. **A-B:** Predicted performance as a function of years of experience for monolingual and immersion learners (A) and non-immersion learners (B). **C-D**: Predicted performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). **E**: Estimated learning rate as a function of age.

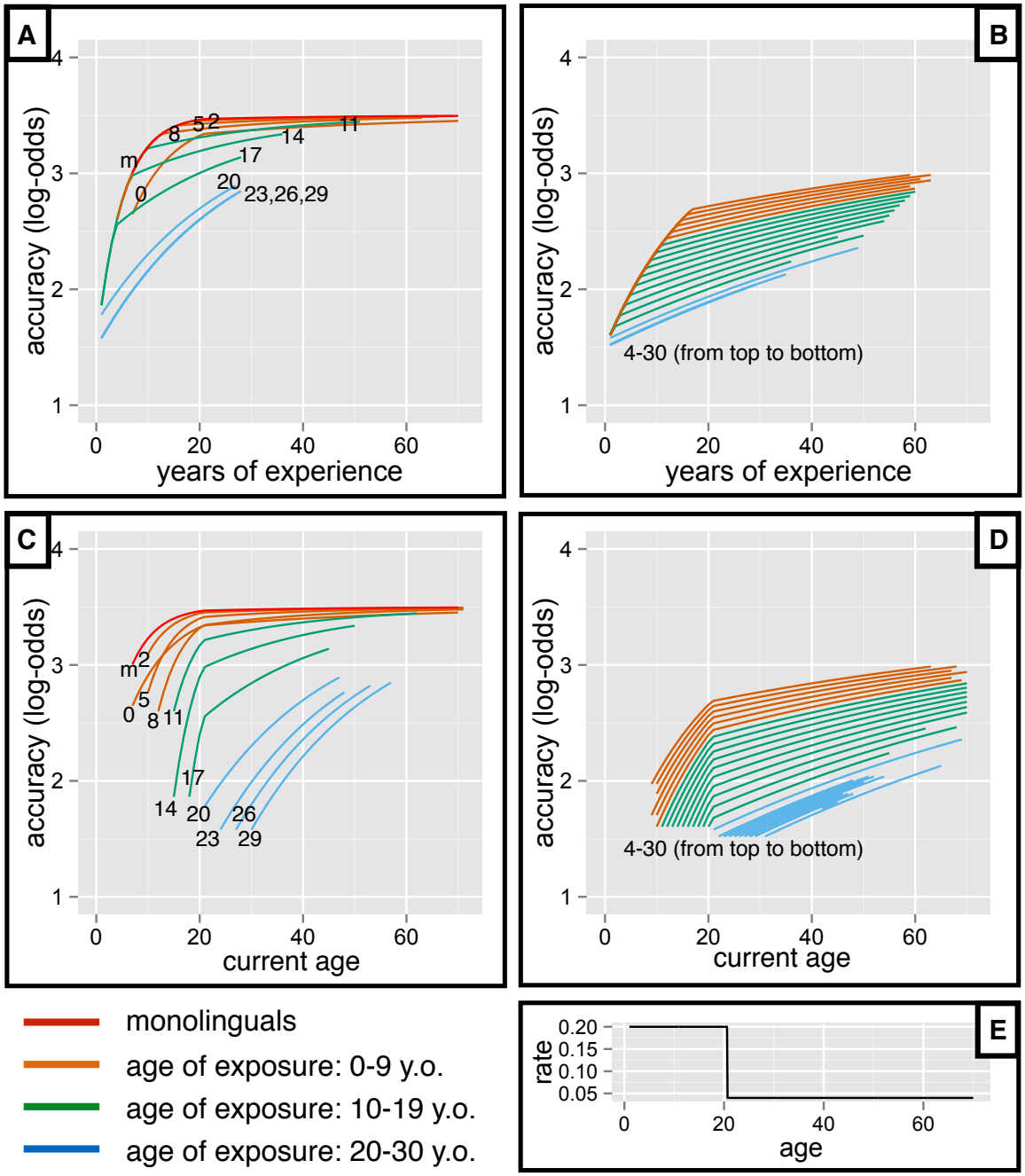


Figure S5. The best-fitting discontinuous rate change model (t_c = 20.7, r_0 = .20, r_1 = .04; E = 1.00, .61, 1.00, .27 for monolinguals, simultaneous bilinguals, later immersion learners, and non-immersion learners, respectively). R^2 = .86. **A-B:** Predicted performance as a function of years of experience for monolingual and immersion learners (A) and non-immersion learners (B). **C-D**: Predicted performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). **E**: Estimated learning rate as a function of age.

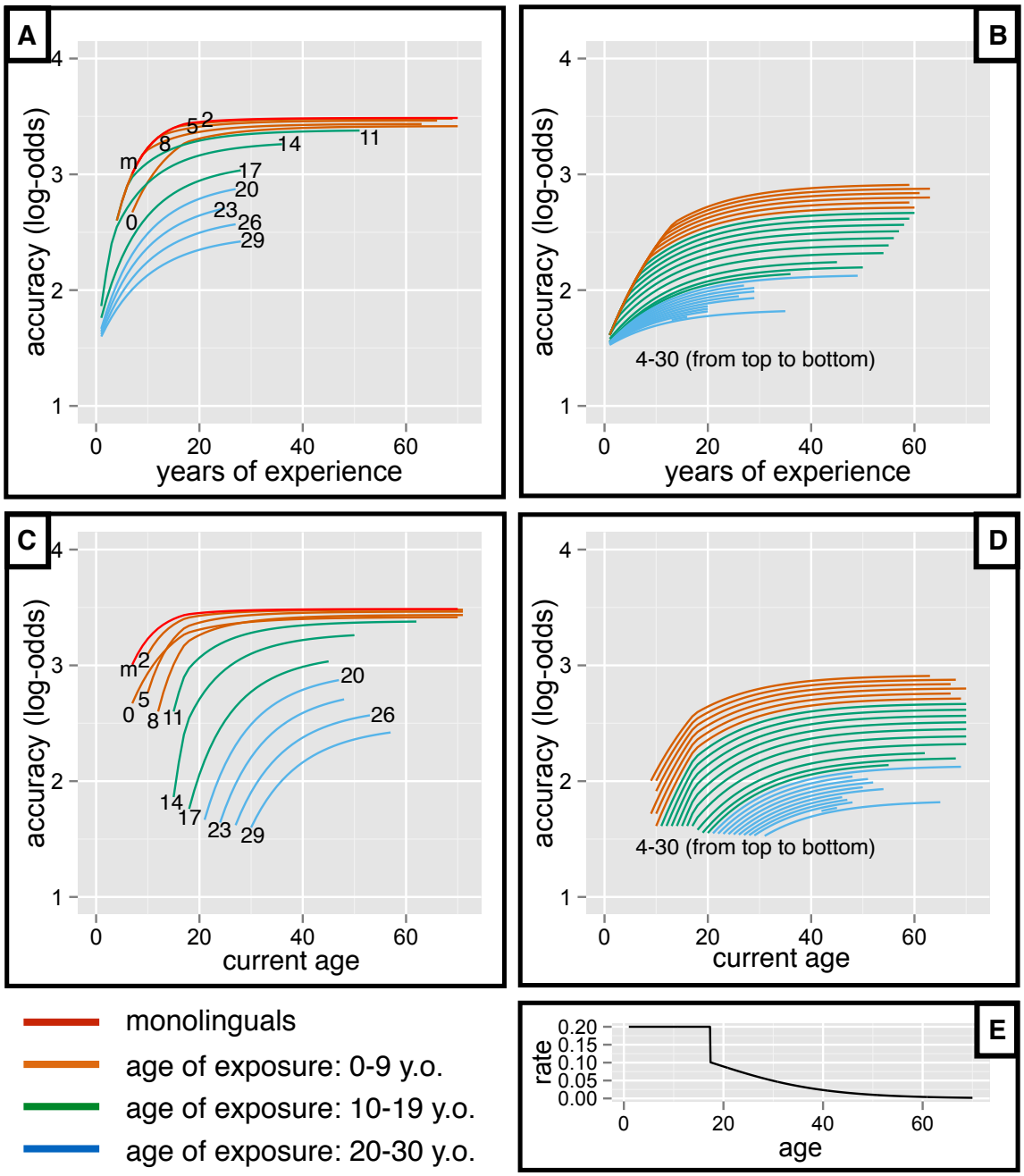


Figure S6. The best-fitting ELSD model (t_c = 17.4, r = .20, α = .09, δ = .18; E = 1.00, .63, 1.00, .29 for monolinguals, simultaneous bilinguals, later immersion learners, and non-immersion learners, respectively). R^2 = .89. **A-B:** Predicted performance as a function of years of experience for monolingual and immersion learners (A) and non-immersion learners (B). **C-D**: Predicted performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). **E**: Estimated learning rate as a function of age.

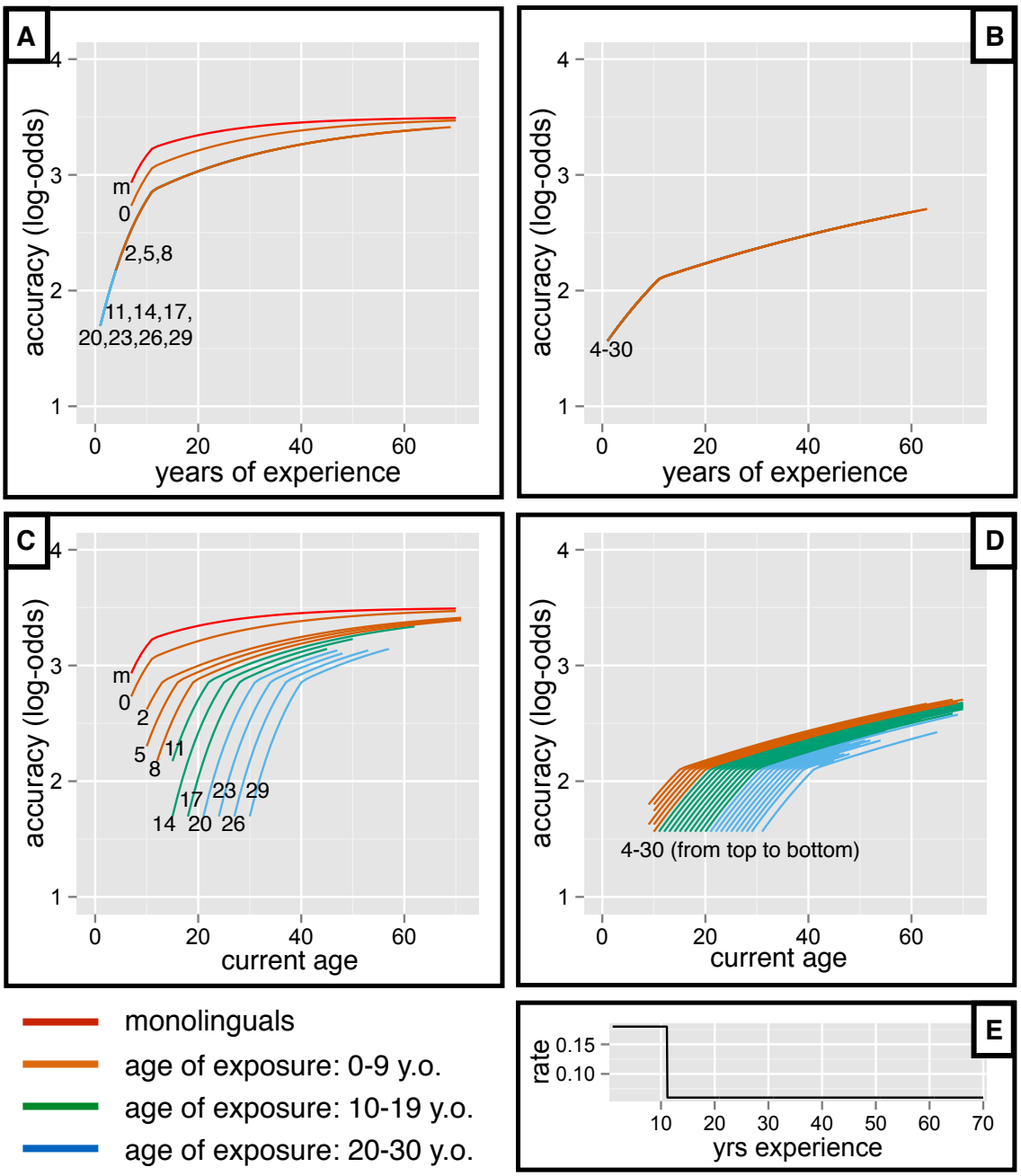


Figure S7. The best-fitting discontinuous rate change model where the discontinuity happens after a set number of years of experience rather than at a set age (t_c = 11.2, r_0 = .18, r_1 = .06; E = 1.00, .76, .57, .18 for monolinguals, simultaneous bilinguals, later immersion learners, and non-immersion learners, respectively). R^2 = .71. **A-B:** Predicted performance as a function of years of experience for monolingual and immersion learners (A) and non-immersion learners (B). **C-D**: Predicted performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). **E**: Estimated learning rate as a function of age.

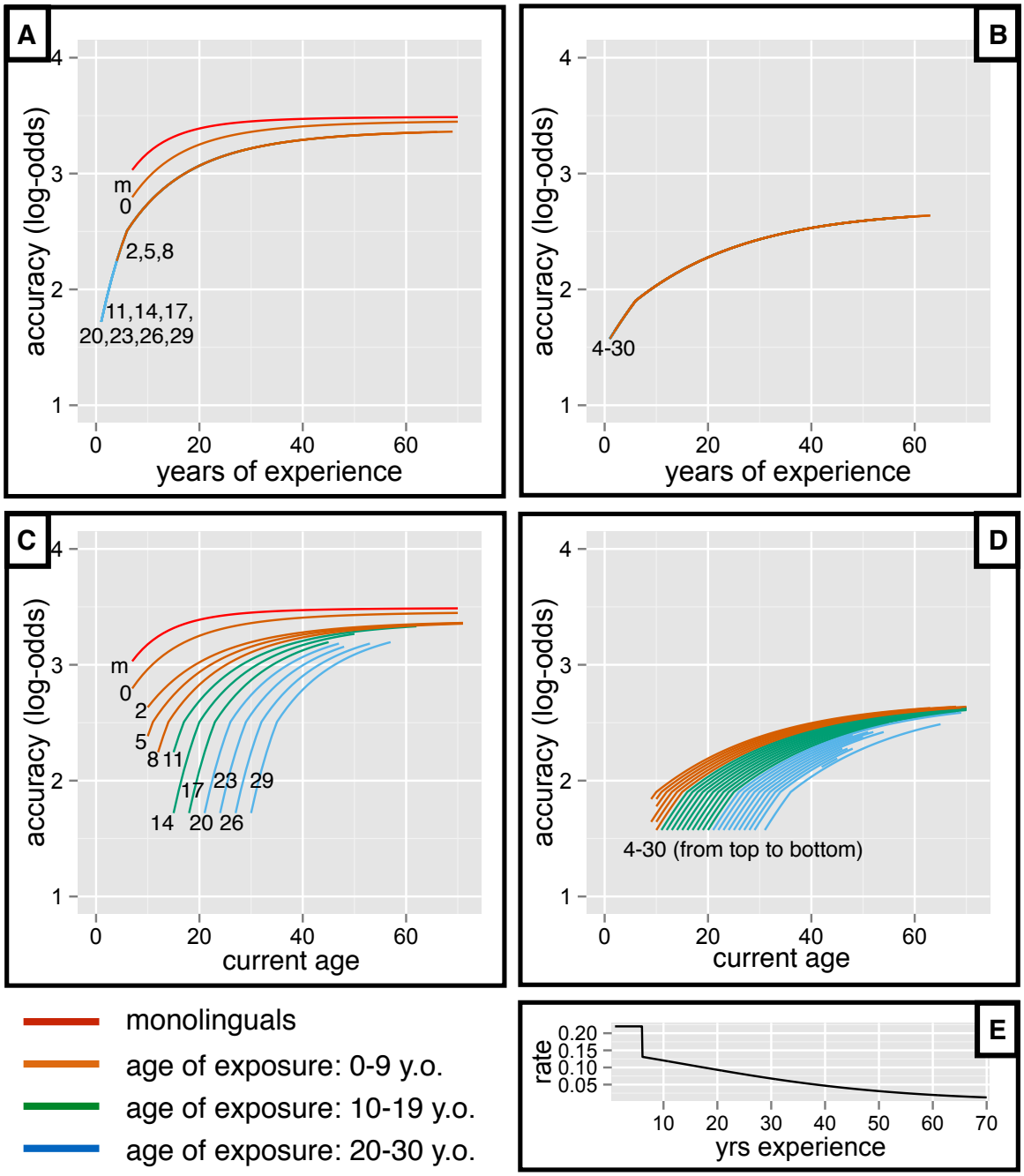


Figure S8. The best-fitting ELSD variant where the discontinuity happens after a set number of years of experience rather than at a set age (t_c = 6.0, r = .22, α = .05, δ = 7.8; E = 1.00, .72, .53, .17 for monolinguals, simultaneous bilinguals, later immersion learners, and non-immersion learners, respectively). R^2 = .70. **A-B:** Predicted performance as a function of years of experience for monolingual and immersion learners (A) and non-immersion learners (B). **C-D**: Predicted performance as a function of age for monolinguals and immersion learners (C) and non-immersion learners (D). **E**: Estimated learning rate as a function of age.

Ultimate attainment.

In order to improve readability of Figure 6, means and standard errors were calculated using a three-year floating window, and curves were restricted to consecutive windows with more than ten subjects. However, data from all subjects with at least 30 years of experience and no more than 70 years of age were included in analyses. This resulted in 107,125 monolinguals available for analysis. Numbers for immersion learners and non-immersion learners are given in the main text.

We identified significant changes in the slope of the ultimate attainment curve using multivariate adaptive regression splines (MARS) (Friedman, 1991) as implemented in the earth package for R (Milborow, 2014). MARS successively breaks linear regression lines into multiple segments, each with its own slope. It then prunes breakpoints that do little to improve fit. To further avoid overfitting, we used 50-fold cross-validation.

In order to ensure that the results were robust to the method of breakpoint estimation, we also considered two other methods that identify breakpoints. The second method (*segmented*) is an iterative search algorithm that finds optimal placement for a specified number of breakpoints (Muggeo, 2014). The locations of breakpoints chosen by this algorithm often depend on the initial first guess as to the location of the breakpoints (the "seed"), which is set by the researcher. We used several procedures to minimize effects of the seed and increase the chances of finding the optimal placement of the breakpoints. For each number of breakpoints, we ran the algorithm with three different sets of starting seeds, choosing the best-fitting result. Moreover, we employed a bootstrap restarting procedure with 25 randomly jittered samples. We fit segmented models with 0, 1, 2, 3, and

4 break points and chose the best-fitting model based on the Bayesian Information Criterion.

The third method (*optimal breakpoint placement*) was a procedure recommended by Vanhove (2013), generalized to multiple breakpoints (in developing this generalization, we are indebted to code written by David Hitchcock of the University of South Carolina, posted at http://www.stat.sc.edu/~hitchcock/raw_piecewise_Rexample705.txt). We considered every possible combination of 0, 1, 2, or 3 breakpoints, with the restriction that the breakpoint must be placed on a whole number of years. For a specific number of breakpoints (e.g., 3), we chose the model with the smallest deviation. We then chose from among the resulting models using the Bayesian Information Criterion. Results for all three sets of analyses are shown in Table S1. Ultimate attainment began to decline rapidly for immersion learners at age of exposure of about 12 in all three analyses and for non-immersion learners at around 9, similar to the estimates obtained by MARS. Of the six analyses, five showed no evidence of a slowing of the decline; the sixth (the optimal analysis applied to the non-immersion learners) showed evidence of a slowing—though still ongoing—decline after an exposure age of 19.

Table S1. *Estimated breakpoints.*

Learners	Method	Breakpoints	Segment Slopes
Immersion	MARS	12	-.009, -.06
	Segmented	11.4 (2.4)	-.007, -.04
	Optimal	12	-.007, -.04
Non-Immersion	MARS	9	+.01, -.06
	Segmented	10.5 (0.2)	-.005, -.07
	Optimal	9, 12, 19	+.01, -.05, -.10, -.02

Note that in all cases the resulting breakpoints are statistically significant in the sense that they result in optimal fits as judged by cross-validation or the Bayesian Information Criterion. However, placing confidence intervals on the placement of the breakpoints or the slopes of the resulting segments is non-trivial and remains an area of active research (segmented provides confidence intervals, but these are likely overfitted, since they assume the number of breakpoints is known). For this reason, in the main text we focus on effect sizes, as measured in terms of the standard deviation of scores by simultaneous bilinguals (immersion learners with an age first exposure of 0).

Permutation analyses. Performance curves (proficiency as a function of years of experience) were plotted for non-immersion learners at each age of exposure from 4 to 30, and also for monolinguals (Figure S3 A-B). Each performance curve was restricted to consecutive ages for which there were at least ten participants in the five-year window, leaving 244,840 monolinguals, 44,412 immersion learners, and 257,998 non-immersion learners. The other details of this analysis are in the main text.

Simulations of Prior Ultimate Attainment Studies

In each simulation, we sampled monolinguals and immersion bilinguals with replacement from our own data. All subjects were required to have at least 30 years of experience with English, and immigrants were required to have minimal exposure to English prior to immigration (following the same definition for "immersion learners" used elsewhere in this paper). For simulating mid-sized studies (N = 275), there were no other restrictions. For simulating large studies (N = 11,371), we matched the number of both monolinguals and immersion bilinguals as in our dataset (given the just-mentioned constraints). For simulating small studies (N = 69), we matched the demographics of

Johnson and Newport's 69 subjects as closely as possible except where it conflicted with the just-mentioned constraints (30 years of experience, limited non-immersion exposure). That is, we sampled 46 non-native speakers and 23 native speakers from our data with replacement; the non-native English speakers spoke a Chinese language or Korean and had immigrated to the United States, and; subjects matched the age of immigration as reported in their paper (see their Table 1). Johnson and Newport provide no demographic information about the native speaking controls. Thus, we selected the 23 native speakers who were between 17 and 22 years old, inclusive, on the assumption that Johnson and Newport's native speakers were undergraduates.

Effect of Analysis Decisions.

Many prior studies have included immigrants who had significant amounts of education in the target language prior to immigration (Hakuta et al., 2003; Johnson & Newport, 1989). This raises an issue: should researchers date the onset of learning to the age at first exposure or the age at immigration (Johnson & Newport, 1989)? Either option introduces imprecision: the first treats immersion and non-immersion learning equally, whereas the latter assumes non-immersion learning is completely ineffective. Thus, either option introduces both noise (since pre-immigration exposure varies between subjects and between studies) and bias (since older immigrants typically have more pre-immigration exposure; see Johnson & Newport, 1989; Flege et al., 1999). Following DeKeyser et al. (2010), we sidestepped this problem by excluding immigrants who had substantial pre-immigration exposure to English.

Similarly, ultimate attainment analyses compare learners who have at least X years of experience. In the absence of solid data, prior researchers employed a variety of cut-offs

(Fig. 5A). Our larger sample allowed us to identify the appropriate cutoff directly: about 30 years (at least for our data). This suggests that the smaller cut-offs in previous studies disadvantaged later-learners, who (for obvious reasons) tend to have less experience and thus are farther from asymptote. Moreover, different studies used different cut-offs, which can introduce variability in ultimate attainment curves (Fig. S9).

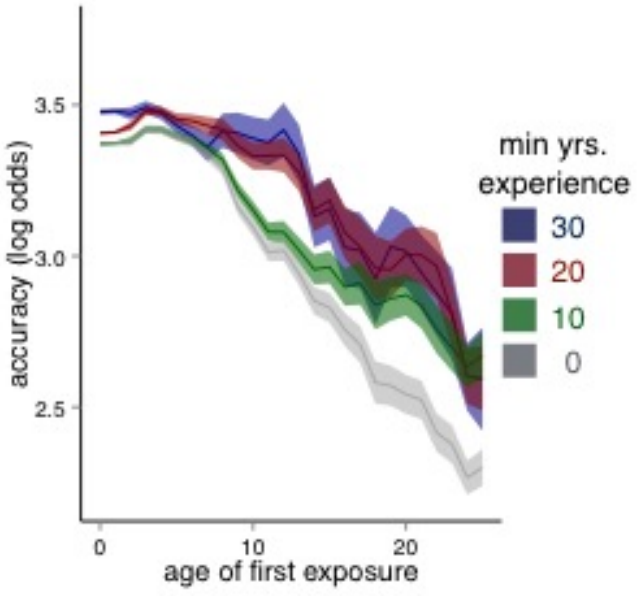


Figure S9. Ultimate attainment curves revealed by our data, using different cut-offs for minimum years of experience.

However, while these analytic decisions can have a significant effect for studies like ours, which have large datasets (Fig. S9), they appear to have little effect on typically-sized studies. We concluded this based on simulating Johnson & Newport's (1989) study with or without the analytic decisions above. That is, in the simulations discussed in the main text, we simulated running Johnson & Newport's study with the same number of subjects hailing from the same countries and arriving at the same ages, but otherwise using our own analytic decisions (limited pre-immigration exposure, and at least 30 years of experience

with English). We ran a second set of simulations in which we matched Johnson & Newport's analytic decisions, matching their subjects' numbers of years in America and number of years of pre-immigration exposure (as reported in their paper). As can be seen in Fig. S10, the range of results was similar, suggesting that these analyses decisions were a fairly minor contributor to the differences across prior studies.

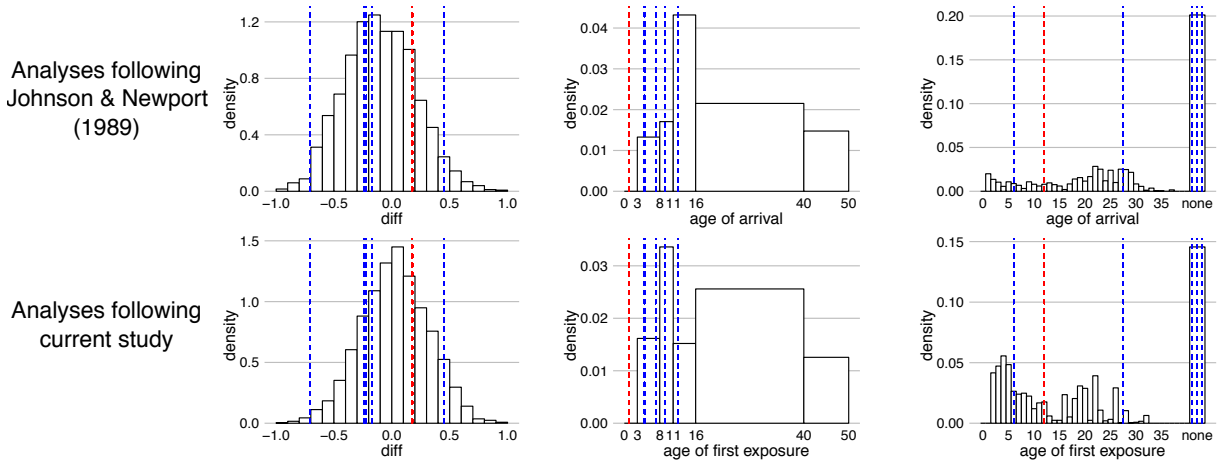


Figure S10. We conducted two sets of 2,500 simulated experiments. In the first set (top row), we drew 69 subjects following the demographics of Johnson & Newport (1989) as closely as possible (including native language, pre-immigration English exposure, and age at test). Likewise, we followed them in conducting analyses in terms of age of arrival rather than age of first exposure. We modified these methods in the second set (bottom row) to match what we used in our own analyses by restricting immigrants to minimal pre-immigration exposure, following the definition used in the main text. Note that these simulations are also reported in Fig. 8 (top row). As in Fig. 8, three analyses were considered. From left to right: correlation between onset age and ultimate attainment prior to 16 years old. minus after 16 years old; first subgroup of subjects to be significantly worse than monolinguals in a t-test; onset age at which performance begins to decline more rapidly, if any. **Blue**: estimates from prior studies. **Red**: estimates from current study. Note that the y-axis varies across panels.

Item Effects.

It is plausible that critical period effects might differ for those aspects of grammar that are typically mastered early in first-language acquisition as opposed to those that are

mastered late. Our data, however, provided little support for this hypothesis: The best-fitting models of learning indicated that learning rate began to slow at approximately the same time for the 47 items that are mastered by the youngest monolingual English-speakers in our sample (ages 7-8) as for the 48 items that are mastered only by the older ones: 17.3 years old and 18.2 years old, respectively.[1]

Likewise, ultimate attainment analysis and permutation analysis of the performance curves both supported this finding, with one complication: Very few immersion learners missed *any* of the early-mastered items. Thus, while breakpoint analyses of ultimate attainment curves (Figure S11) using MARS found that attainment began to drop steeply for non-immersion learners at ages of exposure of 12 years for early-mastered items (B = -.12) and at 9 years for late-mastered items (B = -.07), and for immersion learners at 9 years for late-mastered items (B = -.04), immersion learners remained near the ceiling on early-mastered items regardless of their age of first exposure: Even immersion learners who began learning English at 25 years old missed, on average, fewer than one of the early-mastered items. Thus, although MARS was able to identify a breakpoint (3 years old), performance declined only negligibly after that age (B = -.01). (There is an apparent decline

[1] Items were considered mastered if they were answered correctly by at least 22 of the 23 monolingual English-speakers ages 7-8 (we combined the two youngest age categories in order to achieve sensible N). While this was a somewhat arbitrary choice, it was the only one we considered, mitigating somewhat concerns about *post hoc* analyses. As in the main analyses, the best-fitting models involved sigmoidal rate change (R^2 = .85 for early-mastered items and R^2 = .87 for late-mastered items). Early-mastered items: t_c = 17.3, r = .27, α = .07, δ = -4.1; E = 1.00, .42, 1.00, .22. Late-mastered items: t_c = 18.2, r = .17, α = .10, δ = 2.3; E = 1.00, .65, .95, .33. Recall also that the natural range of the model is (0, 1), and thus we used scale parameters a and b (2 and 1.5, respectively) to map the model's range onto the empirically observed range of scores [in the case of the primary analyses, approx. (1.5, 3.5); cf. Figure S3]. Thus, parameter b had to be adjusted to 2.25 and 1 for the early-mastered and late-mastered items, respectively, each of which has a different empirical range [approx. (2.25, 4.25) for early-mastered items and (1, 3) for late-mastered items].

starting at around 19 years old, but it is not statistically significant.)

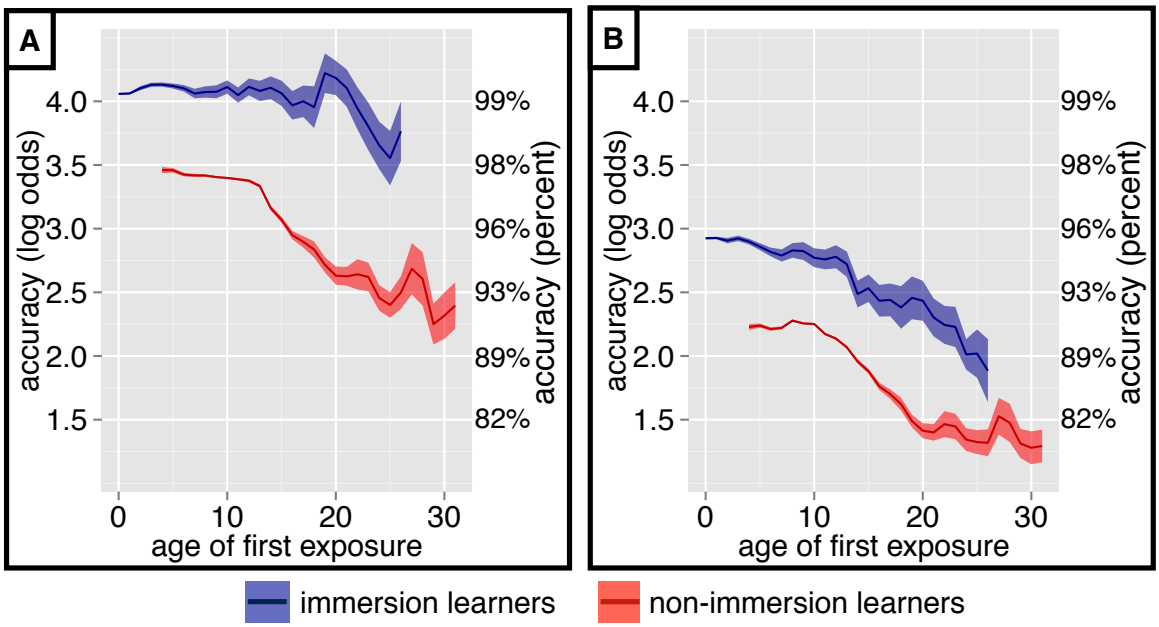


Figure S11. Ultimate attainment for early-mastered items (A) and late-mastered items (B), smoothed for presentation with a three-year floating window. Shadows represent +/- 1 SE. As in Figure 2, data were smoothed by a three-year floating window, and only consecutive windows with more than 10 subjects shown.

Permutation analysis of the learning curves (Figure S12) reveals similar results. Performance curves are reliably shallower for non-immersion learners by an age of first exposure of 10 years for early-mastered items (p = .009) and by 12 years for late-mastered items (p = .03).[2] Immersion learners show significantly shallower performance curves by an exposure age of 11 years for late-mastered items (p = .002).[3] However, for early-

[2] On early-mastered items, there is a significant effect at 7 years old (p = .03), but this disappears for 8 years old (p = .09) and 9 years old (p = .36) and so is likely due to noise; in contrast, every performance curve from 10 years old on shows a significant effect. Note that throughout, these analyses are not corrected for multiple comparisons.

[3] In fact, there is a significant difference at 2 years old (p = .04), but not at 5 (p = .89) or 8 (p = .28), so this is again probably noise. Note again that these analyses are not corrected for multiple comparisons.

mastered items, immersion learners do not show significantly shallower performance curves until an age of first exposure of 23 years. Again, this final result may be influenced by ceiling effects: by 10 years of experience, nearly all these performance curves are above 3.4 (equivalent to making a single error).

We also considered whether there was consistency in item difficulty for different types of learners. In order to avoid ceiling effects, we considered learners with 7-10 years of experience, thus capturing the earliest stage of learning for which we have data for every learner type. Within this group, we compared monolinguals (N = 82), simultaneous bilinguals (N = 35), and immersion learners with exposure ages 1-5 (N = 77), 6-10 (N = 314), 11-15 (N = 287), and 16-20 (N = 82).

For every bilingual group, by-item performance was highly correlated with that of monolinguals: simultaneous bilinguals (r = .75, *logBF* = 35.8, p<.0001), immersion learners with exposure ages 1-5 (r = .81, *logBF* = 47.4, p<.0001), immersion learners with exposure ages 6-10 (r = .81, *logBF* = 46.1, p<.0001), immersion learners with exposure ages 11-15 (r = .77, *logBF* = 39.2, p<.0001), and immersion learners with exposure ages 16-20 (r = .73, *logBF = 37.4*, p<.0001), where *logBF* is the log of the Bayes Factor $p(H_1|d)/(p(H_0|d)$ (see Wagenmakers, 2007). Importantly, the correlation for the latest immersion learners (r = .73) was almost identical to that for the earliest immersion learners (r = .75).

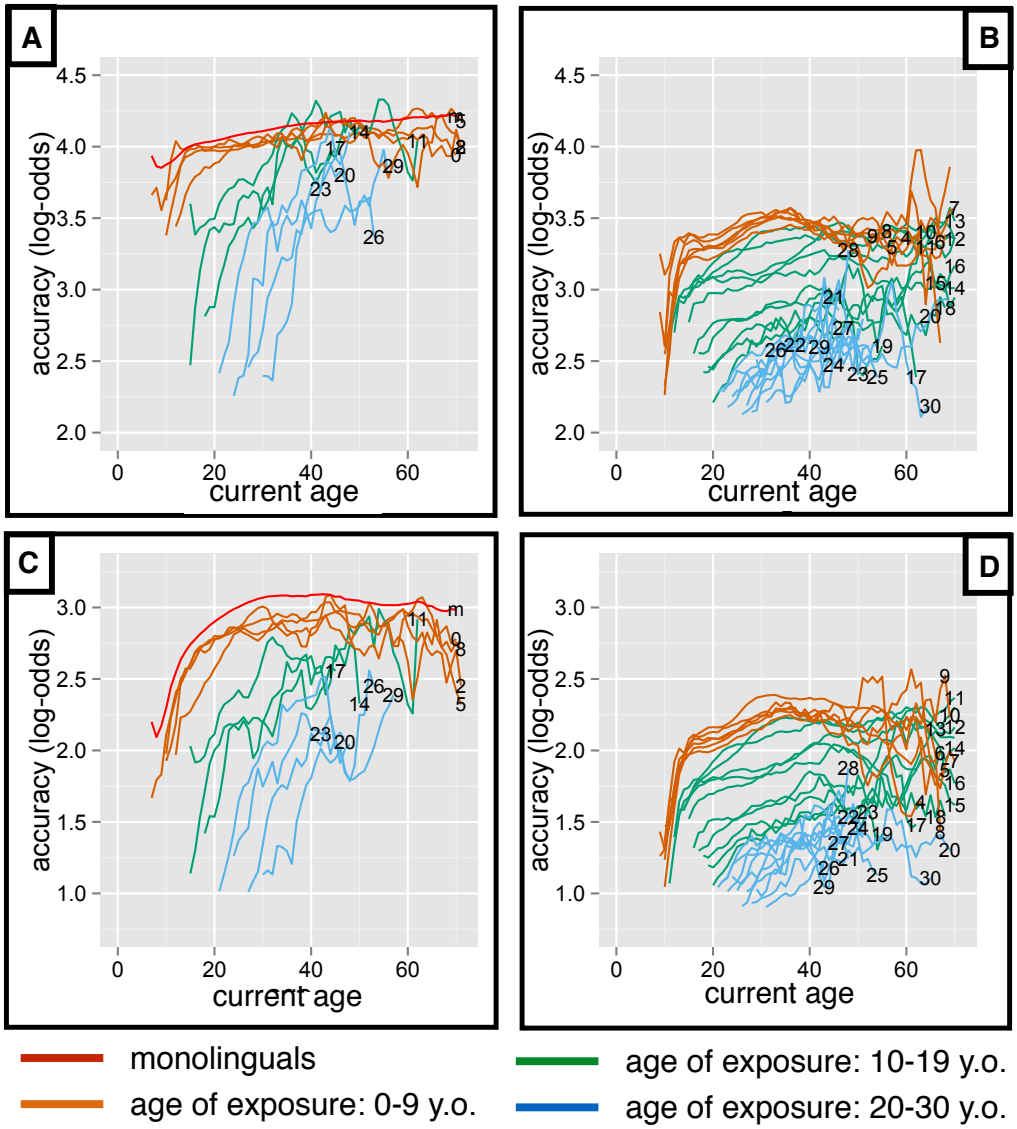


Figure S12. Performance curves for early-learned items are show for immersion learners in (A) and non-immersion learners in (B). Performance curves for late-learned items are shown for immersion learners in (C) and non-immersion learners in (D). Note that the y-axis scale is different for the top two panels vs. the bottom two.

L1 Effects

In this section, we assess evidence that certain aspects of learning of English are significantly different for one of the language groups (Chinese, Western Germanic, etc.; see main text) relative to the others. We focused on immersion learners, where the range in

outcomes is larger. We considered differences in asymptotic performance (ultimate attainment), the length of the optimal period, and the shape of learning curves.

For each type of analysis, we first conducted a series of "power" simulations to determine our ability to detect effects of various sizes.[4] These simulations serve three purposes: 1) determining the likelihood we could detect meaningful differences in our data, given the sample sizes available; 2) providing some guidance on sample size for researchers who are designing follow-up studies, and; 3) providing some intuition into Bayes Factors for readers unfamiliar with them. For the purposes of (1), it would be ideal to use the actual *N*s from our data and take into account the uneven distribution of subjects across conditions. However, we found that this results in complex, confusing graphs (cf. 3), and was not especially helpful for providing guidance on sample sizes (cf. 2), since it merely shows that in most cases we have limited power, rather than indicating how many subjects would be needed for more power. Thus, we elected to use a range of balanced Ns, which we believe will ultimately be the most useful to the reader.

To compare performance across populations, we used Bayes Factors to compare a model where that language group is treated separately from the others (M_1) against the null model that treats all language groups the same (M_0). Bayes Factors represent how much more likely the data are under one model compared to the other:

[4] Strictly speaking, "statistical power" refers to a construct in null hypothesis significance testing (the probability of rejecting the null hypothesis given a particular effect size and sample size). However, there is a fairly straightforward extension to Bayes Factor Analysis: the probability of the data favoring the alternative hypothesis ($logBF>0$) given a particular effect size and sample size. We hope that the reader will forgive this abuse of terminology, since it results in much simpler prose.

$$BF_{10} = \frac{P(D|M_1)}{P(D|M_0)}$$

where the subscripts of BF denote which model is the numerator:

$$BF_{01} = \frac{1}{BF_{10}} = \frac{P(D|M_0)}{P(D|M_1)}$$

Thus $BF_{10} = 3$ means that the data are three times more likely under M_1 relative to M_0. We use the natural logarithm of the Bayes Factor because it makes 0 clearly interpretable:

$\log(BF_{10}) < 0$	more evidence for M_0
$\log(BF_{10}) = 0$	equal evidence for M_1 and M_0
$\log(BF_{10}) > 0$	more evidence for M_1

Bayes Factors have many advantages over p-values, including the fact that they quantify evidence for the null hypothesis (Wagenmakers, 2007). Most importantly, they are guaranteed to select the correct model as N increases to infinity. P-values, in contrast, have a fixed Type I error rate of 0.05. However, Bayes Factors can be complex to calculate. Throughout, we use the BIC approximation to the Bayes Factor (Wagenmakers, 2007). At small sample sizes, this method will tend to favor the null (M_0) more than do other, less tractable alternatives (as sample sizes increase, the difference disappears). However, this matches psychologists' commonly stated preference for favoring the null hypothesis. In any case, the alternatives proved intractable with datasets as large as ours.

Ideally, we would treat both subjects and items as random factors (Baayen, Davidson, & Bates, 2008; Clark, 1973). Unfortunately, it is unclear how to calculate BIC for such models. While some approximations have been suggested, we generally found that this gave us unreliable results for unbalanced designs. Thus, we use fixed effects models throughout.

Asymptotic Performance. We asked whether asymptotic performance (defined as performance by individuals with at least 30 years of experience and no more than 70 years old) differs reliably depending on first language. We first present power analyses assessing our ability to detect meaningful differences, followed by our actual results.

We conducted power analyses through simulation. To simulate the case where there is no difference between groups, we drew two groups of N subjects from the asymptotic simultaneous bilingual data, sampling with replacement. To simulate differences between groups of size *e* (measured as difference in log-odds), we first fit a binomial mixed effects model to the asymptotic simultaneous bilingual data using lme4 in R. We then used predict.merMod from the lme4 package to generate data for two groups of N subjects: one with the original intercept and one with an intercept that differed by *e* from the original. This method allowed us to take into account correlations between items and within subjects. We then recalculated the log-odds of a correct answer for each subject (again, using the empirical logit function) and compared two fixed-effect linear regression models: one with a fixed effect of subject group, and one without.

For every level of N, we conducted 200 simulations for $e = 0$ and 100 simulations for all other *e*s. We included both positive and negative *e*s in order to account for any ceiling effects (large positive *e*s were not possible because of ceiling). The results of our simulations are shown in Fig. S13.

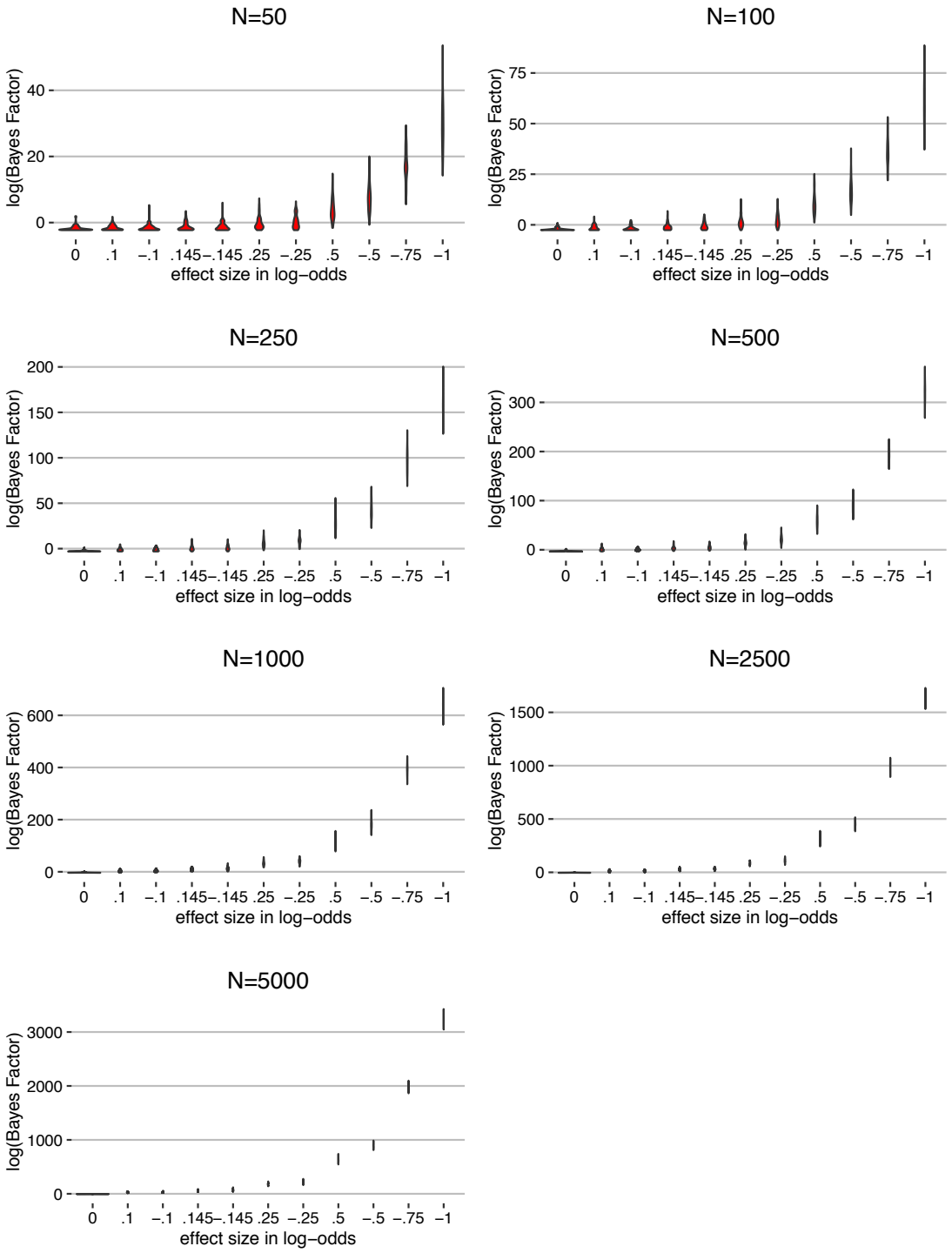


Fig S13. Expected Bayes Factors based on a range of Ns *per condition* (panels) and effect sizes (x-axis). For reference, $e = 0.145$ is the difference between asymptotic simultaneous bilinguals and monolinguals. Note that the y-axis scale varies across panels.

As can be seen, with small N, Bayes Factors favor the null. However, as N increases, the Bayes Factor is extremely likely to favor the correct model. With only 50 subjects per condition, only fairly large effects on the order of 0.5 can be reliably detected. For comparison, 0.5 is approximately half the difference between asymptotic monolinguals and our youngest monolinguals (7 years old). Reliably detecting the difference between asymptotic monolinguals and bilinguals (e = 0.145) requires around 500 subjects per condition.

For comparison, we have included standard power analyses based on p-values in Table S2 (we used binomial mixed effects regression with maximal random slopes). The comparison of Fig. S13 and Table S2 nicely demonstrates the fact that using p-values makes one more likely to reject the null hypothesis—not only when the null is false but also when it is true (Wagenmakers, 2007). Note that when e = 0, the probability of the Bayes Factor erroneously supporting the alternative hypothesis is negligible, even for small N. In contrast, when using p-values, Type I error remains a constant 0.05 independent of N (as by definition).

Table S2. *Power analyses for studying asymptotic behavior, using p-values, for various effect sizes.*

N / condition	0.1	-0.1	0.145	-0.145	0.25	-0.25	0.5	-0.5	-0.75	-1
50	.22	.31	.24	.32	.44	.49	.78	.92	1.0	1.0
100	.38	.33	.36	.53	.63	.73	.98	1.0	1.0	1.0
250	.49	.45	.62	.67	.91	.96	1.0	1.0	1.0	1.0
500	.68	.66	.86	.90	.99	.99	1.0	1.0	1.0	1.0
1000	.84	.84	.97	.97	1.0	1.0	1.0	1.0	1.0	1.0
2500	.99	.97	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

To analyze our data, we divvied up immersion learners into groups by age of first exposure (AoFE): 0, 1-5, 6-10, 11-15, and 16-20. We then analyzed results for any language group with at least 50 subjects in a particular AoFE bin. As described in the main text, for each language group, we asked whether the data would be better fit by assuming a distinct mean for that language group as opposed to all other subjects. As can be seen in Fig. S14, results were highly similar across language groups within a particular AoFE bin. Given this, it is not surprising that Bayes Factor analyses typically supported the null hypothesis (logBF < 0) (Table S3). The exceptions involved superior performance by Romance speakers at AoFE = 0, superior performance by Western Germanic speakers at AoFE 1-5, and superior performance by Chinese speakers at AoFE 6-10. Given the lack of systematicity and the relatively small Bayes Factors, these are most likely spurious. For comparison, we have included traditional p-values for the same analyses, based on binomial mixed effects regression. As expected, these tend to more strongly favor the alternative hypothesis (Wagenmakers, 2007). Note with correction for multiple comparisons, $\alpha = .0037$.

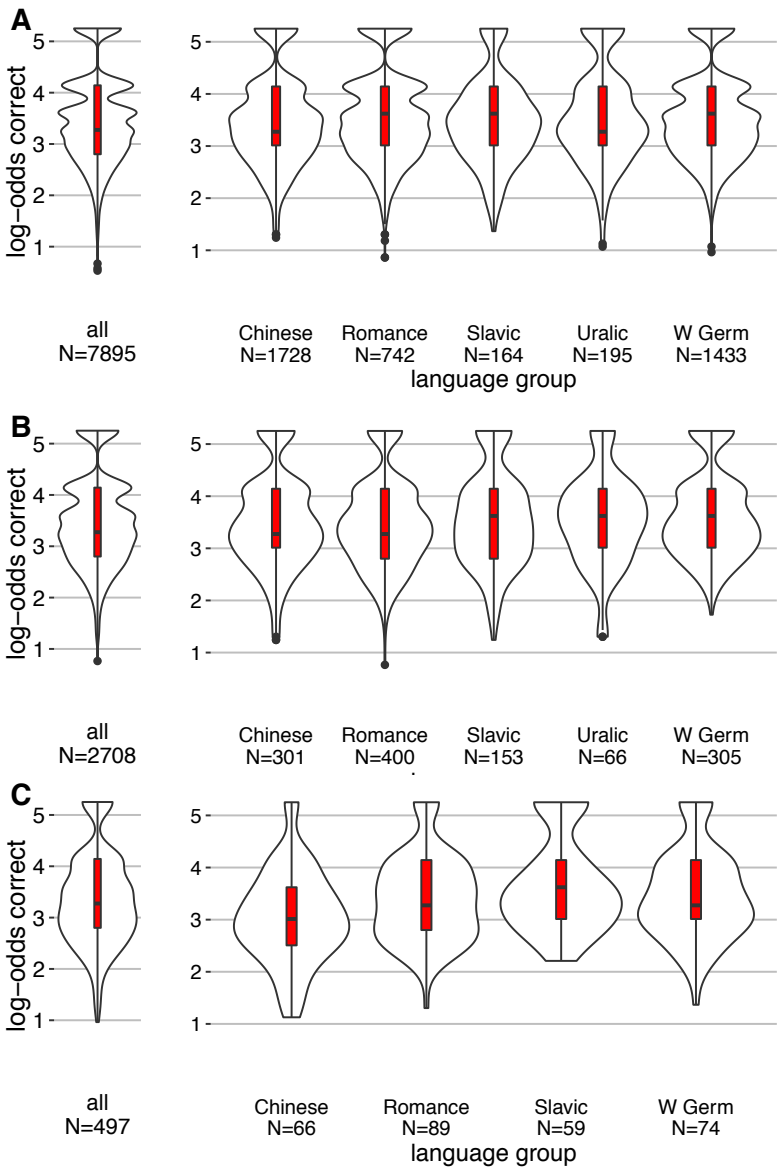


Figure S14. Boxplots (in red) overlaid on violin plots (white) for asymptotic immersion bilinguals overall (left) and for five language families (right). **A**: age of first exposure = 0 (simultaneous bilinguals). **B**: age of first exposure 1-5. **C**: age of first exposure 6-10 (note that for this, there were too few Uralic speakers to include).

Table S3. Evidence that asymptotic performance was significantly different for a language group relative to the rest, using both Bayes Factor and p-value.

AoFE		Chinese	Romance	Slavic	Uralic	West. Germ.
0						
	logBF	-4.4	2.4	-4.5	-3.2	-2.0
	p-val	.295	.000	.254	.450	.000
1-5						
	logBF	-3.7	-3.9	-2.8	-3.6	2.6
	p-val	*.243*	.060	.785	.302	.000
6-10						
	logBF	1.7	-3.0	0.3	NA	-2.5
	p-val	.003	.473	.000	NA	.019

Optimal Period. We first present power analyses assessing our ability to detect meaningful differences, followed by our actual results.

For power analysis, we conducted a series of simulations to determine our ability to detect differences of various sizes. We first fit a segmented regression model with a single breakpoint to the full ultimate attainment dataset using the *segmented* package in R, as described above ("Ultimate Attainment"). For a given sample size (as explained below), we then generated two datasets: one using the original model and one with the breakpoint shifted by *e* years.[5]

We considered nine sample sizes: Ns = 155, 310, 775, 1550, 3100, 7750, 15500, 31000, and 11371 *per condition*. The first eight sample sizes involved N = x subjects per age of first exposure from 0 to 31, where x = 5, 10, 25, 50, 100, 250, 500, and 1000,

[5] The slope of the first segment was adjusted so that the height of the curve at the breakpoint was kept the same across the two models (as described above, that slope was very small, so this decision had a limited effect). The slope after the breakpoint was left unchanged.

respectively. For the final simulation, for each condition we drew exactly the number of subjects at each age of first exposure that was found in our actual ultimate attainment dataset. Thus, this final simulation better represents the unequal sample sizes in our actual data. We conducted 100 simulations for each level of N and *e*, except for *e* = 0, where we ran 200. Results are shown in Fig. S15.

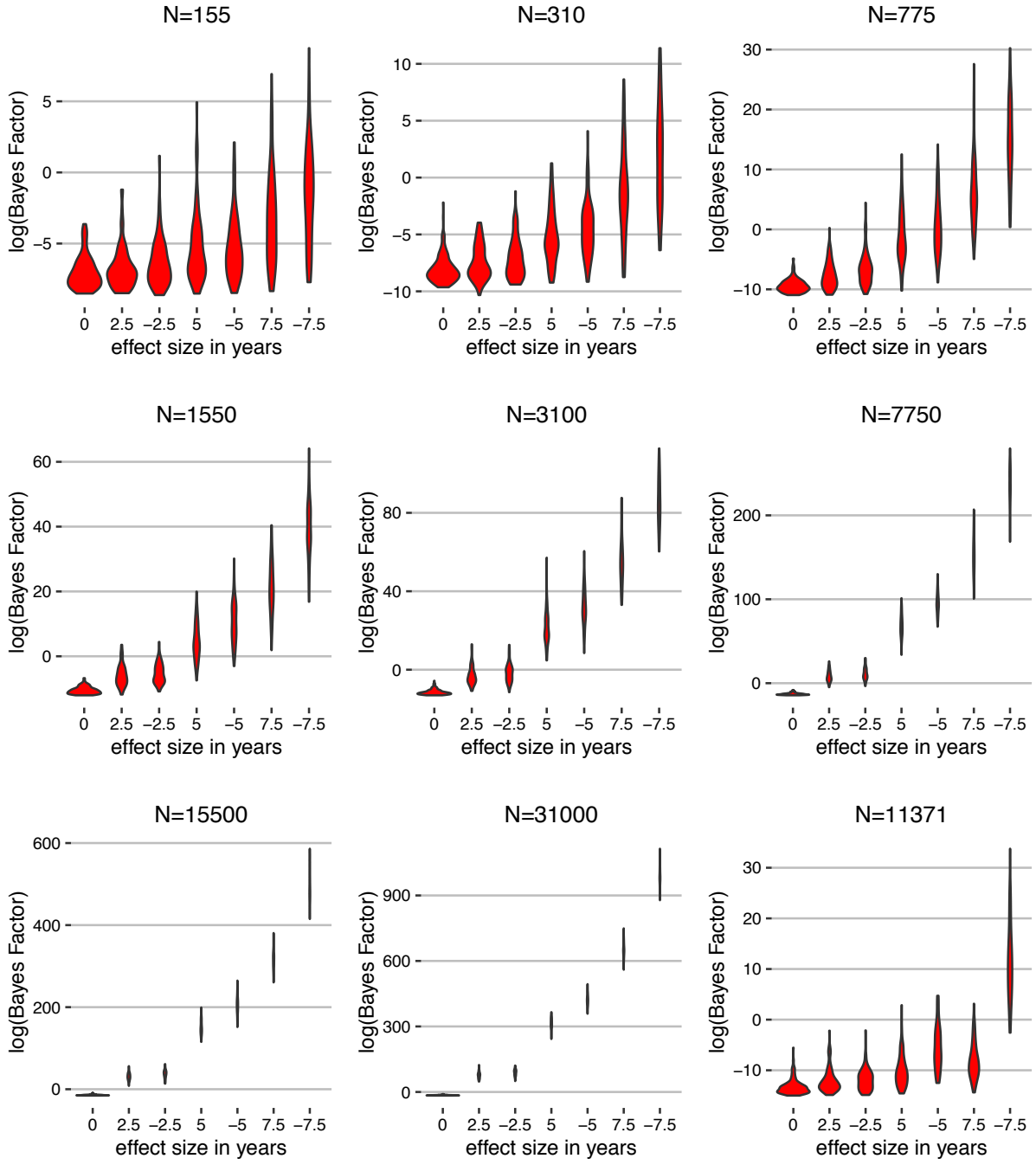


Fig. S15. Violin plots of logged Bayes Factors for 9 different sample sizes (panels) and 7 different effect sizes (x-axis). See text for explanation of how these subjects were distributed across ages of first exposure. Note that the y-axis scale varies across panels.

Thus, based on these simulations, we are unlikely to detect a difference in optimal

period of much less than 7.5 years, which is around 60% of the length of the optimal period measured over all subjects. However, if future researchers find a mechanism for recruiting a subject group more evenly distributed across ages of first exposure, they could get considerably better precision with fewer subjects.

We then analyzed our data. Results did not favor the hypothesis of a separate breakpoint for Chinese (*logBF* = -7.9), Romance (*logBF* = -12.6), Slavic (*logBF* = -9.1), or Turkic (*logBF* = -9.6). The *segmented* package was unable to identify any breakpoint for Uralic or Western Germanic. This seems to be due to large amounts of noise rather than clear evidence against the existence of a breakpoint (see Fig. S16).

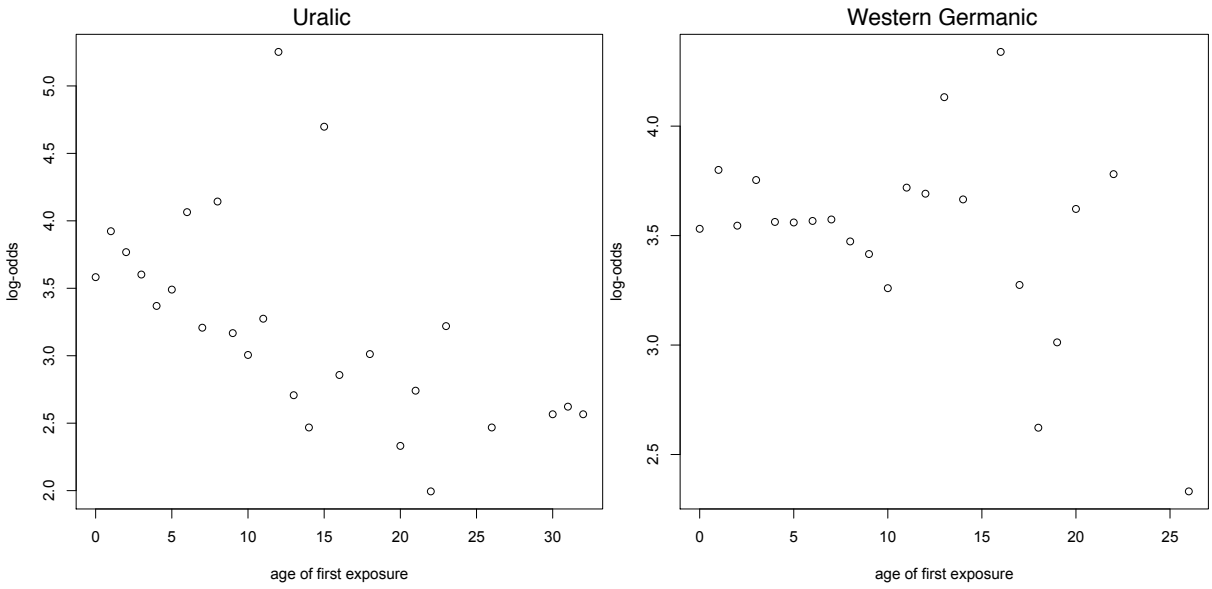


Figure S16. Ultimate attainment as a function of age of first exposure for Uralic (left) and Western Germanic speakers (right). In order to better show the variability, no smoothing was used. Note that the y-axis scale varies across panels.

Learning Curves. We also considered whether first language affected how quickly

English was learned. Ideally, we would measure how long it takes to reach asymptote as a function of first language. However, most parametric curves—including the one we used in our main model—require an infinite amount of time to reach asymptote. We could alternatively measure how long it takes for subjects to get within ε of asymptote for some ε, but then the results may depend on the ε chosen, particularly given that any reasonable ε will be small relative to the amount of statistical noise in the data (we have relatively few subjects at the ages that are most relevant for this analysis).

More information is available if we compare language groups in terms of the shape of their learning curves (the curves relating performance and years of experience). We chose a method that put as few *a priori* constraints on the shape of the learning curve as possible. Specifically, we fit a linear regression on accuracy (in log-odds) with years of experience, language group, and their interaction as *fixed* effects. Thus, no relationship is assumed between performance at age *a* and at age *a+1*. We asked whether this model fit better than one without language group or its interactions. Because our interest is in the learning curve, we restricted analyses to the first thirty years of experience (from analyses above, we know that data after this time is highly similar across the language groups, so including these ages would diminish our ability to detect differences in learning).

We first present power analyses assessing our ability to detect meaningful differences, followed by our actual results.

Our power analyses show that the sensitivity and flexibility of this method comes at a cost: It requires many subjects. This was confirmed in a series of simulations (Fig. S15). In each simulation, we generated two datasets. In the first, we sampled from our simultaneous bilinguals with replacement. For the second dataset, we simulated

completion of learning *e* years earlier by first adjusting "years of experience" in the original data according to the following formula:

$$A = \begin{cases} 7 + ((30 - d) - 7) * \frac{(A-7)}{30-7} & A \leq (30 - d) \\ (30 - d) + (70 - (30 - d)) * \frac{(A-30)}{70-30} & 30 < A \leq 70 \end{cases}$$

This has the effect of "squeezing" the curve prior to 30 years of experience into 30-*e* years of experience (the data after 30 years of experience is "stretched" to compensate). We then sampled the second dataset with replacement.

We considered 9 sample sizes with 5, 10, 15, 20, 25, 50, 75, 100, or 150 subjects per level of years of experience. As Fig. S17 shows, if one group learned 3x faster than the other, we would need around 2,250 subjects *per group* (75 per group per year of experience). To detect "only" a difference of 2x, we need around 4,500 subjects *per group* (150 per group per year of experience). These values are in fact optimistic, in that they assume an even distribution of subjects across years of experience, and thus equal precision at all points in the curve, which is not the case.

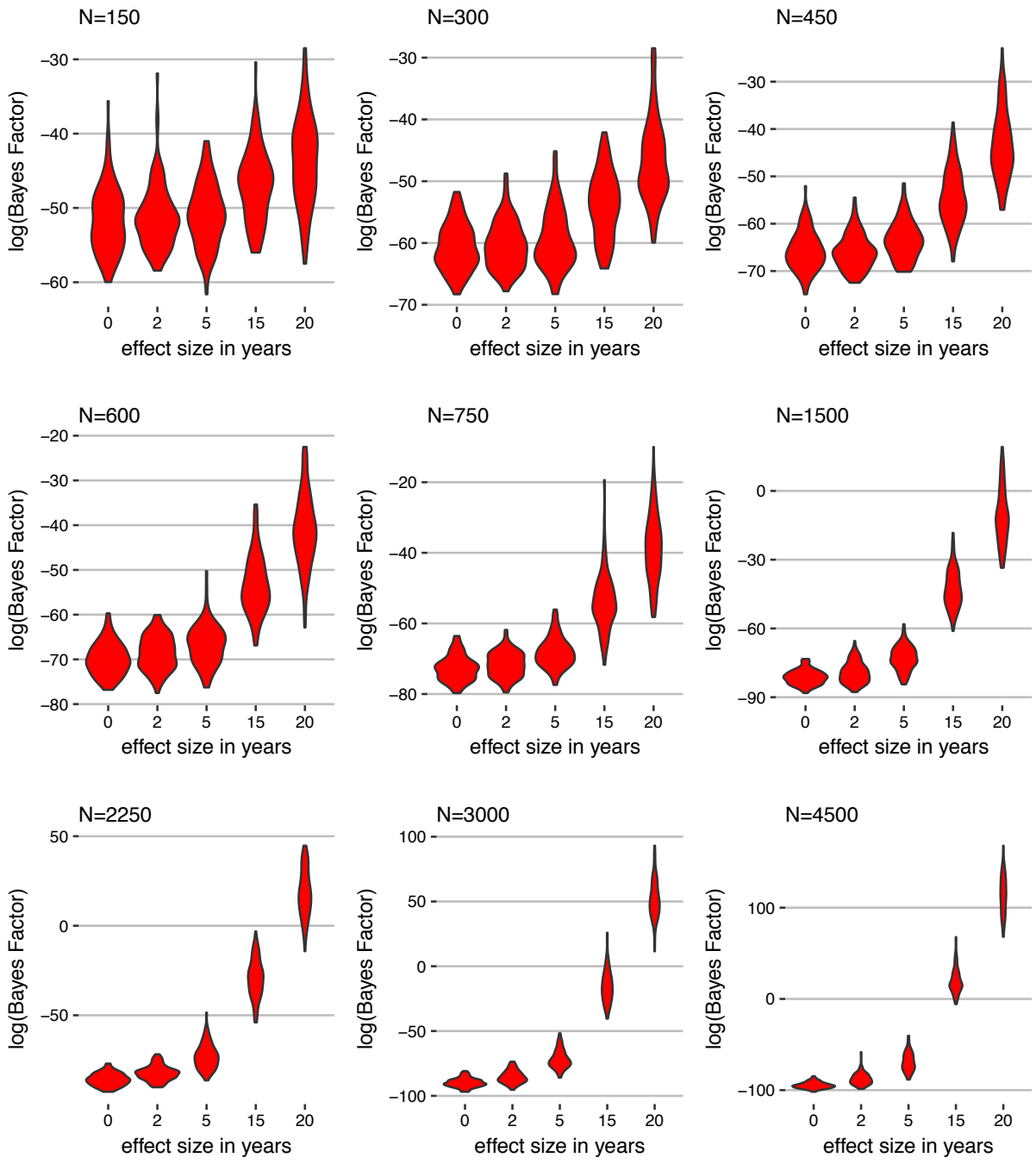


Figure S17. Violin plots for logged Bayes Factors for nine different sample sizes *per condition* (panels) and five different effect sizes (x-axis). The effect size reflects the number of years the learning curve was shifted left (see text). Note that the y-axis scale varies across panels.

Given these simulations, we restricted analyses for subgroups where we had at least 1,500 subjects (approx.) or more. Within simultaneous bilinguals, this included Chinese, Romance, and Western Germanic. Bayes factor analyses suggested no distinction between Romance and the remaining languages (*logBF* = -4.8) or between Western Germanic and the remaining languages (*logBF* = -3.1). There was weak evidence in favor of treating Chinese separately (*logBF* = 2.0). However, the actual difference between learning curves appears to be slight (Fig. S18, left) and may only reflect noise.

Looking at later learners (ages of first exposure from 1-5 years old), only Chinese had sufficiently many subjects. In this case, there was strong evidence for treating Chinese separately from the other groups (*logBF* = 4.4). From inspection of Fig. S18 (right), this seems to reflect somewhat faster initial learning and perhaps an earlier decline. However, as the figure shows, the data are fairly noisy, reflecting the relatively small number of subjects for this type of analysis.

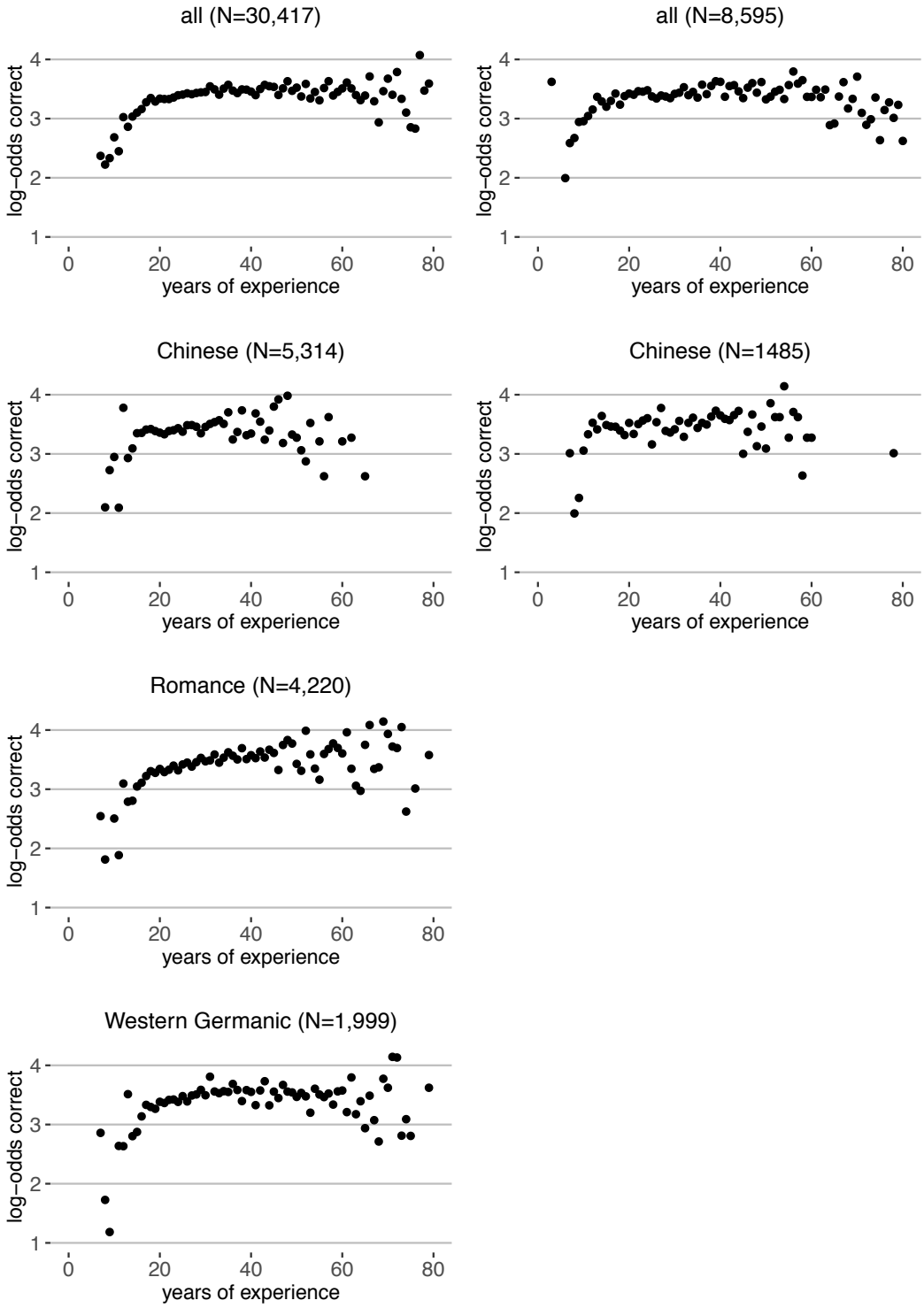


Figure S18. Learning curves (without smoothing) for immersion learners with age of first exposure 0 (**left**) and 1-5 (**right**). Although only the first 30 years are used for analyses, we plot through 80 years old for reference.

Education Differences.

To investigate the effects of education on ultimate attainment, we categorized participants according to whether their highest level of education was *secondary* (high school diploma or less: 578 immersion learners and 4,359 non-immersion learners), *undergraduate* (partial or complete undergraduate studies: 4,411 immersion learners and 6,309 non-immersion learners), or *graduate* (partial or complete graduate studies: 6,382 immersion learners and 18,006 non-immersion learners). A small number of subjects were excluded for not reporting education level. Results are shown in Fig. S19. The overall shapes of the curves are sufficiently similar to one another that no formal analysis was run.

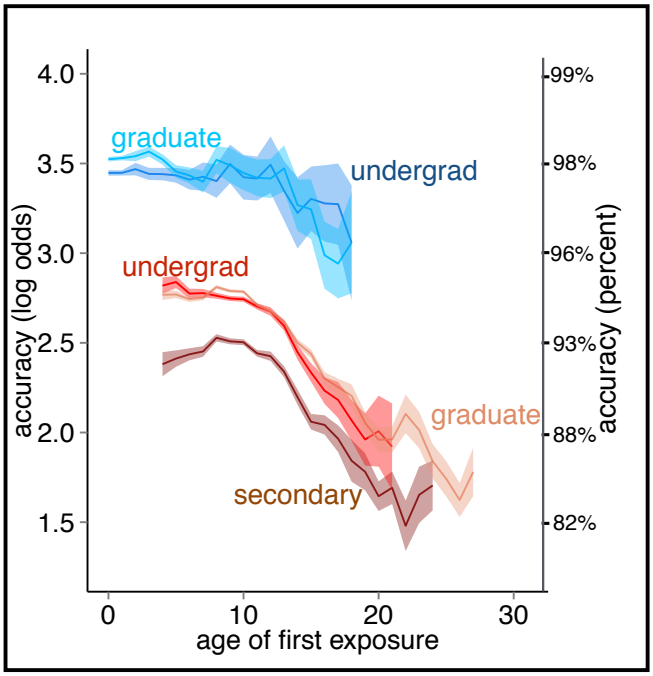


Figure S19. Ultimate attainment for immersion learners and non-immersion learners, by education level, smoothed with a three-year floating window. There were insufficient immersion learners with only secondary education for analysis.

Gender Differences.

Figure S20 separates the ultimate attainment data for male and female participants. The same exclusions used elsewhere were applied (e.g., restricting analyses to consecutive windows with at least 10 subjects). If the offset of the critical period is driven by puberty, one might expect women's learning rate to begin to decline earlier than men's. However, the learning model estimated a slightly *later* onset for a decline in the underlying learning rate in women (19.3 years old) than in men (17.9 years old).[6] Because the underlying learning rate is a theoretical estimate which requires intensive computational resources to derive, it was not feasible to use permutation analysis to determine whether the gender difference in this estimate is statistically significant. But statistically analyzing the gender difference in the age at which the ultimate level of attainment declines is more tractable. Permutation analysis showed that the effect gender on ultimate attainment was not significant either for the immersion learners (12 years old for females, 8 years old for males, $p = .80$) or for the non-immersion learners (11 years old for females, 9 years old for males; $p = .43$).[7] (An alternative would be to use the model-comparison approach used in

[6] Men: $t_c = 17.9$ $r = .16$, $\alpha = .09$, $\delta = .04$; $E = 1.00, .66, .98, .31$. Women: $t_c = 19.3$ $r = .21$, $\alpha = .10$, $\delta = 1.3$; $E = 1.00, .66, .95, .28$.

[7] As with the main analyses, ultimate attainment analyses focused on participants with at least 30 years of experience and who were no older than 70 years of age. 5,110 immersion learners were male, and 6,207 were female. 14,043 non-immersion learners were male, and 15,565 were female. The permutation analyses were conducted by shuffling participants' genders separately for each learner type (immersion, non-immersion) and each age of first exposure. For immersion and non-immersion learners of each gender, MARS analyses were applied, and the youngest age for which the slope was more negative than -.02 was recorded (because often the first segment of the MARS regression is slightly negative, the arbitrary and relatively modest threshold of -.02 was used to define "substantial decline"). This process was repeated 100 times. The two-tailed p-value is the number of iterations for which the absolute difference in the age at which ultimate attainment began substantial decline for men as opposed to women was equal to or greater

"L1 Effects", above. We developed that analysis after the permutation one used here and did not re-run analyses with the new method.)

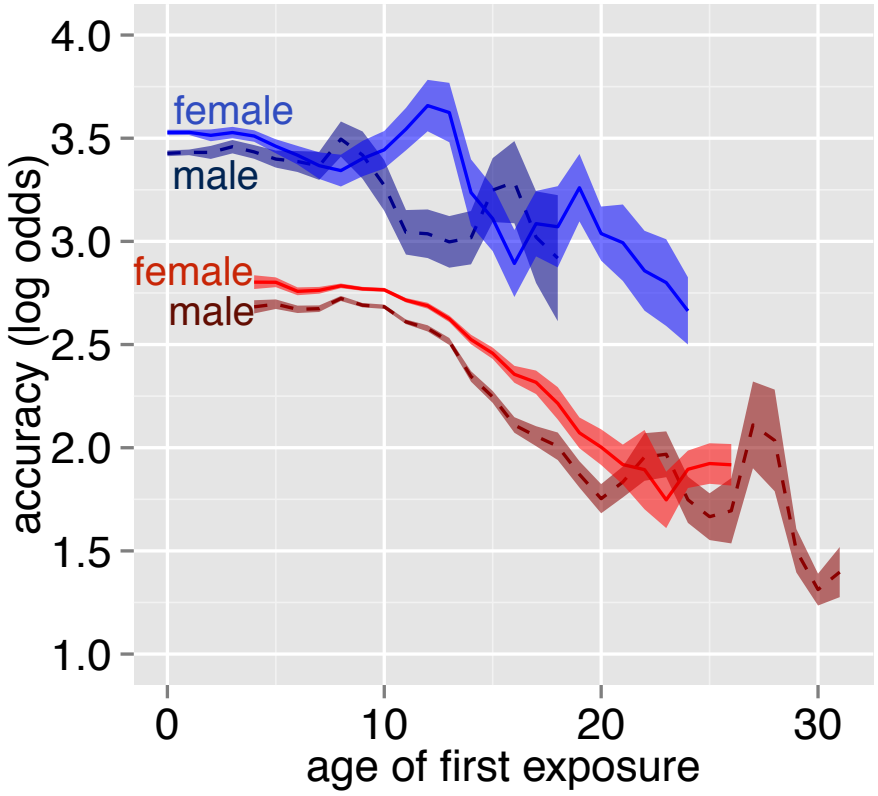


Figure S20. Ultimate attainment curves for men (A) and women (B), smoothed for presentation with a three-year floating window. Blue = immersion learners, Red = non-immersion learners. Shaded areas represent +/- 1 SE. As in Figure 6, data were smoothed by a three-year floating window, and only consecutive windows with more than 10 subjects shown.

It is possible that more sensitive analyses would find statistically significant evidence of a sex difference in the critical period (though it would be in the opposite direction of what is predicted by the sex difference in the onset of puberty). Interestingly, there was a striking sex difference in overall performance: across the entire age range,

---

than the difference actually observed, calculated separately for immersion and non-immersion learners.

women outperformed men ($p < .01$), consistent with a literature showing a female advantage in certain verbal abilities (Geary, 2010). The difference was true of each of the types of learners (monolinguals, immersion learners, and non-immersion learners; $ps < .01$), and across the range of age and experience (see Figure S20). The current data do not speak to the extent to which the difference is the result of biological or cultural causes.

Asymptotic Performance

Figure 5B shows that monolinguals and simultaneous bilinguals reach asymptote on our test at around 30 years of age. The results for later-learners also suggest protracted periods of improvement (Fig. 4). Because Fig. 4 is compact, we have re-plotted some of the curves for immersion learners and non-immersion learners in Figs. S21 and S22, respectively. Note the graphs for immersion learners involve far fewer subjects and so are noisier, even with the smoothing (see caption).

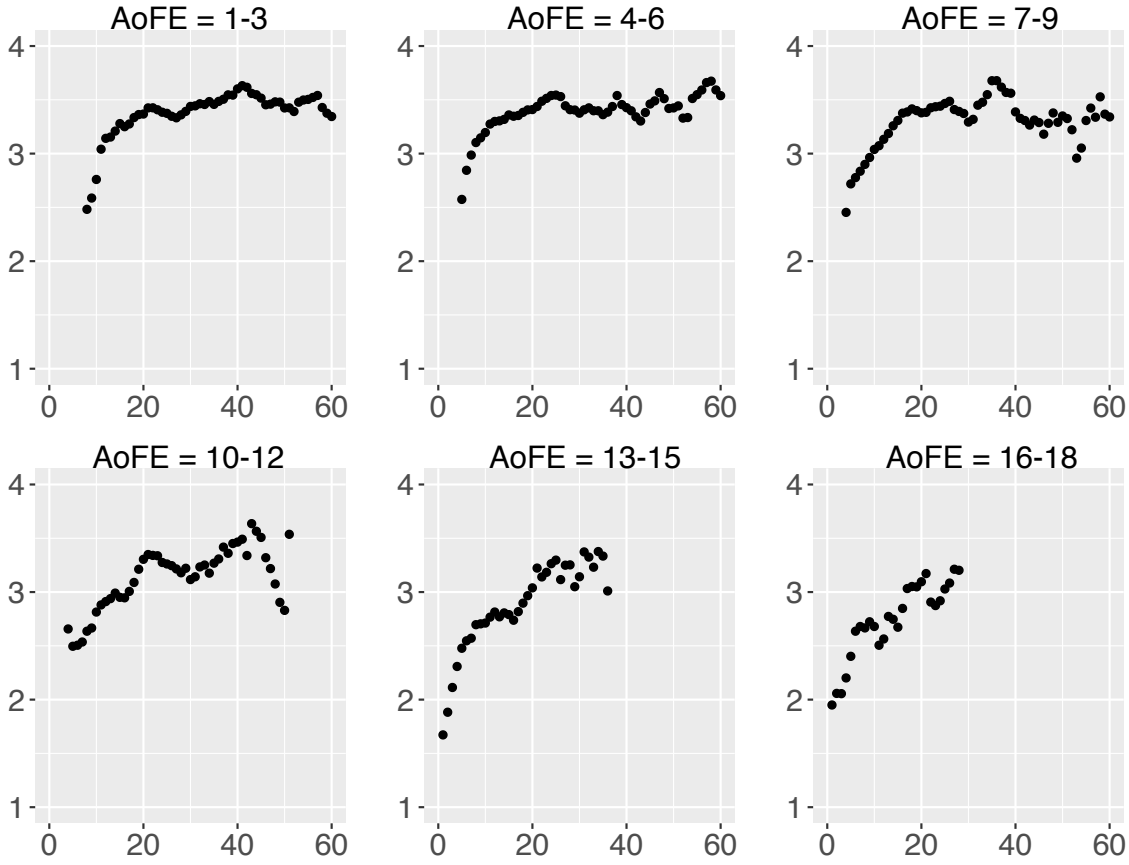


Figure S21. Panels show log-odds accuracy as a function of years of experience for immersion learners, by age of first exposure (AoFE). Graphs involves the same smoothing and exclusions as in Fig. 3. Only the first six curves are shown because later curves are shorter and provide no evidence about asymptote.

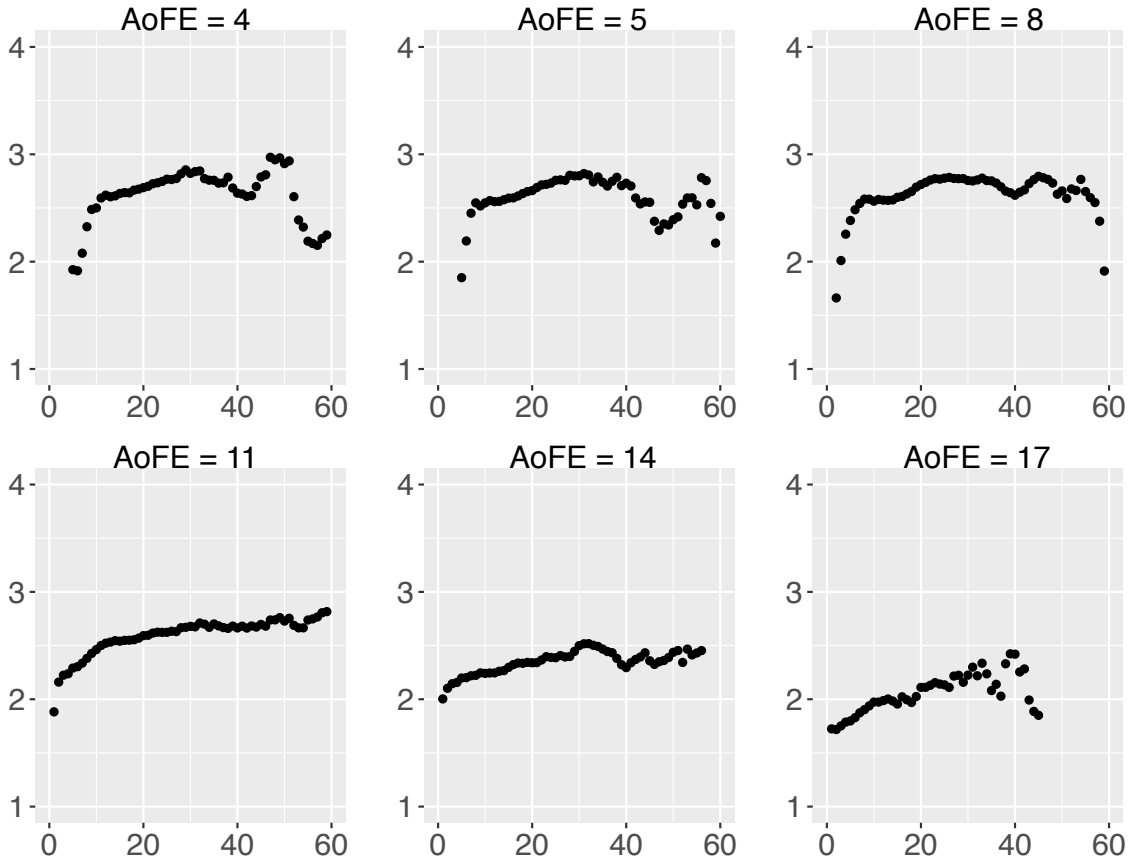


Figure S22. Panels show log-odds accuracy as a function of years of experience for non-immersion learners. Graphs involves the same smoothing and exclusions as in Fig. 3. Six curves are shown corresponding in exposure age to the six in Fig. S12. See also Fig. S1.

Results by Item

The learning curves for monolinguals varied across items. Rather than graph each of these 95 learning curves separately, we have provided the raw data, which the reader can use to generate any visualization of interest. In the meantime, we provide a single graph showing all 95 learning curves. While they are not individually distinguishable, this graph provides some intuition about the by-item variability.

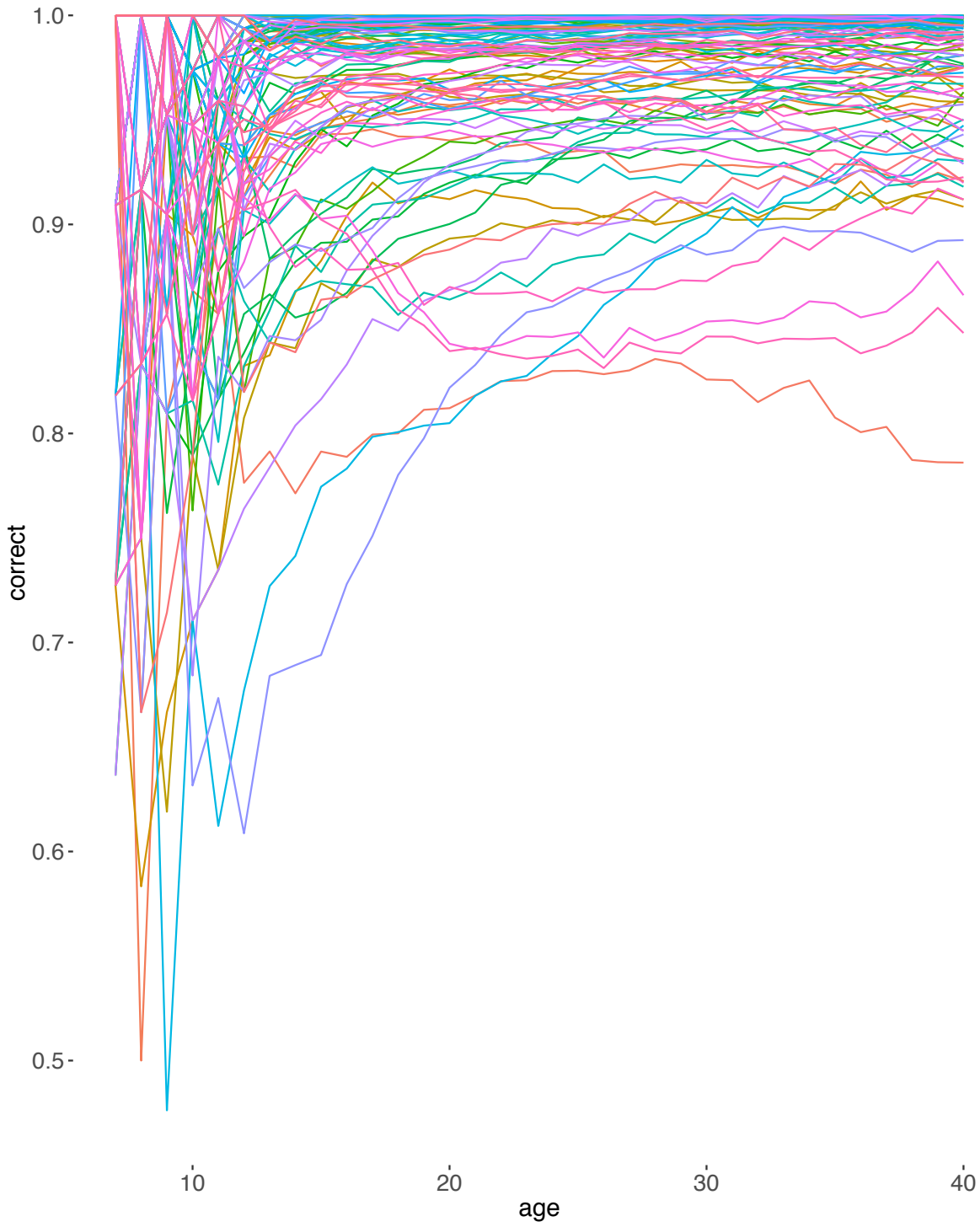


Fig. S23. Accuracy by age for each of the 95 critical items. Note that because we have different numbers of subjects at each age, this is presented in terms of percent correct,

rather than log-odds. (Readers who try creating this graph in log-odds will understand the issue.)

Materials

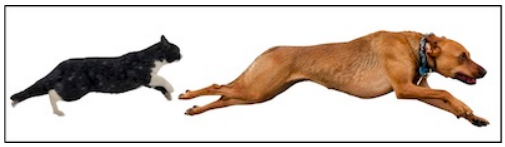
All items are included below. As noted in the main text, where possible we grouped multiple grammaticality judgments into a single multiple-choice question. Thus, Questions 9-35 are in fact 124 distinct questions.

Because the grammaticality judgment task is time-consuming and unsuitable for probing certain grammatical phenomena, we also included items that required matching a sentence to a picture (e.g., to probe topicalization and the application of linking rules). Questions 1-8 are of that format.

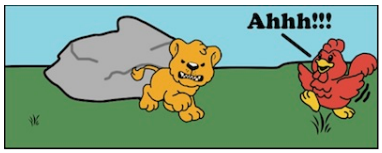
The correct answers are given in the next section.

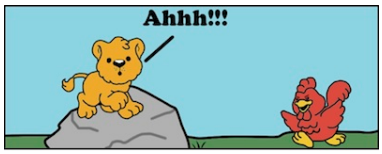
Click on the picture that best matches the sentence.

1. The dog was chased by the cat.

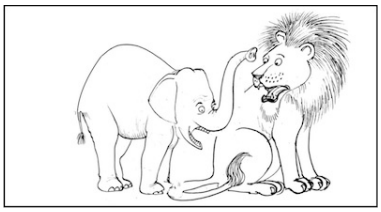


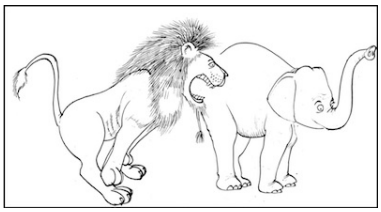
2. It was the chicken that scared the lion.



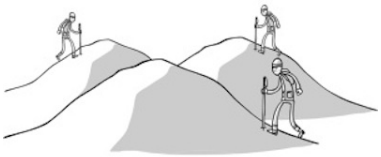


3. It was the lion that the elephant bit.



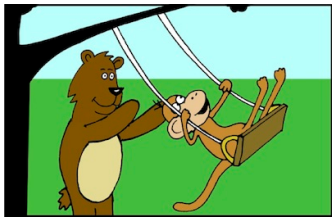


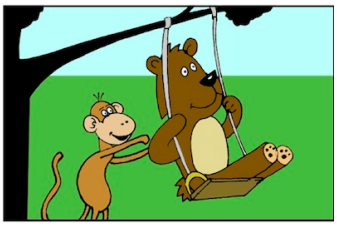
4. Every hiker climbed a hill.



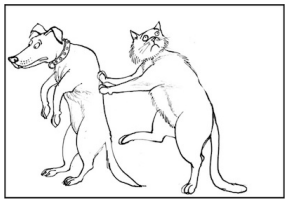
5. It was the tiger that the monkey hugged.

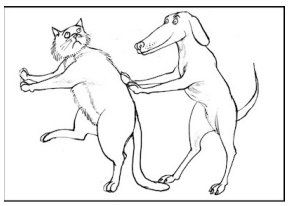
6. It was the monkey that pushed the bear.



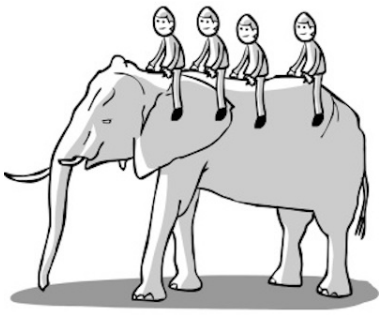


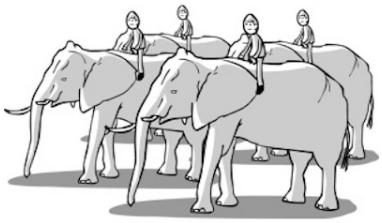
7. The dog was pushed by the cat.





8. Every child rode an elephant.





Four-Alternative Forced Choice[8]

9. Which of the following sentences sounds most natural?

a. I shan't be coming to the party after all.

b. I won't be coming to the party after all.

c. Both

d. Neither

10. Which of the following sentences sounds most natural?

a. What age are you?

b. How age are you?

c. How old are you?

c. What old are you?

[8] Including the above two questions in the analyses is not straightforward, both because they are the only questions that are not a binary forced choice and because some of the options are excluded from the analysis. There are several ways of coding the responses, though the choice is unlikely to have much effect: Accuracy was extremely high for these questions, and so they contribute very little to the variance, and moreover they represent only a tiny fraction of the included data. For simplicity, we elected to analyze the items above as if each option was an independent forced choice (thus, for example, participants are credited for two correct answers if they do *not* select (b) or (c) on Question 9).

Fill in the blank (Choose all that apply)

11. I _____ for 6 hours by dinner time.

a. will have studied

b. will have been studying

c. will had studied

d. will be studying

12. The people _____ angry.

a. is

b. be

c. were

d. are

13. The man _____ arrived yesterday needs a wakeup call at nine.

a. that

b. whom

c. which

d. where

14. We won the game, _____ we did!

a. so

b. yes

c. no

d. although

15. I _____ medicine.

a. studies

b. reads

c. study

d. read

16. He broke his leg, so he is _____.

a. in the hospital

b. in hospital

c. on hospital

d. on the hospital

17. I told Sally I was worried about the exam. She said, "Don't worry. _____"

a. He'll be right!

b. She'll be right!

c. It be okay!

d. It'll be okay!

18. If he _____, he would have helped her.

a. knew

b. had been knowing

c. had known

d. have known

19. My brother and sister _____ playing tennis at 11pm later tonight.

a. are

b. will

c. were

d. was

20. I _____ the story.

a. said

b. replied

c. declared

d. told

21. My grandmother really loved John. She left all her money to _____.

a. he

b. him

c. her

d. it

22. They _____ be traveling, but I'm not sure.

a. may

b. can

c. would

d. have

23. John ____ the library the book.

a. gave

b. donated

c. distributed

d. contributed

24. Sally _____ Mary.

a. laughed

b. happied

c. giggled

d. tickled

25. _____ lives in the White House.

a. A President Obama

b. The President Obama

c. These President Obama

d. President Obama

26. The sun is in _____.

a. the sky

b. a sky

c. an sky

d. sky

27. I believe in _____.

a. these justice

b. justice

c. a justice

d. the justice

28. Sorry to disturb you _____.

a. with the weekend

b. under the weekend

c. at the weekend

d. on the weekend

29. I would _____ go home.

a. like

b. prefer

c. rather

d. want

30. Bill _____ the cup with wine.

a. poured

b. filled

c. drained

d. dripped

31. I play _____ the soccer team.

a. at

b. in

c. on

d. inside

Choose all that are grammatical

32.

a. John agreed the contract.

b. Sally appealed against the decision.

c. I'll write my brother.

d. I'm just after telling you.

e. The government was unable to agree on the budget.

f. I after ate dinner.

g. Who did Sue ask why Sam was waiting?

h. He thought he could win the game.

33.

a. The committee were divided on the question.

b. She resigned Thursday.

c. He said that she is taking a trip.

d. He said that she was taking a trip.

e. Sally swam two miles. Wore a pair of 100 goggles.

f. I'm going to Wisconsin next week.

g. He encouraged her to travels around the world.

h. I'm wanting dessert.

34.

a. I worked for five years.

b. Who did Bill ask why Jane was talking to?

c. Who whom kissed?

d. John went to the store. Bought ice cream.

e. I’m finished my homework.

f. I'm finished with my homework.

g. We did go the beach.

h. He be working Tuesdays.

35.

a. Yesterday, John wanted to won the race.

b. Up the audience's expectations, the critics built.

c. I’m done dinner.

d. He was pulled over by the police for driving 120 miles per hour.

e. He stay working.

f. The dog the man owns barked.

g. I eats dinner.

h. Where is the pen that I gave it to you yesterday?

Scoring

Below we provide the correct answer for the 95 critical items. For the other 37 items, the "correct" answer varied by dialect.

1. Bottom
2. Bottom
3. Top
5. Bottom
6. Bottom
7. Top
9a. Incorrect
9d. Incorrect
10b. Incorrect
10c. Incorrect
11c. Incorrect
11d. Incorrect
12a. Incorrect
12b. Incorrect
12d. Correct
13c. Incorrect
13d. Incorrect
14c. Incorrect
14d. Incorrect
15a. Incorrect
15b. Incorrect
15c. Correct
16c. Incorrect
16d. Incorrect
17a. Incorrect
17c. Incorrect
17d. Correct
18b. Incorrect
18c. Correct
18d. Incorrect
19a. Correct
19b. Incorrect
19c. Incorrect
19d. Incorrect
20a. Incorrect
20b. Incorrect
20c. Incorrect
20d. Correct
21a. Incorrect
21b. Correct

21c. Incorrect
21d. Incorrect
22a. Correct
22b. Incorrect
22c. Incorrect
22d. Incorrect
23c. Incorrect
23d. Incorrect
24a. Incorrect
24b. Incorrect
24c. Incorrect
24d. Correct
25a. Incorrect
25b. Incorrect
25c. Incorrect
25d. Correct
26a. Correct
26b. Incorrect
26c. Incorrect
26d. Incorrect
27a. Incorrect
27b. Correct
27c. Incorrect
27d. Incorrect
28a. Incorrect
28b. Incorrect
29a. Incorrect
29b. Incorrect
29c. Correct
29d. Incorrect
30a. Incorrect
30b. Correct
30c. Incorrect
30d. Incorrect
31a. Incorrect
31d. Incorrect
32e. Correct
32f. Incorrect
32h. Correct
33d. Correct
33e. Incorrect
33f. Correct
33g. Incorrect
34a. Correct
34b. Incorrect
34c. Incorrect

34d. Incorrect
34f. Correct
34h. Incorrect
35a. Incorrect
35b. Incorrect
35d. Correct
35e. Incorrect
35g. Incorrect
35h. Incorrect

Supplementary References

Geary, DC (2010) *Male, female: The evolution of human sex differences*. (American Psychological Association, Washington, DC) 2nd edition.

Muggeo, VMR (2014) Segmented: An R package to fit regression models with broken-line relationships. R version 0.4-0.0. http://CRAN.R-project.org/web/packages/segmented.