

Short Communication

More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence

Tony Chen, Joshua K. Hartshorne *

Department of Psychology, Boston College, USA

ARTICLE INFO

Keywords:

Language acquisition
Critical period
Replication
Item response theory

ABSTRACT

The ability to attain native-like proficiency of a second language is heavily dependent on the age at which learning begins. However, the exact properties of this phenomenon remain unclear, and the literature is divided. Recently, Hartshorne, Tenenbaum, & Pinker presented a novel computational analysis of over 600,000 subjects, estimating that the ability to learn syntax drops at 17.4 years of age [Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177, 263–277]. However, the novelty of the dataset and analyses raises questions and suggests caution [Frank, M. C. (2018). With great data comes great (theoretical) opportunity. *Trends in cognitive sciences*, 22(8), 669–671]. In the present paper, we address several such concerns by employing improved psychometric measurement, calculating confidence intervals, and considering alternative models. We also present data from an additional 466,607 subjects. The results support the prior report of a sharp decline in the ability to learn syntax, commencing at the tail end of adolescence.

1. Introduction

There is a clear negative relationship between the age at which someone begins learning a second language and how well they learn it. Nonetheless, characterizing this relationship — *When does learning ability decline? How rapidly?* — has proven extraordinarily difficult, complicating attempts to explain it (Birdsong, 2013; Birdsong & Molis, 2001; Flege, 2018; Johnson & Newport, 1989; Snow & Hoefnagel-Höhle, 1978; Vanhove, 2013). A primary difficulty is that children's advantage in learning only appears over long time scales: During the first few months of learning, the relationship between age and learning success is inverted, with adults actually learning more quickly (Snow & Hoefnagel-Höhle, 1978).

Because the childhood advantage for language learning is most apparent retrospectively, much of the research for the last 40 years has focused on experienced second-language speakers, comparing their highest level of proficiency attained ("ultimate attainment") against the age at which they started learning (Birdsong, 2013; Flege, 2018). As a practical matter, the results of these studies have been contradictory and thus inconclusive — probably because of extremely low statistical power (Hartshorne, Tenenbaum, & Pinker, 2018; Vanhove, 2013). Even more problematic is that it can be shown mathematically that the relationship

between ultimate attainment and starting age provides little information about how real-time learning ability depends on age (Hartshorne et al., 2018). This problem is analogous to inferring how quickly a runner ran each leg of a race by looking only at their finish time: boiling down an entire trajectory to its beginning and end points leaves many open questions in between.

Hartshorne et al. (2018) — henceforth "HTP" — attempted to address these limitations by analyzing 669,498 responses to an online English grammar quiz. This dataset was highly diverse in terms of the subjects' native language, current age, the age at which they began learning English, and whether they learned in an immersion or non-immersion environment. HTP presented a novel analytic model that capitalized on this diversity to mathematically reconstruct how learning rate changes with age. They found that the rate of syntax learning declined about 50% at 17.4 years old — an age much later than previously supposed (Johnson & Newport, 1989; Pinker, 2000). These results provide a strong challenge to extant theories, all of which either supposed the critical period in early- or mid-childhood, or posited that there is no critical period at all. (Note that we follow HTP in using the term *critical period* as a theory-neutral descriptor of a period during which learning is most successful, regardless of cause.)

Given the novelty of HTP's methods, analyses, results and

* Corresponding author.

E-mail address: joshua.hartshorne@bc.edu (J.K. Hartshorne).

conclusions, caution is warranted. Indeed, a number of questions have been raised, most notably by Frank (2018) (see also Flege, 2018). In the present work, we address several significant concerns.

2. Limitations of HTP

2.1. Measuring uncertainty

HTP reports that learning ability plummets at 17.4 years old. However, it is unclear how precise this estimate is. As Frank (2018) notes, HTP do not provide estimates of uncertainty for any model parameters. In the present work, we use bootstrapping to derive confidence intervals for parameter estimates (Efron & Tibshirani, 1994).

2.2. Measuring knowledge

Frank (2018) notes that HTP interpreted proportion correct in the grammar test as a direct assay of grammatical knowledge. While this is standard practice in psychology, it implicitly treats each question as equally informative. This is rarely true. For instance, one can draw different conclusions from a subject correctly answering a difficult question vs. correctly answering an easy question. Conversely, we should make different inferences about a subject who misses only the most difficult question (they did not know the answer) than a subject who misses only the easiest question (they probably pressed the wrong button accidentally).

As an alternative, Frank (2018) suggests inferring grammatical knowledge using Item Response Theory (IRT). IRT provides a mathematical framework for simultaneously inferring properties of test items and subjects (Embretson & Reise, 2013). Specifically, in the four-parameter IRT model, the probability that a subject with ability θ will answer question Y_i correctly is given by:

$$P(Y_i = 1|\theta) = c_i + \frac{c_i - d_i}{1 + e^{-a_i(\theta - \beta_i)}}$$

where a_i , β_i , c_i , d_i govern the slope, horizontal shift, and lower and upper asymptotes of the curve for item i , respectively. These parameters can distinguish items with very different properties, such as varying levels of difficulty, variable easiness of guessing, and variable strengths of relationships between ability and probability of answering correctly. Note that both the item properties and the subject abilities are latent factors that must be inferred by fitting data to the model.

Using IRT abilities estimates rather than raw accuracy has two potential benefits. First, by accounting for the different properties of items, IRT can measure subject ability/knowledge (θ) more precisely than does raw accuracy (Embretson & Reise, 2013). This is demonstrated in Fig. 1, which analyzes the data from the experiment described below. While

IRT ability estimates and accuracy are highly correlated, the same level of accuracy can correspond to a range of IRT ability estimates, depending on which questions were answered correctly. It is also possible that the greater precision will reveal effects that would otherwise be obscured by noise and thus were missed by HTP. At the very least, it should improve the precision of the model results.

Second, since IRT's inferences about ability are abstracted away from the specific items used, it may be more robust to biases in stimulus selection. For instance, too many easy questions may induce floor effects, whereas too many difficult questions may induce ceiling effects. Using IRT ability estimates should help address these concerns, since these estimates are not directly related to the relative proportion of difficult or easy items.

2.3. Measuring the learning rate curve

Another potential concern is the degree to which any of HTP's findings are artifacts of modeling assumptions. HTP assumed that learning rate r varies as a function of age as

$$r(t) = \begin{cases} r_0 & t \leq t_c \\ r_0 \left(1 - \frac{1}{1 + e^{-\alpha(t - t_c - \delta)}} \right) & t > t_c \end{cases}$$

where t is current age, t_c is the age at which learning began, and t_c is a critical inflection point. Prior to t_c , learning rate is constant (r_0). Afterwards, it declines sigmoidally with shape parameters α and δ , which stretch and shift the sigmoid left or right. Grammatical knowledge is then modeled as a combination of learning rate and input over time (see HTP and Supplementary Materials for details).

HTP found a sharp drop in the learning rate at 17.4 years old. As can be seen in Fig. 2, the model is biased towards relatively sharp drops in ability. This raises questions about the degree to which the sharpness of the drop observed by HTP is artifactual.

To address this concern, we developed a new, more flexible learning rate model that can account for a wider range of shapes. It consists of a segmented sigmoid: two sigmoids with independent shape parameters that are joined at their intersection:

$$r(t) = \begin{cases} r_0 \left(1 - \frac{1}{1 + e^{-a_1(t - d_1)}} \right) & t \leq t_a \\ r_0 \left(1 - \frac{1}{1 + e^{-a_2(t - d_2)}} \right) & t > t_a \end{cases}$$

$$t_a = \frac{a_1 \cdot d_1 - a_2 \cdot d_2}{a_1 - a_2}$$

where a and d are shape parameters that adjust the slope and location of

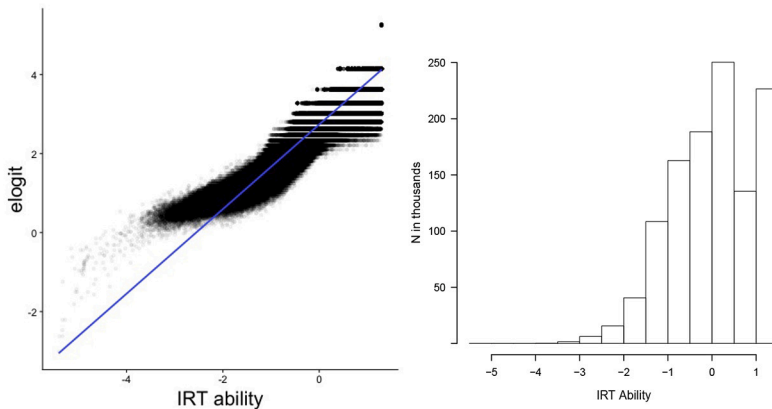


Fig. 1. Left: Comparing two different ability measures for subjects in the present dataset: the measure of accuracy used by HTP (log-odds correct, using the empirical logit function; cf. Jaeger, 2008) and ability estimates (expected value) from a four-parameter IRT model. While there is a clear relationship, for every level of accuracy (elogit), IRT infers a range of abilities. This is because it takes into account which questions the subject answered correctly, weighting different questions according to their inferred difficulty. Right: Histogram of expected ability scores. The distribution is left-skewed, reflecting the fact that monolinguals and experienced bilinguals tend to cluster near ceiling. Note also that it is much less left-skewed than the one described in Frank (2018), who used *maximum* a posteriori ability estimates rather than integrating over uncertainty. While this choice affects the histogram, it does not appreciably affect our primary analyses or conclusions (see Supplementary Materials).

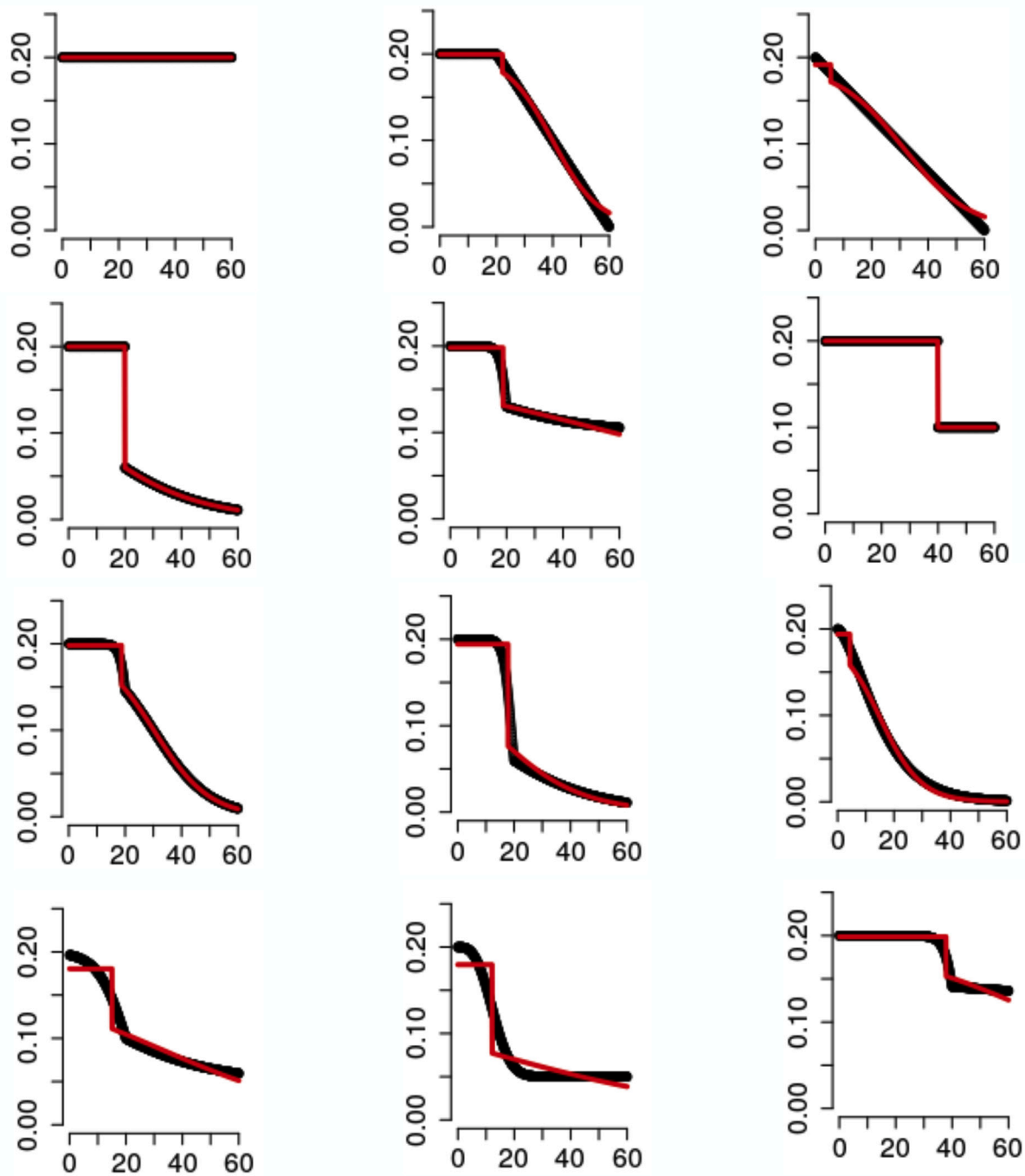


Fig. 2. Best fits of HTP's function for r (red) for a variety of curves (black). As can be seen, while HTP can fit discontinuities fairly well, it sometimes struggles to fit smoother shapes (see esp. bottom row). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the sigmoids, and t_d is the point of intersection between the sigmoids, which can be derived analytically. As shown in Fig. 3, this model is able to fit both smooth and discontinuous curves. [Note that this model does not explicitly represent the end of the critical period (there is no t_c). Nonetheless, the end of the critical period can be straightforwardly defined (see Method)].

We designed several other models that fit our test curves less well and were not considered further. For instance, we considered supplementing the segmented sigmoid curve by allowing the height of the lower asymptote to vary. However, this resulted in slightly worse fits on our test curves at the expense of additional parameters and computational complexity. We also considered a learning rate model based on a

five-parameter sigmoid. While much more elegant than either the HTP learning rate curve or the present one, it did a poor job of fitting discontinuities (see Supplementary Materials).

3. Method

We reanalyze the data from HTP, along with an additional 461,903 subjects that completed the study after data collection concluded. These new subjects were folded into the existing dataset subject to the same data exclusions used in HTP, bringing the total number of subjects up to 1,131,401. This results in a total of 324,160 monolingual English speakers (who learned only English as a child), 41,664 simultaneous

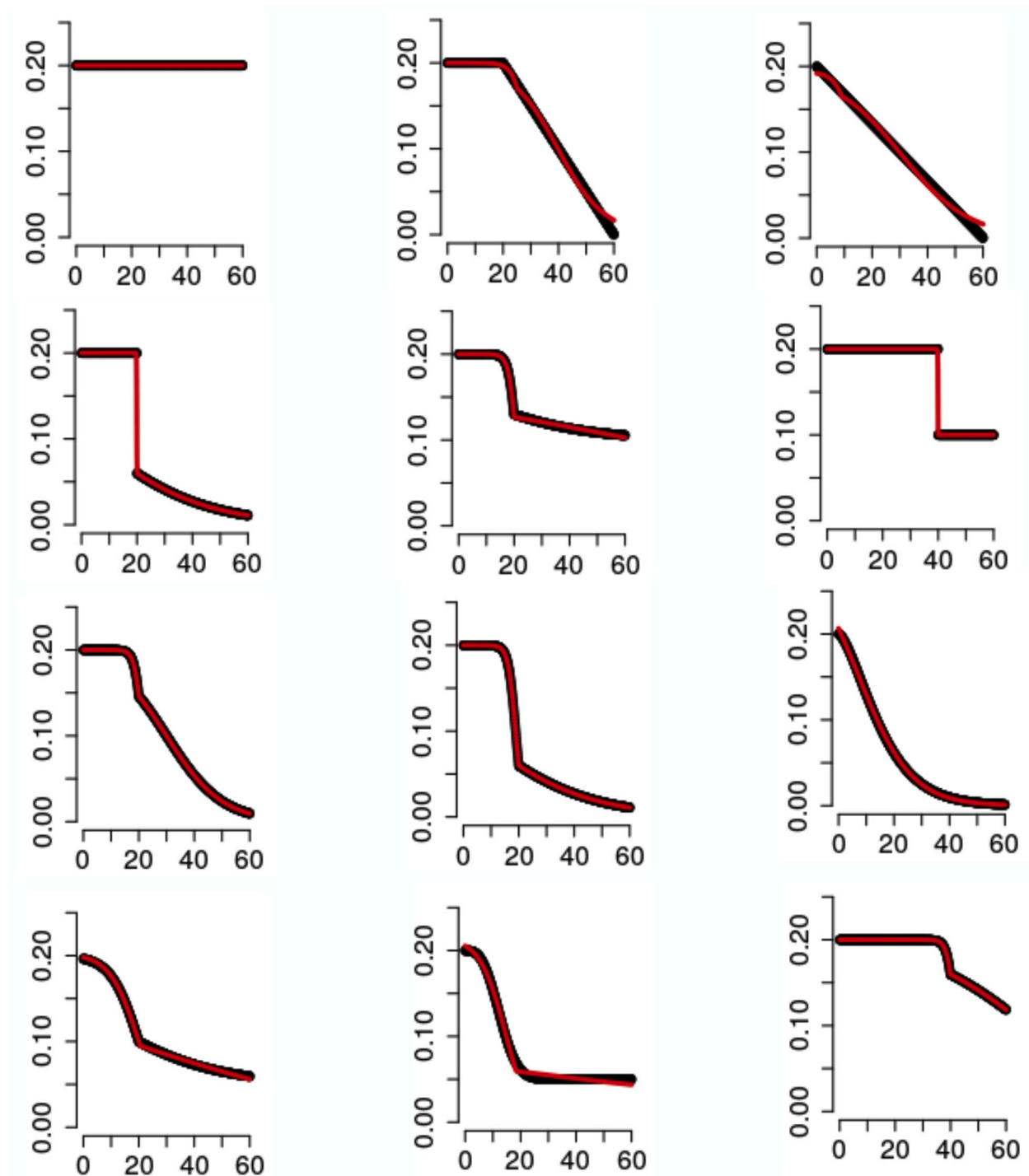


Fig. 3. Best fits of the segmented sigmoid function for r (red) for a variety of curves (black). In comparison to HTP's model, the new model performs comparably on discontinuous curves but markedly better on smooth curves (cf. Fig. 2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bilinguals (who learned both English and at least one other language from infancy), 22,208 immersion learners (late learners who learned English in an English-speaking country), and 564,999 non-immersion learners (late learners who learned English in a non-English-speaking country). Subjects completed a short online quiz consisting of 135 grammaticality judgments, 95 of which were dialect-invariant and are the subject of analysis. Subjects also provided extensive demographic information. For more details about method, subject population, and data exclusions, see HTP. In exploratory analyses, we did not find any

qualitative differences in the results for the original dataset and the full dataset. Thus, only results for the full dataset are described here.

We obtained ability estimates (θ) from an IRT model using the *mirt* package (Chalmers, 2012) (Fig. 1). Because IRT is in fact a family of models, using it requires a number of analytic choices. Following Frank (2018), we used a four-parameter model rather than the more common three-parameter model. The four-parameter model differs from the three-parameter model in that it considers the possibility not just of correctly guessing the right answer but also the probability of

inadvertently giving the wrong answer (Barton & Lord, 1981). This makes it more robust to the occasional stray response (Liao, Ho, Yen, & Cheng, 2012). Second, in inferring ability, we integrated over uncertainty, rather than using *maximum* a posteriori estimates. We feel this choice makes better use of the information in the IRT model and is less sensitive to ceiling effects. Nonetheless, neither of these analytic choices had any appreciable effect on our critical period analyses, which is reassuring (see Supplementary Materials).

We modeled learning using both the original HTP model and the new segmented-sigmoid model. Parameters were fit using differential evolution (Storn & Price, 1997). Following HTP, we fit data based on age bins, rather than the raw data. This avoids overly weighting the

monolinguals, who constitute an outsized proportion of the data. For additional details on model fitting, see Supplementary Materials. Confidence intervals for parameter estimates were obtained by re-fitting the models to 1000 bootstrap samples, in which the dataset was resampled with replacement and re-binned for each bootstrap run.

4. Results and discussion

The inferred learning rate curves are shown in Fig. 4. Cross-validated model fits are shown in Table 1. For parameter estimates with confidence intervals, see Table S3.

Neither the choice of model nor the use of IRT- or elogit-based ability

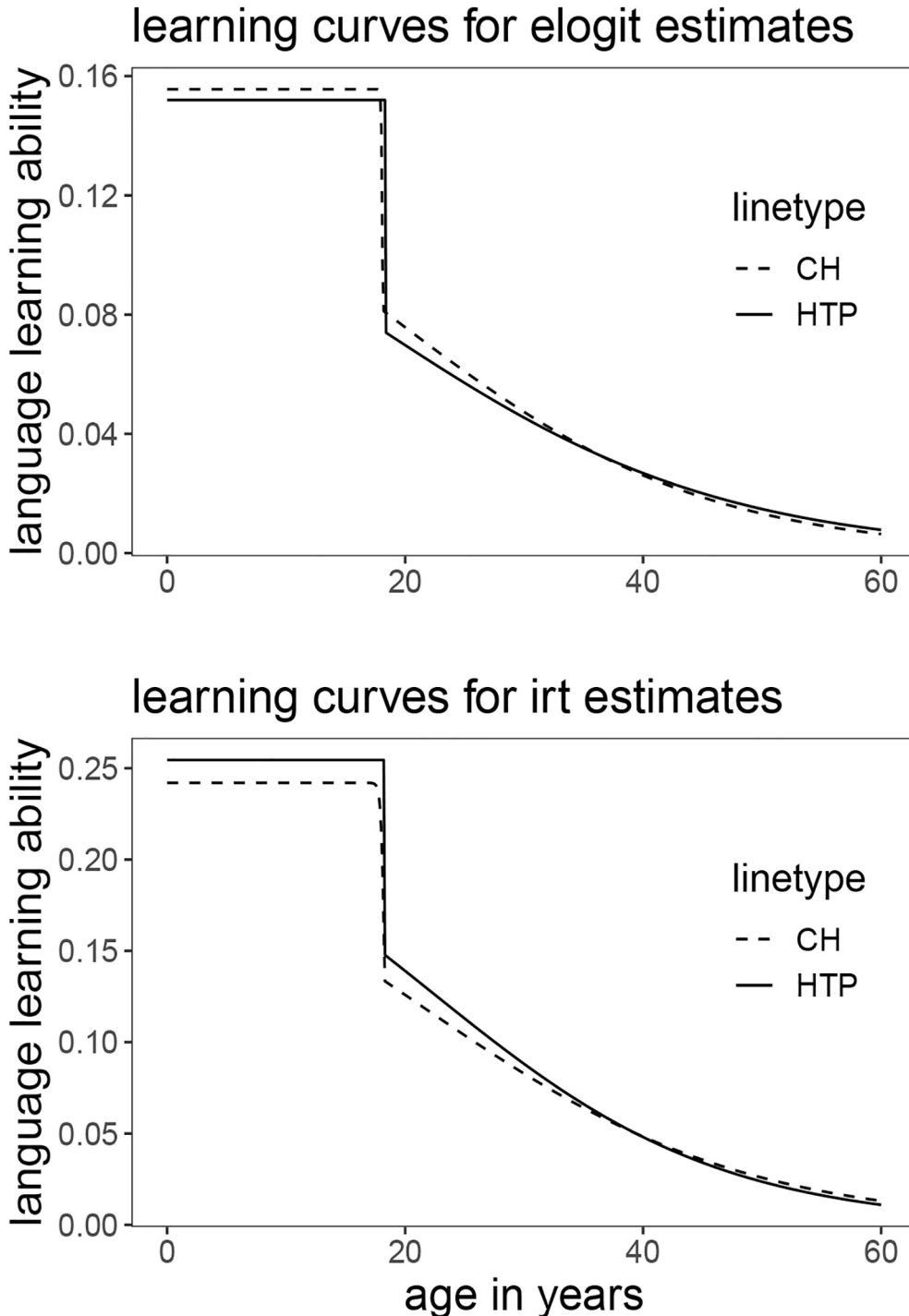


Fig. 4. Best fitting language learning curves for each model type and ability measure. The graph on the top shows the segmented sigmoid (CH) versus the original flat sigmoid (HTP) model fits using empirical log-odds (elogit) estimates, while the graph on the bottom shows the two model fits using the IRT 4pl ability estimates. In each case, the difference between the results of the old HTP model and the new segmented-sigmoid model is minimal. Furthermore, the location of the drop is essentially equal across both model type and ability measure. Note that because IRT and elogit measure ability differently, the units of the y-axes are different.

Table 1

Table of the four different models, their corresponding 10-fold cross-validated R^2 values, and best-estimate critical age (with confidence intervals). In order to directly compare the models, the critical age was defined as when learning rate fell to 95% of its initial rate. As should be clear from Fig. 4, results, other thresholds resulted in nearly identical estimates (see Supplementary Materials).

Ability Estimate	Model Type	R^2	Critical age (95% CI)
IRT	New model (segmented-sigmoid)	0.89	17.87 [17.01, 18.61]
IRT	Old model (HTP)	0.89	18.25 [17.50, 18.60]
Elogit	New model (segmented-sigmoid)	0.90	17.97 [16.96, 19.29]
Elogit	Old model (HTP)	0.90	18.34 [17.66, 19.41]

measures had much effect on the shape of the learning rate curve (Fig. 4). Of critical interest is the end of the critical period. In HTP's model, this is represented as an explicit parameter (t_c), but our model does not, due to its monotonic and asymptotic shape. Thus for *all* models, we defined the critical age as the age at which learning rate declined to 95% its original rate. (Other thresholds provided similar results; see Supplementary Methods.) These quantitative results confirm the qualitative impression of Fig. 4: Any differences in estimated critical age across models are well within the bounds of uncertainty (Table 1; see also Fig. 5). (Estimates and confidence intervals for all model parameters

can be found in Table S3).

Thus, it appears that despite new data, a new model, and a new analysis method, the results largely match those of HTP. In particular, HTP's finding of a decline in syntax-learning ability at 17–18 years of age does not appear to be an artifact of HTP's model preferring sharp discontinuities or of their statistical method for estimating grammatical knowledge. Moreover, the estimate was quite precise. Not only are the error bars fairly tight for our estimate, but nearly doubling the sample size barely affects the results: Using HTP's analyses, we obtain a point estimate for the full dataset (18.4) that is within one year of the point estimate for the original dataset (17.4). The robustness of these results across analysis methods and datasets should increase confidence in the robustness of the conclusions.

These results also address a concern about ceiling effects. As noted above, Frank (2018) worried that the fact that monolinguals and experienced bilinguals score near ceiling on the grammar test is consistent with ceiling effects. These concerns may be mitigated somewhat by our finding that ability scores are somewhat less left-skewed if we take into account uncertainty in the ability estimates (see above and Supplementary Materials). They are further mitigated by HTP's finding that the critical period was the same for the easiest items (which have a much lower ceiling) as for the hardest items. Nonetheless, the general point that the results might have been different if different items were used can only be fully addressed by running more studies with more items (we return to this issue below).

The present results do not, of course, remove all worries or answer all

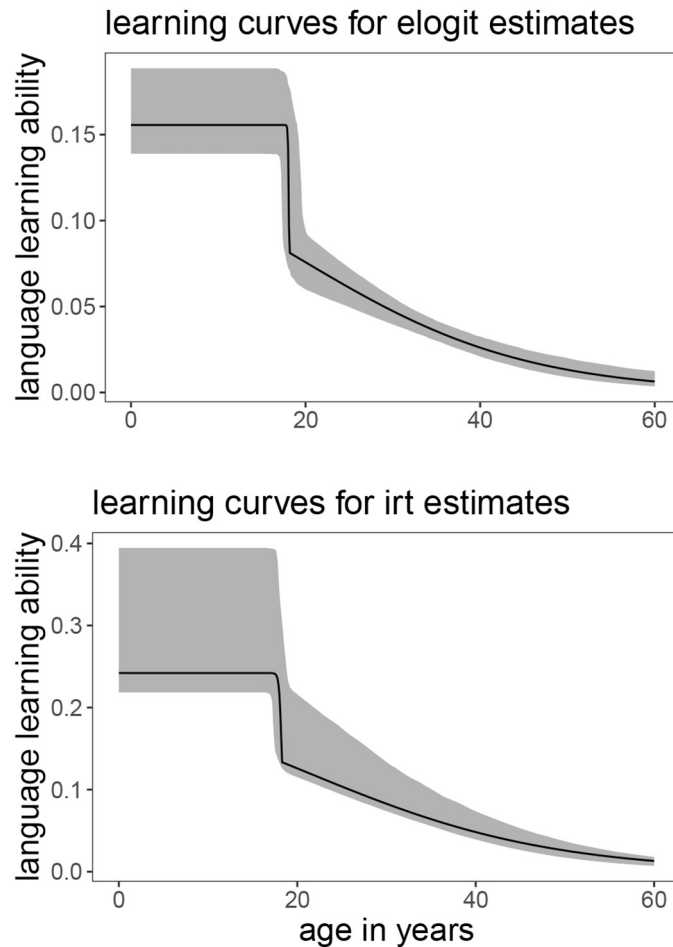


Fig. 5. Learning curves for the best fitting segmented sigmoid (CH) model using elogits (top) and 4pl IRT estimates (bottom), along with 95% confidence bands (shown in gray). The black line in both figures is the best fitting curve, and corresponds to the dashed lines in Fig. 4. Although the model is somewhat uncertain about the initial height of the learning curve for both types of ability estimates, the location of the drop is similar among all bootstrapped learning curves. Note that because IRT and elogit measure ability differently, the units of the y-axes are different.

questions. The data may be biased or confounded. Our results may not generalize beyond the target language (English). While our stimuli were meant to provide a broad assay of English syntax, we cannot know just how representative they are without a better understanding of syntax itself. There are some reasons for reassurance — HTP found a similar critical period for early-acquired and late-acquired grammatical patterns, and we found similar results when using IRT models rather than accuracy, thus abstracting away (some of) the peculiarities of the stimuli — but only more studies can provide certainty. Similarly, our subjects may be unusual — successful bilingualism is not randomly assigned, nor is age of first exposure (see also [Flege, 2018](#)) — and there may be confounds with education and socio-economic status ([Frank, 2018](#)). Because the ideal solution — random assignment — is not feasible, the best path forward is to investigate investigate specific, testable hypotheses about potential biases and confounds.

In terms of theoretical issues, [Frank \(2018\)](#) notes that HTP's model (and ours) assumes a finite amount of grammar to be learned. This is a common assumption and simplifies the mathematics, but it is not beyond challenge ([Chater & Christiansen, 2018](#)). For instance, construction grammar approaches posit that syntax consists of stored, (semi-)generalizable patterns not too different in nature from vocabulary items ([Croft, 2001](#); [Fillmore, 1988](#); [Goldberg, 2003, 2006](#)). Like vocabulary, this set is in principle unbounded. It is unclear whether a construction grammar-inspired model of the present data would lead to different conclusions about critical periods is unclear. Unfortunately, we have not yet identified a tractable way of instantiating such a model, so we leave it to future work.

Similarly, we elide any effects of the first language on the second. On the analytic side, our IRT models assume that item difficulty is independent of the speaker's native language, which is not the case (English tenses are harder for Mandarin speakers than Spanish speakers). This simplification is probably reasonable for the present study: the sheer diversity of subjects and items renders this imprecision more a source of noise than bias, and addressing it would make any such IRT model enormously complex, prone to overfitting, and at risk of circularity. However, addressing the interaction of subject and item properties would allow more precise results and has theoretical implications. One question of particular interest is whether the critical period is later for languages that are more similar to one's first language: a Spanish speaking native might be able to start later, and perhaps learn faster, than an equivalent Mandarin speaker. Unfortunately, while the diverse set of native languages in our data set means that results are not overly dependent on any particular native language, it also means we cannot easily compare native languages. In particular, the binning procedure used in HTP and in this paper — which prevents the uneven distribution of data across ages from biasing model fits — is inefficient, and applying it to subsets of the data results in too many bins with too few subjects. Even with Russian, the largest non-English group in our sample ($N = 135,185$), fully 46% of bins would be excluded for insufficient subjects, relative to the full data set. We are currently investigating more data-efficient alternatives such as first fitting the data with Bayesian regression splines, but fully developing this new analytic method is beyond the scope of the present investigation. Note that HTP did investigate the effect of native language on level of ultimate attainment, age at the end of the optimal learning period, and the shape of the learning curve, ruling out any large effects (there was insufficient power to detect mid-sized or small effects).

Relatedly, our analyses assume the critical period is the same for all grammatical phenomena. There are theoretical reasons to suspect that it might not be ([Johnson & Newport, 1991](#)). Testing this possibility will require a more targeted study comparing different aspects of grammar. Because the present grammar test was intended as a holistic assessment of grammatical knowledge as a whole, it involves a wide variety of questions addressing different phenomena (often multiple phenomena in a single question), and thus does not lend itself to precisely measuring critical periods for specific grammatical phenomena.

We also have not addressed the contention that observed changes in learning rate are attributable to changes in quantity and quality of the learner's input. [Flege, 2018](#) speculates that our findings might be driven by older learners being less likely to adopt English as a primary language. While this is reflected in the data ([Fig. 6](#)), causality could run either direction, and the decline does not obviously correspond to our findings (e.g., an inflection point in late adolescence), though perhaps under a more sophisticated theory, it would. This highlights the need for well-specified theories of the connection between input and learning success that make quantitative, testable predictions, as well as a clearer picture of how input changes with the age of the learner.

Whatever the results of these continued investigations, the present (and future) findings strongly challenge extant theory. As noted by HTP, theorists have developed theories to explain why the critical period ends in early- or mid-childhood (or they have argued against critical periods entirely). Thus, theories have focused on events in childhood: synaptic pruning in the first few years of life, growing working memory capacity, competition from a first language, or hormonal changes coinciding with puberty, etc. By definition, none of these are fully consistent with our data. Moreover, the open questions above suggest the phenomenon may be far too complex for such relatively straightforward explanations. Thus, while our discussion above points to the need for new data, it also points to a desperate need for new theory (for initial steps in this direction, see [Hernandez, Bodet III, Gehm, & Shen, 2020](#)).

Author note

We thank Eric Gu, Michael Frank, Steven Pinker, Joshua Tenenbaum, members of the Princeton language acquisition community, the audience at the 2019 CUNY Conference on Human Sentence Processing, and three anonymous reviewers. JKH designed the experiment and collected the data. Both authors contributed to analysis, interpretation, and drafting. All data, materials, and code are available at <https://osf.io/vab8j/>.

Author contributions

JKH conceptualized the study. Both authors contributed to the methodology, formal analysis, and writing.

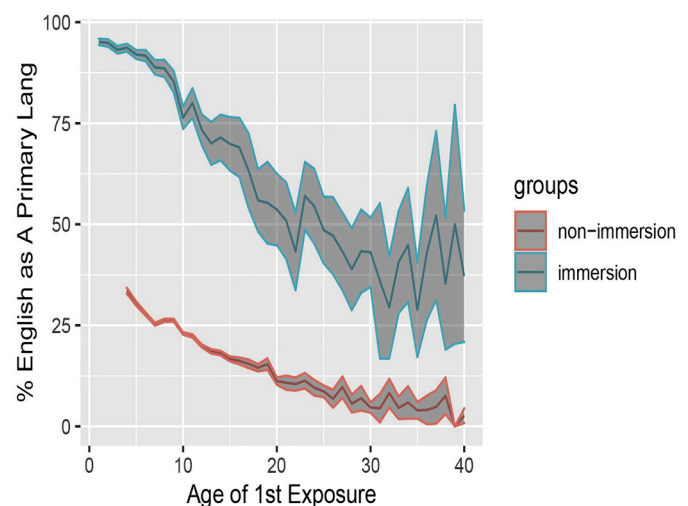


Fig. 6. Probability that subject reports English to be one of their primary languages, as a function of the age at which they began learning English, separately for immersion and non-immersion learners. Both show a continuous decline through at least 30 years of age, with no obvious discontinuities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2021.104706>.

References

- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1) (i–8).
- Birdsong, D. (2013). *The critical period hypothesis for second language acquisition: Tailoring the coat of many colors* (pp. 43–50). Cham: Springer International Publishing.
- Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language*, 44(2), 235–249.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chater, N., & Christiansen, M. H. (2018). Language acquisition as skill learning. *Current Opinion in Behavioral Sciences*, 21, 205–208.
- Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford, UK: Oxford University Press.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC press.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Hove, UK: Psychology Press.
- Fillmore, C. J. (1988). The mechanisms of “construction grammar”. In , 14. *Annual meeting of the Berkeley linguistics society* (pp. 35–55).
- Flege, J. (2018). A non-critical period for second-language learning. In A. M. Nyvad, & M. Hejná (Eds.), *A sound approach to language matters. In honor of ocke-schwen bohn*. Aarhus University.
- Frank, M. C. (2018). With great data comes great (theoretical) opportunity. *Trends in Cognitive Sciences*, 22(8), 669–671.
- Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5), 219–224.
- Goldberg, A. E. (2006). *Constructions at work*. Oxford, UK: Oxford University Press.
- Hartshorne, J. K., Tenenbaum, J., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177, 263–277.
- Hernandez, A. E., Bodet, J. P., III, Gehm, K., & Shen, S. (2020). What does a critical period for second language acquisition mean?: Reflections on hartshorne et al. (2018). *Cognition*, 206, 104478.
- 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.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60–99.
- Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: The status of subadjacency in the acquisition of a second language. *Cognition*, 39(3), 215–258.
- Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, 40(10), 1679–1694.
- Pinker, S. (2000). *The language instinct*. New York, NY: William Morrow & Co.
- Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, 1114–1128.
- Storn, R., & Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359.
- Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLoS One*, 8(7), Article e69172.