

# Second-language acquisition research on critical period effects: caveats and statistical fallacies

Jan Vanhove

Department of Multilingualism  
University of Fribourg

## Abstract

The critical period hypothesis (CPH) holds that the function between learners' age and their susceptibility to L2 input is nonlinear. This paper revisits the indistinctness found in the literature with regard to the CPH's scope and predictions. Even when its scope is clearly delineated and its predictions are spelt out, CPH studies, with few exceptions, use analytical (statistical) tools that are irrelevant with respect to the predictions made. We discuss statistical fallacies common in CPH research and illustrate an alternative analytical method (piecewise regression) by means of a reanalysis of a recent paper purporting to have found cross-linguistic evidence in favour of the CPH (DeKeyser, Alfi-Shabtay & Ravid, 2010). The data and computer commands used are provided as supplementary materials.

*Keywords:* critical period hypothesis, second-language acquisition, statistical fallacies, reanalysis, piecewise regression

It is undisputed that, in the long term and in immersion contexts, second-language learners starting acquisition early in life—and staying exposed to input and thus learning throughout several years or decades—tend to outperform later learners. Apart from being misinterpreted as an argument in favour of early foreign language instruction, which takes place in wholly different circumstances, this general age effect is also sometimes taken as evidence for a so-called 'critical period' (CP) for second-language acquisition (SLA). Derived from biology, the CP concept was famously introduced into the field of language acquisition by Penfield and Roberts (1959) and refined by Lenneberg (1967). Lenneberg argued that language acquisition needs to take place between age two and puberty—a period which he believed to coincide with the lateralisation process of the brain. (Neurological research seems to suggest that different time frames exist for the lateralisation process of different language

---

Jan Vanhove, Department of Multilingualism, University of Fribourg, Rue de Rome 1, CH-1700 Fribourg, Switzerland. Phone: +41 26 300 7141. Fax: +41 26 300 9717. E-mail: [jan.vanhove@unifr.ch](mailto:jan.vanhove@unifr.ch).

Last update: May 2nd, 2013.

No competing interests. No external funding.

I am obliged to Irmtraud Kaiser (University of Fribourg) for her help in drafting this manuscript.

functions. Most, however, close before puberty (Singleton, 2007).) However, Lenneberg mostly drew on findings pertaining to first language development in deaf children, feral children or children with serious cognitive impairments in order to back up his claims. For Lenneberg, the critical period concept is concerned with the implicit “automatic acquisition” (Lenneberg, 1967, p. 176) in immersion contexts and does not preclude the possibility of learning a foreign language after puberty, albeit with much conscious effort and typically less success.

SLA research adopted the critical period hypothesis (CPH) and applied it to second and foreign language learning, resulting in a host of studies. In its most general version, the CPH for SLA states that the ‘susceptibility’ or ‘sensitivity’ to language input varies as a function of age, with adult L2 learners being less susceptible to input than child L2 learners. Importantly, the age–susceptibility function is hypothesised to be nonlinear. Moving beyond this general version, we find that the CPH is conceptualised in a multitude of ways (see Singleton, 2005). This state of affairs requires scholars to make explicit their theoretical stance and assumptions (Long, 2005), but has the obvious downside that critical findings risk being mitigated as posing a problem to only one aspect of one particular conceptualisation of the CPH, whereas other conceptualisations remain unscathed. This overall vagueness concerns two areas in particular, viz. the delineation of the CPH’s scope and the formulation of testable predictions. Delineating the scope and formulating falsifiable predictions are, needless to say, fundamental stages in the scientific evaluation of any hypothesis or theory, but the lack of scholarly consensus on these points seems to be particularly pronounced in the case of the CPH. This article therefore first presents a brief overview of differing views on these two stages. Then, once the scope of their CPH version has been duly identified and empirical data have been collected using solid methods, it is essential that researchers analyse the data patterns soundly in order to assess the predictions made and that they draw justifiable conclusions from the results. As we will argue in great detail, however, the statistical analysis of data patterns as well as their interpretation in CPH research—and this includes both critical and supportive studies and overviews—leaves a great deal to be desired. Reanalysing data from a recent CPH-supportive study, we illustrate some common statistical fallacies in CPH research and demonstrate how one particular CPH prediction can be evaluated.

### **Delineating the scope of the critical period hypothesis**

First, the age span for a putative critical period for language acquisition has been delimited in different ways in the literature (see Singleton, 2005). Lenneberg’s (1967) critical period stretched from two years of age to puberty (which he posits at about 14 years of age), whereas other scholars have drawn the cutoff point at 12, 15, 16 or 18 years of age (see Muñoz & Singleton, 2011, for a discussion). Unlike Lenneberg, most researchers today do not define a starting age for the critical period for language learning. Some, however, consider the possibility of the critical period (or a critical period for a specific language area, e.g. phonology) ending much earlier than puberty (e.g. age 9 years according to Penfield & Roberts, 1959, or as early as 12 months in the case of phonology according to Ruben, 1997).

Second, some unclarity remains as to the setting that is relevant to the CPH. Does the critical period constrain implicit learning processes only, i.e. only the untutored lan-

guage acquisition in immersion contexts or does it also apply to (at least partly) instructed learning? Most researchers agree on the former (according to DeKeyser, 2000), but much research has included subjects who have had at least some instruction in the L2.

Third, there is no consensus on what the scope of the CP is as far as the areas of language that are concerned. Most researchers agree that a CP is most likely to constrain the acquisition of pronunciation and grammar and, consequently, these are the areas primarily looked into in studies on the CPH (Birdsong, 2006, p. 12). Some researchers have also tried to define distinguishable CPs for the different language areas of phonetics, morphology and syntax and even for lexis (see Long, 2007, for an overview).

Fourth and last, research into the CPH has focused on ‘ultimate attainment’ (UA) or the ‘final’ state of L2 proficiency rather than on the rate of learning. From research into the rate of acquisition (e.g. Krashen, Long & Scarcella, 1979; Snow & Hoefnagel-Höhle, 1977, 1978), it has become clear that the CPH cannot hold for the rate variable. In fact, it has been observed that adult learners proceed faster than child learners at the beginning stages of L2 acquisition. Though theoretical reasons for excluding the rate can be posited (the initial faster rate of learning in adults may be the result of more conscious cognitive strategies rather than to less conscious implicit learning, for instance), rate of learning might from a different perspective also be considered an indicator of ‘susceptibility’ or ‘sensitivity’ to language input. Nevertheless, contemporary SLA scholars generally seem to concur that UA and not rate of learning is the dependent variable of primary interest in CPH research. These and further scope delineation problems relevant to CPH research are discussed in more detail by, among others, Birdsong (2006), DeKeyser and Larson-Hall (2005), Long (2007) and Muñoz and Singleton (2011).

### Formulating testable hypotheses

Once the relevant CPH’s scope has satisfactorily been identified, clear and testable predictions need to be drawn from it. At this stage, the lack of consensus on what the consequences or the actual observable outcome of a CP would have to look like becomes evident. As touched upon earlier, CPH research is interested in the end state or ‘ultimate attainment’ (UA) in L2 acquisition because this “determines the upper limits of L2 attainment” (Birdsong, 2006, p. 10). The range of possible ultimate attainment states thus helps researchers to explore the potential maximum outcome of L2 proficiency before and after the putative critical period.

One strong prediction made by some CPH exponents holds that post-CP learners cannot reach native-like L2 competences. Identifying a single native-like post-CP L2 learner would then suffice to falsify all CPHs making this prediction. Assessing this prediction is difficult, however, since it is not clear what exactly constitutes sufficient nativelikeness, as illustrated by the discussion on the actual nativelikeness of highly accomplished L2 speakers (Abrahamsson & Hyltenstam, 2009; White & Genesee, 1996). Indeed, there exists a real danger that, in a quest to vindicate the CPH, scholars set the bar for L2 learners to match monolinguals increasingly higher—up to Swiftian extremes. Furthermore, the usefulness of comparing the linguistic performance in mono- and bilinguals has been called into question (see Cook, 1992; Grosjean, 1989; Muñoz & Singleton, 2011). Put simply, the linguistic repertoires of mono- and bilinguals differ by definition and differences in the behavioural outcome will necessarily be found, if only one digs deep enough.

A second strong prediction made by CPH proponents is that the function linking age of acquisition and ultimate attainment will not be linear throughout the whole lifespan. Before discussing how this function would have to look like in order for it to constitute CPH-consistent evidence, we point out that the ultimate attainment variable can essentially be considered a cumulative measure dependent on the actual variable of interest in CPH research, i.e. susceptibility to language input, as well as on such other factors like duration and intensity of learning (within and outside a putative CP) and possibly a number of other influencing factors.<sup>1</sup> It seems astonishing to us, however, that the distinction between level of sensitivity to language input and level of ultimate attainment is rarely made in the literature. Implicitly or explicitly (Birdsong, 2005, p. 111), these two are more or less equated and the same mathematical functions are expected to describe the two variables if observed across a range of starting ages of acquisition.

But even when the susceptibility and ultimate attainment variables are equated, there remains controversy as to what function linking age of onset of acquisition and ultimate attainment would actually constitute evidence for a critical period. Most scholars agree that not any kind of age effect constitutes such evidence. More specifically, the age of acquisition–ultimate attainment function would need to be different before and after the end of the CP (Birdsong, 2006). According to Birdsong (2006), three basic possible patterns proposed in the literature meet this condition. These patterns are presented in Figure 1. The first pattern describes a steep decline of the age of onset of acquisition (AOA)–ultimate attainment (UA) function up to the end of the CP and a practically non-existent age effect thereafter. Pattern 2 is an “unconventional, although often implicitly invoked” (Birdsong, 2006, p. 17) notion of the CP function which contains a period of peak attainment (or performance at ceiling), i.e. performance does not vary as a function of age, which is often referred to as a ‘window of opportunity’. This time span is followed by an unbounded decline in UA depending on AOA. Pattern 3 includes characteristics of patterns 1 and 2. At the beginning of the AOA range, performance is at ceiling. The next segment is a downward slope in the age function which ends when performance reaches its floor. Birdsong (2006) points out that all of these patterns have been reported in the literature. On closer inspection, however, he concludes that the most convincing function describing these age effects is a simple linear one. Hakuta, Bialystok and Wiley (2003) sketch further theoretically possible predictions of the CPH in which the mean performance drops drastically and/or the slope of the AOA–UA proficiency function changes at a certain point.

Although several patterns have been proposed in the literature, it bears pointing out that the most common explicit prediction corresponds to Birdsong’s (2006) first pattern, as

---

<sup>1</sup>As Newport (1991) correctly points out, the behavioural outcome, i.e. ultimate attainment, can be assumed to be integrative to the susceptibility function. Other things being equal, ultimate attainment will therefore decrease as susceptibility decreases. However, decreasing ultimate attainment levels in and by themselves represent no compelling evidence in favour of a CPH. The form of the integrative curve must therefore be predicted clearly from the susceptibility function. Additionally, we point out that the age of acquisition–ultimate attainment function can take just about *any* form when other things are not equal, e.g. duration of learning (Does learning last up until time of testing or only for a more or less constant number of years or is it dependent on age itself?) or intensity of learning (Do learners always learn at their maximum susceptibility level or does this intensity vary as a function of age, duration, present attainment and motivation?). The integral of the susceptibility function could therefore be of virtually unlimited complexity and its parameters could be adjusted to fit any age of acquisition–ultimate attainment pattern.

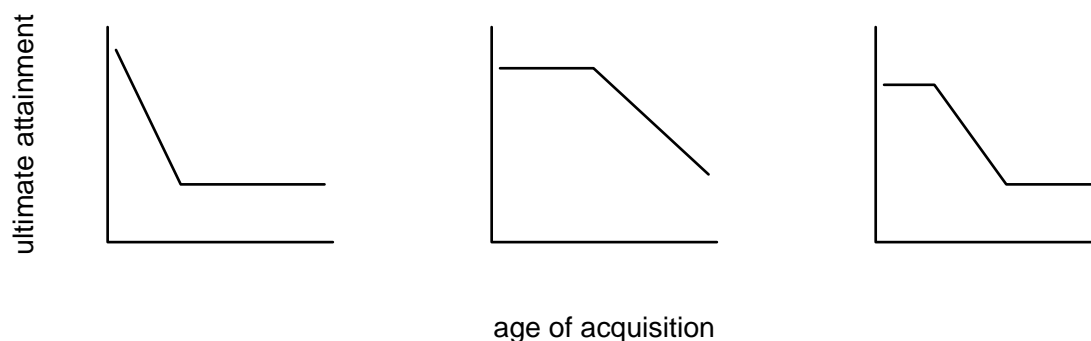


Figure 1. Three possible critical period effects (based on Birdsong, 2006, Fig. 2.).

exemplified by the following crystal-clear statement by one of the foremost CPH proponents:

[A] strong negative correlation between age of acquisition and ultimate attainment throughout the lifespan (or even from birth through middle age), the only age effect documented in many earlier studies, is not evidence for a critical period . . . [T]he critical period concept implies a break in the AoA–proficiency function, i.e., an age (somewhat variable from individual to individual, of course, and therefore an age range in the aggregate) after which the decline of success rate in one or more areas of language is much less pronounced and/or clearly due to different reasons. (DeKeyser, 2012, p. 445)

DeKeyser and before him among others Johnson and Newport (1989) thus conceptualise only one possible pattern which would speak in favour of a critical period: a clear negative age effect before the end of the critical period and a much weaker (if any) negative correlation between age and ultimate attainment after it. This ‘flattened slope’ prediction has the virtue of being much more tangible than the ‘potential nativelikeness’ prediction: Testing it does not necessarily require comparing the L2-learners to a native control group and thus effectively comparing apples and oranges. Rather, L2-learners with different AOAs can be compared amongst themselves without the need of categorising them by means of a native-speaker yardstick, the validity of which is inevitably going to be controversial (see Abrahamsson & Hyltenstam, 2009, pp. 292–293, for a brief discussion). In what follows, we will concern ourselves solely with the ‘flattened slope’ prediction, arguing that, despite its clarity of formulation, CPH research has generally used analytical methods that are irrelevant for the purposes of actually testing it.

### Inferring non-linearities: an overview

In this section we present a non-exhaustive overview of studies that have either claimed to have found evidence relevant to the ‘flattened slope’ prediction or that have been cited by others in this context. These studies can be split up in three broad and partially overlapping categories. The first category consists of studies in which statistical tools to compare means or proportions, e.g. *t*- and  $\chi^2$ -tests and ANOVAs, were used. Studies in which the correlation coefficients of the AOA–UA relationship were compared between

younger and older arrivals make up the second category. Lastly, studies in the third category used regressional methods to address the ‘flattened slope’ prediction. We will demonstrate that the analyses used in the first two categories rest on statistical fallacies, rendering them useless for the purposes of addressing the ‘flattened slope’ prediction. Regression models, we argue, present the only valid alternative, provided they are fitted correctly and interpreted judiciously.

### Group mean or proportion comparisons

The first broad category consists of studies in which the AOA continuum is discretised into bins (e.g. AOA 3–7, 8–10, 11–15 and 17–39 years in Johnson & Newport, 1989), whose UA scores or nativelikeness ratings are subsequently compared together and sometimes with those of native speakers using a series of  $t$ - or  $\chi^2$ -tests or an ANOVA. Inferences about discontinuities<sup>2</sup> in the AOA–UA function are then made on the basis of whether such comparisons reach significance or not. A recent paper by Abrahamsson and Hyltenstam (2009) is a case in point. The authors split up the AOA continuum into five bins (AOA < 1–5, 6–11, 12–17, 18–23 and 24–47 years), carried out an ANOVA with pairwise post-hoc tests on nativelikeness ratings and inferred the presence of a critical point in adolescence on the basis thereof:

[T]he main differences can be found between the native group and all other groups—including the earliest learner group—and between the adolescence group and all other groups. However, neither the difference between the two childhood groups nor the one between the two adulthood groups reached significance, which indicates that the major changes in eventual perceived nativelikeness of L2 learners can be associated with adolescence. (Abrahamsson & Hyltenstam, 2009, p. 270)

Similar group comparisons aimed at investigating the effect of AOA on UA have been carried out by both CPH advocates and skeptics, among whom Bialystok and Miller (1999, pp. 136–139), Birdsong and Molis (2001, p. 240), Flege (1999, pp. 120–121), Flege, Yeni-Komshian and Liu (1999, pp. 85–86), Johnson (1992, p. 229), Johnson and Newport (1989, p. 78), McDonald (2000, pp. 408–410) and Patkowski (1980, pp. 456–458). Not all of these authors drew direct conclusions about the AOA–UA function on the basis of these groups comparisons, but their group comparisons have been *cited* as indicative of a CPH-consistent non-continuous age effect, as exemplified by the following quote by DeKeyser (2012):

Where group comparisons are made, younger learners always do significantly better than the older learners. The behavioral evidence, then, suggests a non-continuous age effect with a “bend” in the AoA–proficiency function somewhere between ages 12 and 16. (DeKeyser, 2012, p. 448)

The first problem with group comparisons like these and drawing inferences on the basis thereof is that they require that a continuous variable, AOA, be split up into discrete

<sup>2</sup>The terms ‘discontinuity’ and ‘non-continuity’ are often used in CPH research, even though the predicted patterns (see Figure 1) do not contain discontinuities in the mathematical sense. In mathematics, a discontinuity is a ‘jump’ in the function (see Weisstein, n.d.).

bins. More often than not, the boundaries between these bins are drawn in an arbitrary fashion, but what is more troublesome is the loss of information and statistical power that such discretisation entails (see Cohen, 1983, for the extreme case of dichotomisation). If we want to find out more about the relationship between AOA and UA, why throw away most of the AOA information and effectively reduce the UA data to group means and the variance in those groups?

Second, we strongly suspect that the underlying assumption when using  $t$ - and  $\chi^2$ -tests and ANOVAs to infer the shape of the underlying AOA–UA function is one of the gravest fallacies in all of statistics: the belief that non-significant test results indicate that the group means or proportions are essentially identical. To quote Schmidt (1996), this notion is “the most devastating of all to the research enterprise” (p. 126). Yet, judging by the snippet quoted above, Abrahamsson and Hyltenstam’s (2009) reasoning seems to be that the lack of a statistical difference between the childhood groups and between the adulthood groups indicates that these groups perform at roughly the same level, whereas the presence of a statistical difference between the adolescence group and all other groups indicates a steep drop in perceived nativelikeness. Such reasoning ignores the issue that when the default null hypothesis of no difference is adopted as or integrated into the research hypothesis, the statistical power of the tests, i.e. the probability of finding a statistically significant difference when the actual population means differ by a prespecified minimum effect size, should be substantially higher than what tends to be the case in the social sciences (Sedlmeier & Gigerenzer, 1989).

In order to illustrate the gravity of this problem, we computed the power that Abrahamsson and Hyltenstam (2009) would actually have had to detect a significant difference between their two childhood groups ( $n_1 = 53$ ,  $n_2 = 54$ ) if the underlying population effect size were, in fact, medium-sized ( $d = 0.50$ , see Cohen, 1992). These power computations were carried out with the `pwr.t2n.test()` function in the `pwr` package for R (Champely, 2009).<sup>3</sup> We found that Abrahamsson and Hyltenstam’s (2009) power was about 0.73 assuming a two-tailed  $t$ -test with  $\alpha$  fixed at 0.05. While this is better than what is typically found in social science papers (Sedlmeier & Gigerenzer, 1989), it still means that in 27 % of cases, even a medium-sized effect would have gone undetected. Since Abrahamsson and Hyltenstam (2009) used post-hoc tests that corrected the individual  $\alpha$ -levels downwards to maintain the familywise Type I error rate, their actual power was even lower.<sup>4</sup> In the case of Johnson and Newport’s (1989) oft-cited study, which claimed that participants with AOAs between 3 and 7 years ( $n = 7$ ) did not behave differently from native speakers ( $n = 23$ ) and on that basis surmised the presence of a non-continuity, this lack of power is even more pronounced at a mere 0.20, assuming a medium-sized effect size and a two-tailed test with  $\alpha$  fixed at 0.05. This means that in a whopping 80 % of cases a medium-sized effect would have gone undetected. Note that Sedlmeier and Gigerenzer (1989) suggest that researchers have a power level of 0.95 before they accept null hypotheses, which is equivalent to the typical requirement of needing a  $p$ -value lower than 0.05 before rejecting the null hypothesis in favour of a non-null research hypothesis, but which would require about 105 participants

<sup>3</sup>We refer to the appendix for more information about the software. For a highly accessible introductory text to power analysis, we refer to Cohen’s (1992) *Power primer*.

<sup>4</sup>Note that we do not blame Abrahamsson and Hyltenstam (2009) for maintaining the familywise  $\alpha$  level. Our point is merely that our power computations are generous.

per group (assuming  $d = 0.50$ ).

Thus, within an ‘orthodox’ frequentist framework, group mean or proportion comparisons are fine for establishing that a difference *does* likely exist between two groups (though subject to a host of caveats, see Cohen, 1994; Ioannidis, 2005; Simmons, Nelson & Simonsohn, 2011; Wetzels et al., 2011; and many others),<sup>5</sup> but using them to infer that a difference does *not* exist is highly suspect. The only reliable inference that they by themselves allow in CPH research is that younger learners tend to outperform older learners in some domains of language (e.g. pronunciation and syntax), which all scholars implied in the debate essentially agree on. In sum, inferring the precise shape of a bivariate relationship using *t*-tests, ANOVAs or  $\chi^2$ -tests is at the very least cumbersome and prone to errors.

### Comparison of correlation coefficients

The second broad category, which is not mutually exclusive with the first category, consists of studies that address the discontinuity hypothesis by computing and comparing correlation coefficients between AOA and UA for two or more AOA subgroups. In a sense, this approach represents an improvement over group mean or proportion comparisons as the AOA data are treated as a continuous variable. Nevertheless, this approach, too, rests on a fallacious assumption, namely that differences in correlation coefficients are indicative of differences in slopes. We suspect that the correlation-based approach dates back to Johnson and Newport (1989), who split up their participants into two AOA-defined groups and found that UA as measured using a GJT correlated strongly and significantly in the early arrivals (age 3–15,  $n = 23$ ,  $r = -0.87$ ) but not in the older arrivals (age 17–39,  $n = 23$ ,  $r = -0.16$ ). Johnson and Newport took this to suggest that “language learning ability slowly declines as the human matures and plateaus at a low level after puberty” (p. 90).

Correlation-based inferences about slope discontinuities have similarly explicitly been made by CPH advocates and skeptics alike, e.g. Bialystok and Miller (1999, pp. 136 and 140), DeKeyser (2012, p. 448), DeKeyser et al. (2010, pp. 425 and 430) and Flege et al. (2006, pp. 166 and 169). Others did not explicitly infer the presence or absence of slope differences from the subset correlations they computed (e.g. Birdsong & Molis, 2001; DeKeyser, 2000; Flege et al., 1999; Johnson, 1992), but their studies nevertheless featured in overviews discussing discontinuities (DeKeyser & Larson-Hall, 2005, Tables 5.1 and 5.2; DeKeyser, 2012, Tables 27.1, 27.2 and 27.3). Indeed, the most recent overview draws a strong conclusion about the validity of the CPH’s ‘flattened slope’ prediction on the basis of these subset correlations:

In those studies where the two groups are described separately, the correlation is much higher for the younger than for the older group, except in Birdsong and Molis (2001), where there was a ceiling effect for the younger group. This global picture from more than a dozen studies provides support for the non-continuity of the decline in the AoA–proficiency function, which all researchers agree is a hallmark of a critical period phenomenon. (DeKeyser, 2012, p. 448)

In Johnson and Newport’s (1989) specific case, their correlation-based inference that UA levels off after puberty happened to be largely correct: the GJT scores are more or

<sup>5</sup>But note that the framework itself is heavily criticised, see Gill (1999), Hunter (1997), Rozeboom (1960) and Schmidt (1996), to name just a few.



less randomly distributed around a near-horizontal trend line (see Birdsong & Molis, 2001, p. 240, Fig. 3). Ultimately, however, it rests on the fallacy of confusing correlation coefficients with slopes, which seriously calls into question conclusions such as DeKeyser's (2012).

For clarity's sake, let's briefly review the difference between correlation coefficients and slopes. The slope of a function is defined as the increment with which and the direction in which the value on the  $y$ -axis changes when the value on the  $x$ -axis is increased by one increment. In a linear regression model of the form  $\hat{y} = \alpha + \beta x$ ,  $\alpha$  is the value of  $\hat{y}$  (i.e. the expected  $y$ -value according to the model) when  $x = 0$ , i.e. the intercept. The coefficient that  $x$  takes in this equation,  $\beta$ , represents the slope of the regression function, i.e. it expresses how  $\hat{y}$  changes when  $x$  is increased by one increment. In principle,  $\beta$  can take any value between negative and positive infinity.<sup>6</sup>

The Pearson correlation coefficient,  $r$ , on the other hand, expresses the strength of the linear relationship between two variables. It is bound between  $-1$  (perfect negative relationship) and  $1$  (perfect positive relationship). If  $r$  equals  $-1$  or  $1$ , a straight line captures all the data points; the closer  $r$  comes to zero, the farther from such a linear line the data points are scattered. In simple linear functions,  $\beta$  and  $r$  are linked to each other in that  $\beta$  is  $r$  times the ratio of the sample standard deviations of the  $y$ - and  $x$ -variables:  $\beta = r \frac{s_y}{s_x}$ . Crucially, however, the relationships between two pairs of variables can be characterised by the same functional regression form but still have radically different  $r$  coefficients, and the other way around (see Figure 2).

What this boils down to is that a hypothesis concerning the slope of a function must be addressed by comparing  $\beta$  coefficients computed using regression techniques rather than by comparing correlation coefficients. But then why are the AOA–UA correlations typically weaker in the older arrivals than in the younger ones? Assuming, for the sake of the argument, that the slope of the AOA–UA function is identical in both groups (Eq. 1), we can substitute the  $\beta$  coefficients for the correlation coefficients times the ratio of the relevant sample standard deviations (Eq. 2).

$$\beta_{\text{old}} = \beta_{\text{young}} \tag{1}$$

$$r_{\text{old}} \frac{s_{\text{UA}(\text{old})}}{s_{\text{AOA}(\text{old})}} = r_{\text{young}} \frac{s_{\text{UA}(\text{young})}}{s_{\text{AOA}(\text{young})}} \tag{2}$$

It can then straightforwardly be deduced that, other things equal, the AOA–UA correlation in the older group decreases as the UA variance in the older group increases relative to the UA variance in the younger group (Eq. 3).

$$r_{\text{old}} = r_{\text{young}} \frac{s_{\text{UA}(\text{young})}}{s_{\text{UA}(\text{old})}} \frac{s_{\text{AOA}(\text{old})}}{s_{\text{AOA}(\text{young})}} \tag{3}$$

Lower correlation coefficients in older AOA groups may therefore be largely due to differences in UA variance, which have been reported in several studies (Birdsong & Molis,

---

<sup>6</sup>In this context  $\alpha$  and  $\beta$  do not refer to the probability of committing Type I and Type II errors, respectively.

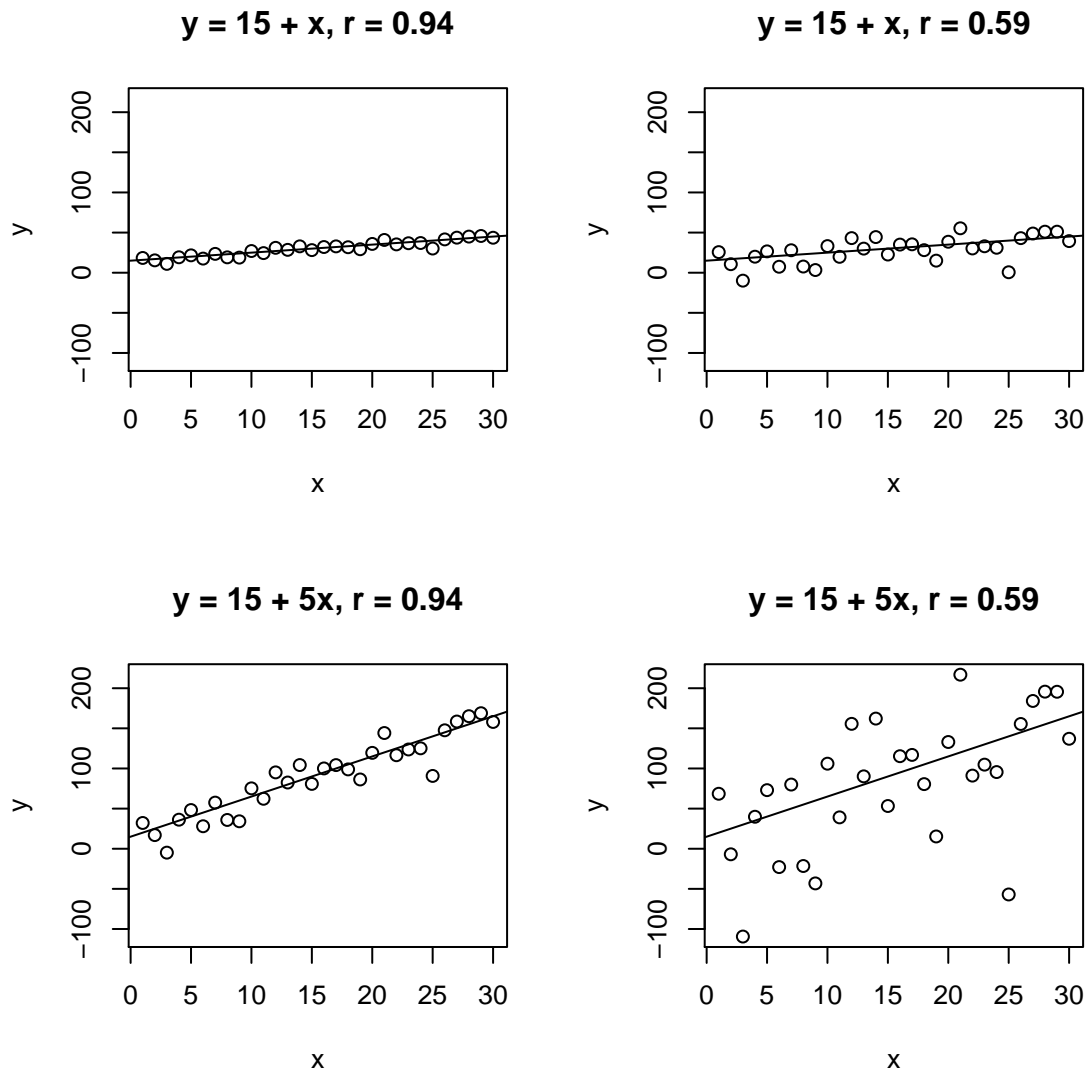


Figure 2. Illustration of the difference between correlation coefficients and slopes. Relationships on the same row were generated by the same underlying function ( $\hat{y} = 15 + x$  and  $\hat{y} = 15 + 5x$ , respectively) but are characterised by different correlation coefficients ( $r = 0.94$  and  $r = 0.59$ , respectively). The inverse is true for relationships in the same column.

2001; Flege et al., 1999; Johnson, 1992; Johnson & Newport, 1989; see Marinova-Todd, Marshall & Snow, 2000, p. 19, for additional references). Greater variability in UA with increasing age is likely due to factors other than age proper (Hyltenstam & Abrahamsson, 2003), such as the concomitant greater variability in exposure to literacy, degree of education, motivation and opportunity for language use, and by itself represents evidence neither in favour of nor against the CPH.

### Regression approaches

Having demonstrated that neither group mean or proportion comparisons nor correlation coefficient comparisons can directly address the ‘flattened slope’ prediction, we turn to the studies in which regression models were computed with AOA as a predictor variable and UA as the outcome variable. Once again, this category of studies is not mutually exclusive with the two categories discussed above.

In a large-scale study using self-reports and approximate AOAs derived from a sample of the 1990 U.S. Census, Stevens (1999) found that the probability with which immigrants from various countries stated that they spoke English ‘very well’ decreased curvilinearly as a function of AOA. Stevens (1999) noted that this development is similar to the pattern found by Johnson and Newport (1989) but that it contains no indication of an “abruptly defined ‘critical’ or sensitive period in L2 learning” (p. 569). We point out that she modelled the self-ratings using an ordinal logistic regression model in which the AOA variable was logarithmically transformed. Technically, this is perfectly fine. One should, however, be careful not to read too much into the non-linear curves found. In logistic models, the outcome variable itself is modelled linearly as a function of the predictor variables and is expressed in log-odds. In order to compute the corresponding probabilities, these log-odds are transformed using the logistic function. Consequently, even if the model is specified linearly, the predicted probabilities will not lie on a perfectly straight line when plotted as a function of any one continuous predictor variable. Similarly, when the predictor variable is first logarithmically transformed and then used to linearly predict an outcome variable, the function linking the predicted outcome variables and the untransformed predictor variable is necessarily non-linear. Thus, non-linearities follow naturally from Stevens’s (1999) model specifications. Moreover, CPH-consistent discontinuities in the AOA–UA function *cannot* be found using her model specifications as they did not contain any parameters allowing for this.

Using data similar to Stevens’s (1999), Bialystok and Hakuta (1999) found that the link between the self-rated English competences of Chinese- and Spanish-speaking immigrants and their AOA could be described by a straight line. In contrast to Stevens (1999), Bialystok and Hakuta (1999) used a regression-based method allowing for changes in the function’s slope, viz. locally weighted scatterplot smoothing (LOWESS). Informally, LOWESS is a non-parametrical method that relies on an algorithm that fits the dependent variable for small parts of the range of the independent variable whilst guaranteeing that the overall curve does not contain sudden jumps (for technical details, see Cleveland, 1979). Hakuta et al. (2003) used an even larger sample from the same 1990 U.S. Census data on Chinese- and Spanish-speaking immigrants (2.3 million observations). Fitting LOWESS curves, no discontinuities in the AOA–UA slope could be detected. Moreover, Hakuta et al. (2003) found that piecewise linear regression models, i.e. regression models containing a parameter

that allows a sudden drop in the curve or a change of its slope, did not provide a better fit to the data than did an ordinary regression model without such a parameter.

Summarising, Bialystok and Hakuta (1999) and Hakuta et al. (2003) found no evidence supporting a CPH account for the AOA–self-ratings relationship. The pertinence of these studies to the CPH has, however, been questioned for a number of reasons. These concern (a) the exclusion of immigrants who reported that they only spoke English at home from the data set (Stevens, 2004), (b) the possibility that the immigrants *believed* that second-language competence decreases monotonically as a function of age of learning and that the self-ratings are shaped by this belief (DeKeyser, 2006; DeKeyser & Larson-Hall, 2005), (c) the coarseness of the AOA variable retrieved from the census (DeKeyser, 2006; Stevens, 2004), and (d) the assumption that the self-ratings could be considered a continuous variable (Stevens, 2004). We recognise the potential of all four points to obscure a CP effect in the AOA–UA function. That said, we fail to grasp another point of Stevens’s (2004) criticism of Hakuta et al.’s (2003) study. This point concerns the use of comparing simple linear regression fits to fits of piecewise linear regressions. She argues that since the AOA–proficiency relationship is negative when viewed over the whole lifespan, there is hardly any variance left to be explained by the breakpoints. This is, of course, the whole point of the enterprise: parsimony dictates that if the breakpoints do not add sufficiently to the model fit, they should be left out! That said, the necessity of including a breakpoint in the model can be assessed by means other than the coefficient of determination ( $R^2$ ), e.g. relative goodness-of-fit measures such as the Akaike Information Criterion (Akaike, 1974) or the Bayesian Information Criterion (Schwarz, 1978) or  $F$ -tests. Such measures can in principle indicate better model fits even if the increase in  $R^2$  is minimal.

To our knowledge, regression models capable of highlighting non-linearities have only been modelled in two studies looking into the relationship between AOA and UA variables extracted using tasks rather than self-ratings. Flege et al. (1999) measured UA in English for 240 Korean participants using foreign-accent ratings and a grammaticality judgement task (GJT). They fitted both linear and cubic functions to the AOA–UA data. The cubic function explained somewhat more variance than did the linear function for the foreign-accent ratings (increase in  $R^2$ : 1.9 %), but follow-up analyses failed to find support for a non-linearity in puberty. A cubic function likewise explained somewhat more variance compared to a linear function for the GJT scores (increase in  $R^2$ : 1.2 %), but this time follow-up analyses revealed a change in slope an AOA of about 12 years. In our opinion, however, Flege et al.’s (1999) follow-up analyses are not quite ideal as they entail fitting models on AOA-defined subsets and checking whether the cubic term still contributed significantly to the model fit in those subsets; we refer the reader to the original publication for details on this procedure.<sup>7</sup>

Instead, we prefer the analytical approach used by Birdsong and Molis (2001), who, like Hakuta et al. (2003), fitted piecewise linear regression models and checked whether the breakpoint parameter contributed enough to the model to offset the resultant loss of

---

<sup>7</sup>Moreover, pinpointing the location of a slope change in a cubic function is mathematically speaking impossible: the function’s slope changes continuously (expressed by the first derivative, which itself is a continuous quadratic function) as does the rate by which it changes (expressed by the second derivative, which is a continuous linear function). One could pinpoint the AOA at which the change in slope starts to slow down or speed up (i.e. the point at which the sign of the second derivative changes), but one should be aware that one is dealing with a continuous phenomenon.

parsimony. Birdsong and Molis’s (2001) study was a replication of Johnson and Newport’s (1989) but used Spanish L1 speakers ( $n = 61$ ) rather than Korean- and Chinese-speaking participants. These authors found a breakpoint in the AOA–UA slope that contributed significantly to the model fit, but this breakpoint was located at AOA 27.5 years—well beyond a putative critical period. Reanalysing Johnson and Newport’s (1989) data, the authors further found that a breakpoint could improve the model fit for this data set, too. This time, however, the breakpoint was located at AOA 18 years. Importantly, the breakpoints had different functions in the two data sets: whereas it marked the beginning of a flatter part of the curve in the Johnson and Newport (1989) data set (as in the left panel of Figure 1), it actually marked the onset of a *steeper* part of the curve in the Birdsong and Molis (2001) study (as in the middle panel of Figure 1). In other words, the age effect in UA actually became more pronounced for the older arrivals. Birdsong and Molis (2001) did not mention by how much  $R^2$  increased when breakpoint parameters were included in their models.

To sum up, we have argued at length that regression approaches are superior to group mean and correlation coefficient comparisons for the purposes of testing the ‘flattened slope’ prediction. Acknowledging the reservations vis-à-vis self-estimated UAs, we still find that while the relationship between AOA and UA is not necessarily perfectly linear in the studies discussed, the data do not lend unequivocal support to this prediction. In the following section, we will reanalyse data from a recent empirical paper on the CPH by DeKeyser et al. (2010).

The first goal of this reanalysis is to further illustrate some of the statistical fallacies encountered in CPH studies. Second, by making our computer code available we hope to demonstrate how the relevant regression models, viz. piecewise regression models, can be fitted and how the AOA representing the optimal breakpoint can be identified. Lastly, the findings of this reanalysis will contribute to our understanding of how AOA affects UA as measured using a GJT.

### Reanalysis of DeKeyser et al. (2010)

We chose to reanalyse a recent empirical paper on the CPH by DeKeyser et al. (2010; henceforth DK et al.). This paper lends itself well to a reanalysis since it exhibits two highly commendable qualities: the authors spell out their hypotheses lucidly and provide detailed numerical and graphical data descriptions. Moreover, the paper’s lead author is very clear on what constitutes a necessary condition for accepting the CPH: a non-linearity in the age of onset of acquisition (AOA)–ultimate attainment (UA) function, with UA declining less strongly as a function of AOA in older, post-CP arrivals compared to younger arrivals (DeKeyser & Larson-Hall, 2005; DeKeyser, 2012). Lastly, it claims to have found cross-linguistic evidence from two parallel studies backing the CPH and should therefore be an unsuspected source to CPH proponents.

DK et al. present data from comparable investigations into the relationship between AOA and UA in morphosyntactic judgements in two groups of adult Russian-speaking immigrants who had started learning English ( $n = 76$ ) or Hebrew ( $n = 62$ ) as an L2 at different ages in North America and Israel, respectively. Participants were split up into three AOA groups: those who emigrated before the age of 18 (young), those between the ages of 18 and 40 (middle) and those who arrived after age 40 (old). The task used was a 200-item oral

grammaticality judgement task (GJT). In addition, all participants took a verbal aptitude test. For further details, we refer to the original publication.

The authors set out to test the following hypotheses:

Hypothesis 1: For both the L2 English and the L2 Hebrew group, the slope of the age of arrival–ultimate attainment function will not be linear throughout the lifespan, but will instead show a marked flattening between adolescence and adulthood.

Hypothesis 2: The relationship between aptitude and ultimate attainment will differ markedly for the young and older arrivals, with significance only for the latter. (DK et al., p. 417)

Both hypotheses were purportedly confirmed, which in the authors' view provides evidence in favour of CPH. The problem with this conclusion, however, is that it is based on a comparison of correlation coefficients. As we have argued above, correlation coefficients are not to be confused with regression coefficients and cannot be used to directly address research hypotheses concerning slopes, such as Hypothesis 1. In what follows, we will reanalyse the relationship between DK et al.'s AOA and GJT data in order to address Hypothesis 1. Additionally, we will lay bare a problem with the way in which Hypothesis 2 was addressed. The extracted data and the computer code used for the reanalysis are provided as supplementary materials, allowing anyone interested to scrutinise and easily reproduce our whole analysis and carry out their own computations (see appendix).

## Data extraction

DK et al. provided high-resolution scatterplots, downloadable from the journal's website, to illustrate the relationship between AOA and GJT performance. Using the open source program `g3data`<sup>8</sup>, we extracted the data underlying these scatterplots. For the Israel study, we chose to round off the AOA data to the nearest integer, as was the case in the North America study, rather than to the first decimal, as in the original.

In order to verify whether we did in fact extract the data points to a satisfactory degree of accuracy, we computed summary statistics for the extracted AOA and GJT data and checked these against the descriptive statistics provided by DK et al. (p. 421, Table 1, and p. 427, Table 3). As Table 1 of the present paper shows, the extracted data are virtually identical to the original data. In addition, we computed the correlation coefficients for the AOA–GJT relationship for the whole AOA range and for AOA-defined subgroups and checked these coefficients against those reported by DK et al. (p. 423 and p. 428). Again, the numbers provided by DK et al. and those computed by us were highly similar, as Table 2 shows. We conclude that we did indeed manage to extract the AOA and GJT data for both the North America and the Israel study to a high degree of accuracy.

---

<sup>8</sup>Downloadable from <https://github.com/pn2200/g3data>. After identifying the  $x$ - and  $y$ -axes, users can just click on the data points presented in a scatterplot, thereby extracting the  $x$ - and  $y$ -coordinates of the selected points.

Table 1

*Descriptive statistics for the North America and Israel studies as reported by DK et al. (2010, p. 421, Table 1 and p. 427, Table 3) and for the data extracted from DK et al.'s scatterplots.*

			range	mean	SD
North America	DK et al.	AOA	5–71	32.54	18.01
		GJT	104–198	150.76	27.32
	Extracted data	AOA	5–71	32.54	18.01
		GJT	104–198	150.76	27.32
Israel	DK et al.	AOA	4.12–65.2	30.57	16.94
		GJT	101–196	149.58	26.33
	Extracted data	AOA <sup>a</sup>	4–65	30.55	16.95
		GJT	101–196	149.58	26.33

<sup>a</sup>The extracted AOA data for the Israel study were rounded off to the nearest integer.

Table 2

*Correlation coefficients in the North America and Israel studies for the relationship between AOA and GJT over the whole AOA range and within subgroups defined by AOA as reported by DK et al. (2010, pp. 423 and 428) and computed for the extracted data. Figures between brackets represent the number of participants in each cell.*

		overall	young	middle	old
North America	DK et al.	-.80 (76)	-.69 (20)	-.44 (26)	-.27 (30)
	Extracted data	-.80 (76)	-.69 (20)	-.45 (26)	-.27 (30)
Israel	DK et al.	-.79 (62)	-.48 (17)	-.37 (32)	-.53 (13)
	Extracted data <sup>a</sup>	-.79 (62)	-.46 (17)	-.37 (32)	-.54 (13)

<sup>a</sup>The extracted AOA data for the Israel study were rounded off to the nearest integer.

### **Modelling the link between age of onset of acquisition and ultimate attainment**

We first replotted the AOA and GJT data we extracted from DK et al.'s scatterplots and added non-parametric scatterplot smoothers in order to investigate whether any changes in slope in the AOA–GJT function could be revealed, as per Hypothesis 1. Figures 3 and 4 show this not to be the case. Indeed, simple linear regression models that model GJT as a function of AOA provide decent fits for both the North America and the Israel data, explaining 65 % and 63 % of the variance in GJT scores, respectively. The parameters of these models are given in Table 3.

These straightforward analyses do not reveal any support for Hypothesis 1. Still, for the sake of completeness, let us turn to the issue of determining whether slope differences between AOA groups, if they do indeed exist, are substantial enough to invoke a critical period. What needs to be established is whether including multiple slopes in a model contributes sufficiently to the fit of the model to the data to offset the loss of parsimony associated with a simpler one-slope model. To this end, we computed linear regression models allowing of breakpoints in the regression slope ('piecewise regression models'). Similarly to ordinary

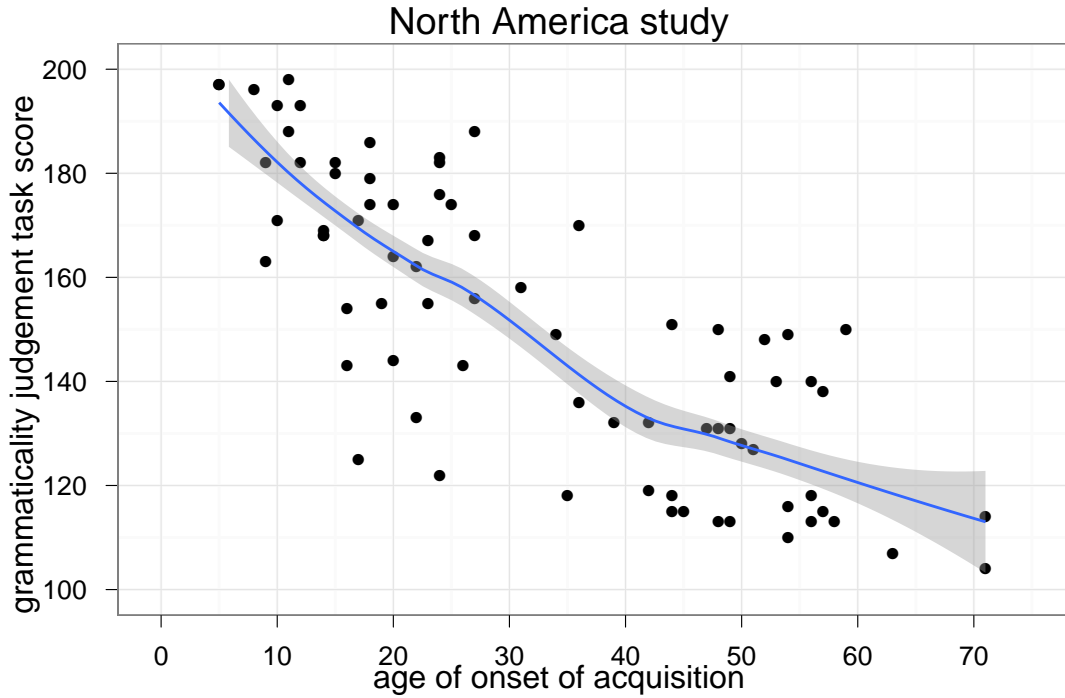


Figure 3. Scatterplot of the AOA–GJT relationship in the North America study with a non-parametric scatterplot smoother added. The scatterplot itself is a near-perfect replication of DK et al.’s Fig. 1.

regression, piecewise regression models the outcome variable  $y$  as a function of an overall intercept  $\alpha$  and a slope parameter  $\beta$  linking it to a predictor variable  $x$ . In contrast to ordinary regression, however, the  $\beta$  parameter of a piecewise regression model changes as a function of a binary indicator variable, which indicates whether  $x$  lies before or beyond the breakpoint to be modelled,  $\psi$ :

$$\hat{y} = \begin{cases} \alpha + \beta_1 x & \text{if } x \leq \psi \\ \alpha + \beta_2 x & \text{if } x > \psi \end{cases} \quad (4)$$

To ensure that both segments are joined at the breakpoint, the predictor variable is first centred at the breakpoint value, i.e. the breakpoint value is subtracted from the

Table 3

Linear regression models containing no breakpoints. GJT is modelled as a function of AOA. For ease of comparison with the breakpoint models, AOA was centred at 18 years.

	intercept $\pm$ SE	slope $\pm$ SE	$R^2$	$F$ -test of model fit
North America	168.50 $\pm$ 2.42	−1.22 $\pm$ 0.10	0.65	$F(1, 74) = 135.3, p < 0.001$
Israel	165.00 $\pm$ 2.57	−1.23 $\pm$ 0.12	0.63	$F(1, 60) = 100.4, p < 0.001$



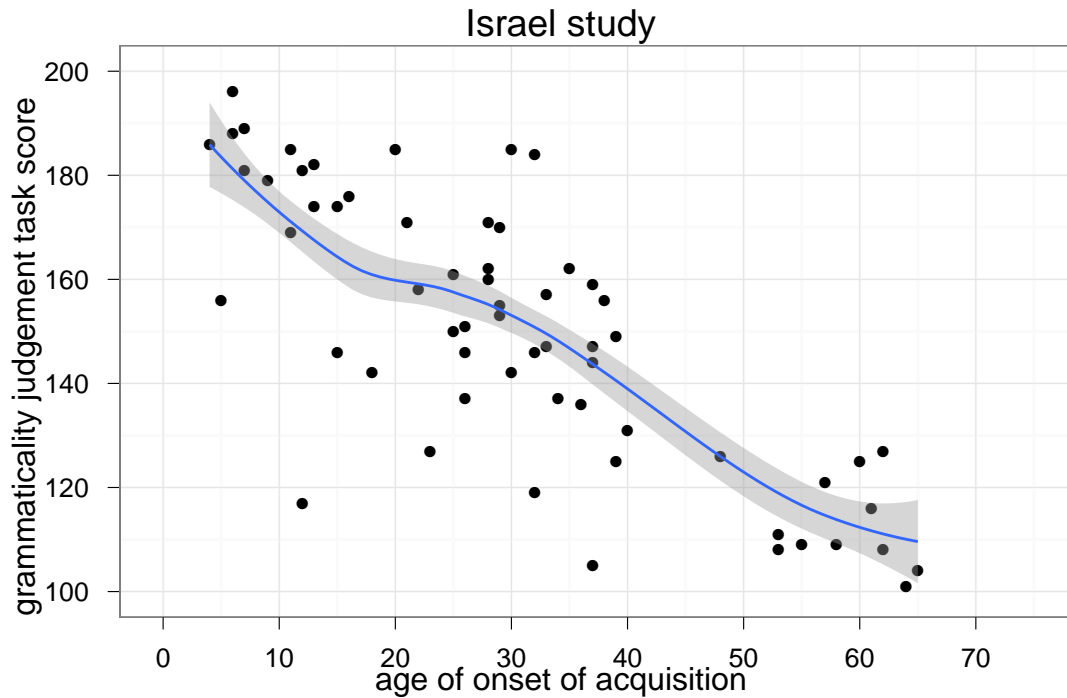


Figure 4. Scatterplot of the AOA–GJT relationship in the Israel study with a non-parametric scatterplot smoother added. The scatterplot itself is a near-perfect replication of DK et al.’s Fig. 5.

original predictor variable values. For a blow-by-blow account of how such models can be fitted in R, see Baayen (2008, pp. 214–222).

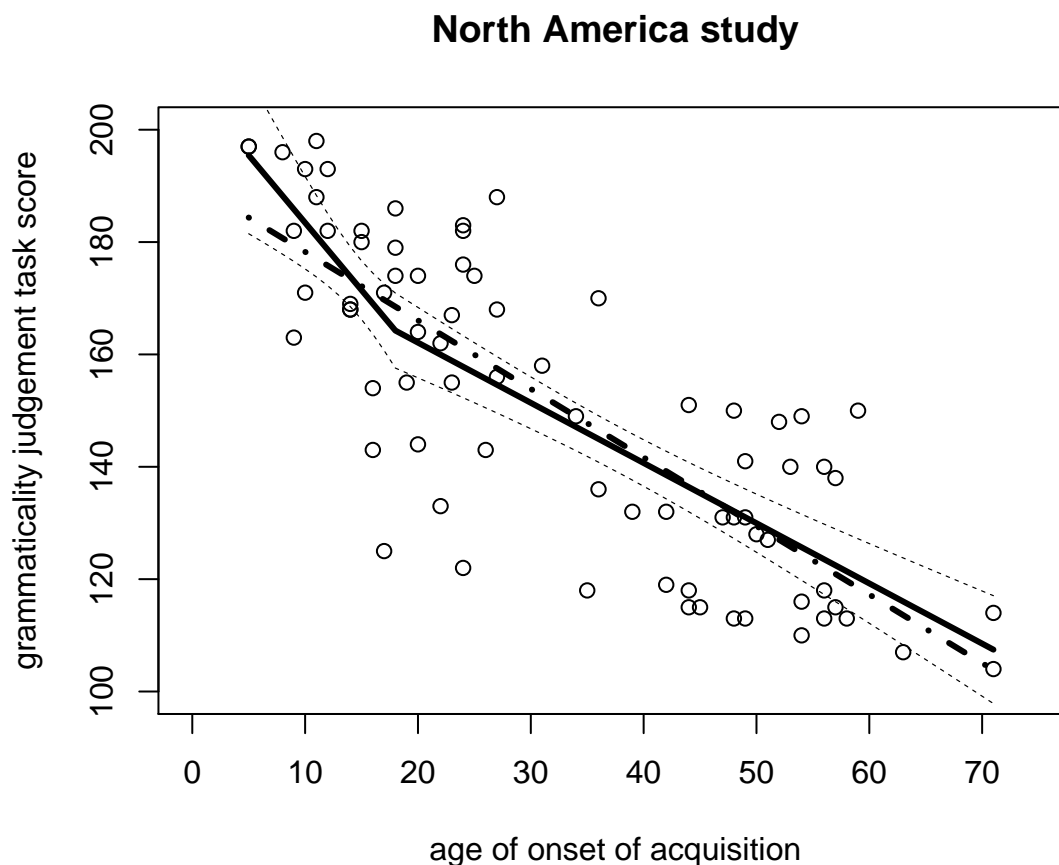
For the first models, we set the breakpoint at AOA 18, the cut-off used by DK et al. The models’ details are presented in Table 4.

At first glance, the coefficients for the North America study in Table 4 might appear to confirm Hypothesis 1: the slope linking AOA and GJT is flatter for participants with  $AOA > 18$  than for those with  $AOA \leq 18$ . However, as Figure 5 illustrates, the regression line of a model without a breakpoint falls well within the 95 % confidence interval of the regression line of the breakpoint model. Thus, the breakpoint parameter may be superfluous. For the Israel data, the change in slope at the breakpoint is hardly perceptible and the regression lines

Table 4

Linear regression models containing breakpoints at AOA 18. GJT is modelled as a function of AOA. Following Baayen (2008, pp. 214–222), AOA was centred at 18 years.

	intercept $\pm$ SE	slope $\pm$ SE (AOA $\leq$ 18)	slope $\pm$ SE (AOA $>$ 18)	$R^2$	$F$ -test of model fit
North America	164.24 $\pm$ 3.35	−2.40 $\pm$ 0.66	−1.07 $\pm$ 0.13	0.66	$F(2, 73) = 71.4, p < 0.001$
Israel	165.07 $\pm$ 3.90	−1.21 $\pm$ 0.62	−1.23 $\pm$ 0.17	0.63	$F(2, 59) = 49.4, p < 0.001$

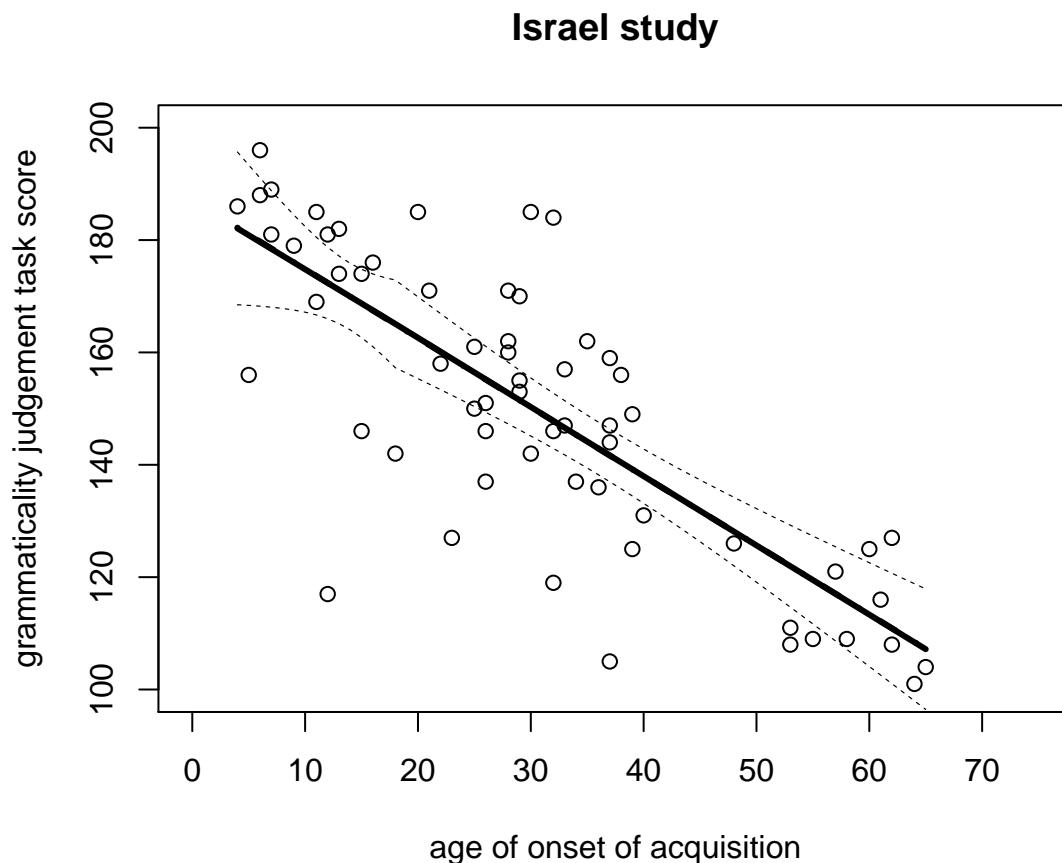


*Figure 5.* Regression lines for the North America data. Solid: regression with breakpoint at AOA 18 (dashed lines represent its 95 % confidence interval); dot-dash: regression without breakpoint.

plotted in Figure 6 overlap almost completely. In both cases, the inclusion of a breakpoint parameter at AOA 18 leads to an at most negligible increase in variance accounted for ( $R^2$ ). Unsurprisingly, then, formal  $F$ -tests confirm that the simpler models, i.e. the ones that do not include breakpoints at AOA 18, are to be preferred on the grounds of parsimony (North America:  $F(1, 73) = 3.27$ ,  $p = 0.07$ ; Israel:  $F(1, 59) = 0.15$ ,  $p = 0.98$ ).<sup>9</sup>

Having ascertained that the inclusion of a breakpoint in the regression to mark the end of a putative critical period at AOA 18 years does not improve the fit of the model to the data, we are still left with the possibility that placing the breakpoint *elsewhere* might do so (see our discussion on the CPH's scope above). Following Baayen (2008), we computed

<sup>9</sup>These  $F$ -tests compare whether the residual sums of squares associated with the more complex model is smaller than the residual sums of squares associated with the simpler model. As such, they are one-tailed tests. Halving the  $p$ -value of the North America model comparison to take into account that DK et al. predicted the direction of the change in slope at the breakpoint and thereby achieving significance at  $\alpha = 0.05$  is therefore unsound.



*Figure 6.* Regression lines for the Israel data. Solid: regression with breakpoint at AOA 18 (dashed lines represent its 95 % confidence interval); dot-dash (hardly visible due to near-complete overlap): regression without breakpoint.

a series of regression models for both data sets in which the position of the breakpoints varied between AOA 5 and 19 years. Breakpoints at AOA  $< 5$  could not be fitted for lack of data points and breakpoints at AOA  $> 19$  would hardly be indicative of a critical period that ends before maturity. For each fitted model, the deviance ( $d^2$ ) was computed, i.e. the sum of the squared differences between the actual data points and the values predicted by the model ( $d^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ). The smaller the deviance, the better the model fits the data. Thus, the optimal breakpoint is the one that results in the model with the smallest deviance. The fitted breakpoints and their associated deviances are plotted in Figure 7. For the North America study, a breakpoint at AOA 16 years is optimal; for the Israel study, the optimal breakpoint lies at AOA 6 years.

We refitted the regression models with and without breakpoints at AOA 16 (North America) and 6 (Israel). The models' details are presented in Tables 5 and 6. The conclusions are largely, though not completely identical compared to when the breakpoints were fixed at AOA 18. For the North America study, the slope flattens after the break-

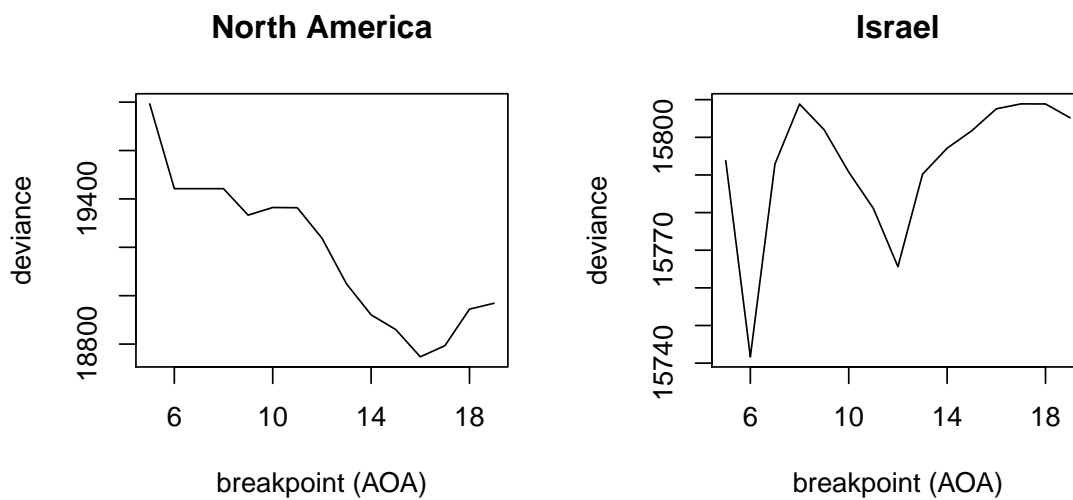


Figure 7. Deviances for regression models assuming breakpoints as a function of the position of the breakpoints.

point, but as Figure 8 shows, the regression line for a model without a breakpoint still falls entirely within the 95 % confidence interval of the breakpoint model and the increase in  $R^2$  is small. Nonetheless, an  $F$ -test returns a borderline significant  $p$ -value, which can be taken as support for the breakpoint model ( $F(1, 73) = 4.07, p = 0.047$ ).<sup>10</sup> For the Israel study, both regression lines again show almost complete overlap (Figure 9) and the increase in  $R^2$  is negligible. Unsurprisingly, an  $F$ -test yields no support for the breakpoint model ( $F(1, 59) = 0.25, p = 0.62$ ).

In sum, a regression model that allows for changes in the slope of the the AOA–GJT function to account for putative critical period effects provides a somewhat better fit to the North American data than does an everyday simple regression model. The improvement in model fit is marginal, however, and including a breakpoint does not result in any detectable improvement of model fit to the Israel data whatsoever. Breakpoint models therefore fail to provide solid crosslinguistic support in favour of critical period effects: across both data sets, GJT can satisfactorily be modelled as a linear function of AOA.<sup>11</sup>

<sup>10</sup>Two other relative goodness-of-fit measures likewise produce borderline results: the breakpoint model has a slightly better (i.e. lower) Akaike Information Criterion (Akaike, 1974) value than the simpler model (642.3 and 644.4, respectively), but a slightly worse (i.e. higher) Bayesian Information Criterion (Schwarz, 1978) value (651.6 and 651.4, respectively).

<sup>11</sup>Ideally, regression models are fitted on homoscedastic data, meaning that the variance around the model's predictions does not vary as a function of the value of those predictions. This condition is not fully met in the present data sets and, indeed, heteroscedasticity seems to be endemic in research on age effects: UA variance is typically larger in the older age groups. We therefore fitted robust regression models (see Fox, 2002) both with and without breakpoints using the `r1m()` function in the `MASS` package for R (Ripley, Hornik, Gebhardt & Firth, 2012). The parameters of these models were highly similar to those of their parametric counterparts (see supplementary materials).

Table 5

*Regression models without and with breakpoints at AOA 16 for the North America data. GJT is modelled as a function of AOA. Following Baayen (2008, pp. 214–222), AOA was centred at 16 years.*

(a) Regression model without breakpoint				
intercept $\pm$ SE	slope $\pm$ SE	$R^2$	$F$ -test of model fit	
$170.94 \pm 2.56$	$-1.22 \pm 0.10$	0.65	$F(1, 74) = 135.3, p < 0.001$	
(b) Regression model with breakpoint at AOA 16				
intercept $\pm$ SE	slope $\pm$ SE (AOA $\leq$ 16)	slope $\pm$ SE (AOA $>$ 16)	$R^2$	$F$ -test of model fit
$166.69 \pm 3.27$	$-2.86 \pm 0.82$	$-1.08 \pm 0.12$	0.67	$F(2, 73) = 72.5, p < 0.001$

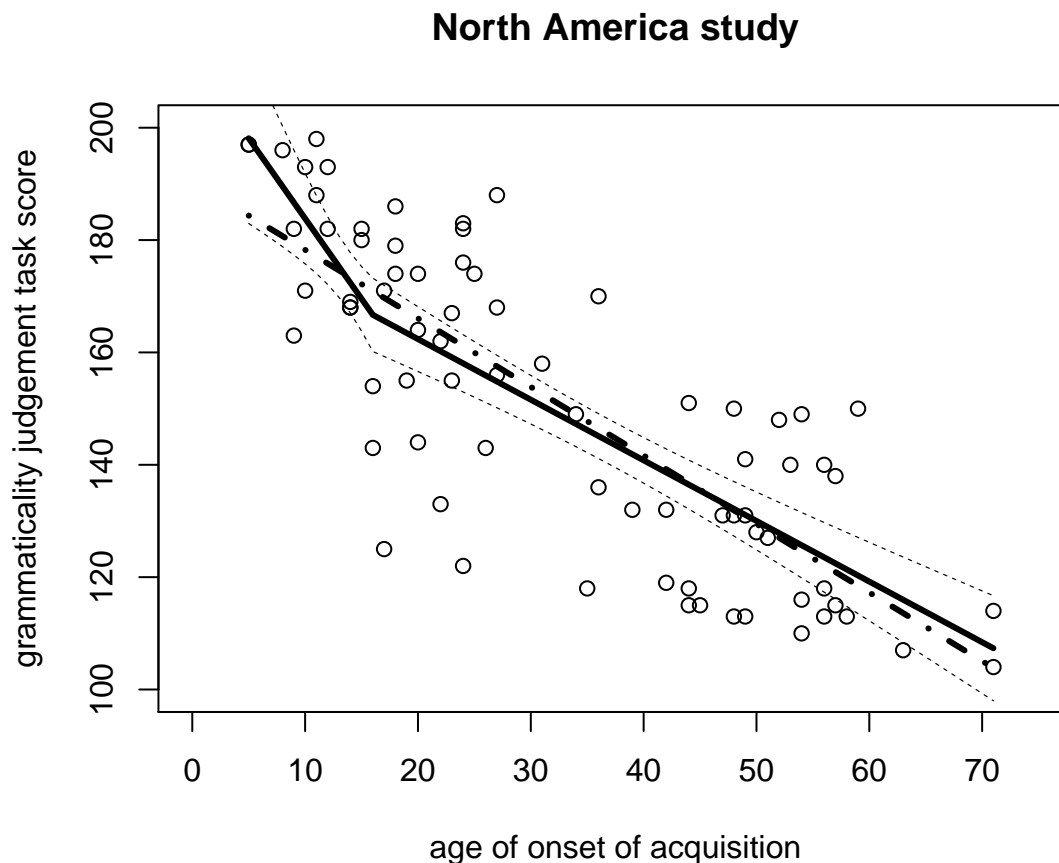
Table 6

*Regression models without and with breakpoints at AOA 6 for the Israel data. GJT is modelled as a function of AOA. Following Baayen (2008, pp. 214–222), AOA was centred at 6 years.*

(a) Regression model without breakpoint				
intercept $\pm$ SE	slope $\pm$ SE	$R^2$	$F$ -test of model fit	
$179.75 \pm 3.65$	$-1.23 \pm 0.12$	0.63	$F(1, 60) = 100.4, p < 0.001$	
(b) Regression model with breakpoint at AOA 6				
intercept $\pm$ SE	slope $\pm$ SE (AOA $\leq$ 6)	slope $\pm$ SE (AOA $>$ 6)	$R^2$	$F$ -test of model fit
$180.37 \pm 3.87$	$2.62 \pm 7.67$	$-1.25 \pm 0.13$	0.63	$F(2, 59) = 49.7, p < 0.001$

### Partialling out age at testing

As we have argued above, correlation coefficients cannot be used to test hypotheses about slopes. When the correct procedure is carried out on DK et al.'s data, no changes in the AOA–GJT function could be detected. In addition to comparing the zero-order correlations between AOA and GJT, however, DK et al. computed partial correlations in which the variance in AOA associated with the participants' age at testing (AAT; a potentially confounding variable) was filtered out. They found that these partial correlations between AOA and GJT, which are given in Table 7, differed between age groups in that they are stronger for younger than for older participants. This, DK et al. argue, constitutes additional evidence in favour of the CPH. At this point, we can no longer provide our own analysis of DK et al.'s data, seeing as the pertinent data points were not plotted. Nevertheless, the detailed descriptions by DK et al. strongly suggest that the use of these partial correlations is highly problematic. Most importantly, and to reiterate, correlations (whether zero-order or partial ones) are actually of no use when testing hypotheses concerning slopes. Still, one may wonder why the partial correlations differ across age groups. Our surmise is that these



*Figure 8.* Regression lines for the North America data. Solid: regression with breakpoint at AOA 16 (dashed lines represent its 95 % confidence interval); dot-dash: regression without breakpoint.

differences are at least partly the by-product of an imbalance in the sampling procedure.

As indicated in DK et al.'s Tables 2 and 4 (pp. 424 and 430), AOA and AAT are correlated to the point of near-unity ( $r = 0.97$  and  $0.98$ ). AAT and GJT, too, are highly correlated ( $r = -0.78$  and  $-0.77$ ) and the AAT-GJT correlations are nearly identical in magnitude to the AOA-GJT correlations ( $r = -0.80$  and  $-0.79$ ). In other words, AOA and AAT essentially represent the same variable when the whole AOA range is considered. However, DK et al. did not compute their partial correlations on the basis of the whole AOA continuum. Rather, they looked at each AOA slice separately. Crucially, however, the AOA-AAT correlation was not constant across AOA groups. In the North America study, for instance, AOA and AAT correlated less strongly in the youngest arrivals ( $r = 0.41$ ) than in the older arrivals ( $r = 0.88$  and  $0.83$ ); in the Israel study, the correlations were rather more comparable ( $r = 0.79$ ,  $0.88$  and  $0.98$ ). What DK et al. did not take into account is that it is these differences in the strength of the AOA-AAT correlation, which for the purposes of testing the CPH are completely uninteresting, that are largely responsible for the differences

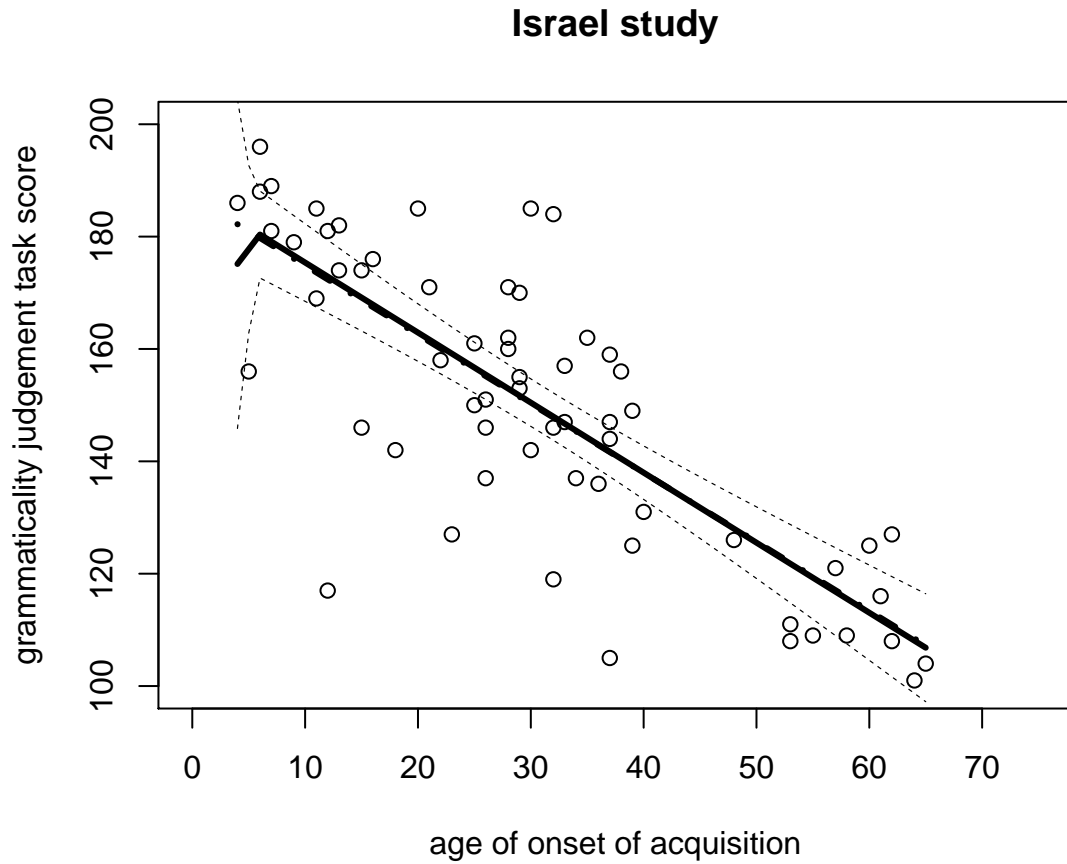


Figure 9. Regression lines for the Israel data. Solid: regression with breakpoint at AOA 6 (dashed lines represent its 95 % confidence interval); dot-dash (hardly visible due to near-complete overlap): regression without breakpoint.

Table 7

Partial correlation coefficients for the relationship between AOA and GJT with AAT controlled for. Coefficients for separate age groups as reported by DK et al., pp. 423 and 429; overall coefficients were computed on the basis of the data in DK et al., Tables 2 and 4. Figures between brackets represent the number of participants in each cell.

	overall	young	middle	old
North America	-.29 (76)	-.71 (20)	-.17 (26)	-.12 (30)
Israel	-.28 (62)	-.51 (17)	-.12 (32)	-.33 (13)

Table 8

*Correlation coefficients for the relationship between aptitude and GJT as reported by DK et al., pp. 425 and 429. Figures between brackets represent the number of participants in each cell.*

	overall	young	middle	old
North America	.21 (76)	.11 (20)	.44 (26)	.33 (30)
Israel	.00 (62)	-.37 (17)	.45 (32)	.14 (13)

in the strength of the partial correlations: partial correlations decrease as the correlations between the ‘independent’ variables increase (see Blalock, 1963, p. 235, Table 2). The partial correlation between  $x$  and  $y$  controlling for  $z$  ( $r_{xy.z}$ ) is computed solely on the basis of the underlying zero-order correlations  $r_{xy}$ ,  $r_{xz}$  and  $r_{yz}$ :

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \quad (5)$$

Decreasing partial correlations with increasing zero-order correlations between the independent variables ( $r_{xz}$ ) follow naturally from this function. This is the most straightforward explanation of why the differences in the partial correlations are smaller between all the groups in the Israel study compared to the North America study: the AOA–AAT correlations in the Israel study are high in all age groups, not just in two of them.

The upshot of this brief discussion is that the partial correlation differences reported by DK et al. are at least partly the result of an imbalance in the sampling procedure: AOA and AAT were simply less intimately tied for the young arrivals in the North America study than for the older arrivals with L2 English or for all of the L2 Hebrew participants. In an ideal world, we would like to fix AAT or ascertain that it at most only weakly correlates with AOA. This, however, would result in a strong correlation between AOA and another potential confound variable, length of residence, bringing us back to square one. Allowing for only moderate correlations between AOA and AAT might improve our predicament somewhat, but even in that case, we should tread lightly when making inferences on the basis of statistical control procedures (see Christenfeld, Sloan, Carroll & Greenland, 2004).

### Estimating the role of aptitude

Having shown that Hypothesis 1 could not be confirmed, we now turn to Hypothesis 2, which predicts a differential role of aptitude for UA in SLA in different AOA groups. More specifically, it states that the correlation between aptitude and GJT performance will be significant only for older arrivals. The correlation coefficients of the relationship between aptitude and GJT are presented in Table 8.

The problem with both the wording of Hypothesis 2 and the way in which it is addressed is the following: it is assumed that a variable has a reliably different effect in different groups when the effect reaches significance in one group but not in the other. This logic is fairly widespread within several scientific disciplines (see e.g. Nieuwenhuis, Forstmann & Wagenmakers, 2011, for a discussion). Nonetheless, it is demonstrably fallacious (see



Gelman & Stern, 2006). Here we will illustrate the fallacy for the specific case of comparing two correlation coefficients.

The  $p$ -value associated with a correlation coefficient is solely a function of three factors: (a) the strength of the correlation, i.e.  $r$ , (b) the number of pairs of correlated observations, i.e.  $n$ , and (c) whether one wishes to conduct a one-tailed or a two-tailed test.<sup>12</sup> Thus, a two-tailed test will reveal that a correlation coefficient of 0.28 for 50 pairs of observations is significant at the 0.05 level ( $p = 0.049$ ), whereas a coefficient of 0.27 fails to reach significance ( $p = 0.06$ ). But it would be clearly wrong to conclude that the miniscule difference between the two correlation coefficients therefore has to be significant, too. We take it that most researchers share this insight when  $p$ -values hover around the 0.05 threshold. Importantly, however, even more substantially different correlation coefficients are not necessarily ‘significantly different’ from one another either, not even if their associated  $p$ -values indicated that one were ‘highly significant’ (e.g.  $r = 0.40$ ,  $n = 50$ ,  $p < 0.01$ ) and the other were not even close to significance (e.g.  $r = 0.10$ ,  $n = 50$ ,  $p > 0.48$ ).

The correct procedure for comparing independent correlation coefficients (see Field, 2009, p. 191, and Olkin & Finn, 1995) consists of converting the correlation coefficients to  $z$ -scores using Fisher’s  $r$ -to- $z$  transformation (Eq. 6), computing the  $z$ -statistic of the difference between the two converted correlation coefficients (Eq. 7) and checking this  $z$ -statistic for significance. In two-tailed tests, the difference of two correlation coefficients reaches significance at  $\alpha = 0.05$  for  $z = 1.96$ ; in one-tailed tests for  $z = 1.64$ .

$$z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \quad (6)$$

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \quad (7)$$

Plugging in the numbers for the substantially different correlation coefficients of 0.10 and 0.40 introduced above, we find that the difference between these coefficients is actually not significant ( $z = 1.57$ ). When the correct method is applied to DK et al.’s correlation coefficients, comparisons of the young arrivals against the older arrivals only revealed a single difference in correlation strength, even when carrying out one-tailed tests: the correlation between aptitude and GJT in young arrivals in Israel ( $r = -0.37$ ,  $n = 17$ ) differs significantly from the one in middle-aged arrivals ( $r = 0.45$ ,  $n = 32$ ;  $z = 2.68$ , one-tailed  $p < 0.01$ ). For all other comparisons,  $z$  varied between 0.46 and 1.28.

Apart from not being replicated in the North America study, does this difference actually show anything? We contend that it does not: what is of interest are not so much the correlation coefficients, but rather the interactions between AOA and aptitude in models predicting GJT. These interactions could be investigated by fitting a multiple regression model in which the postulated CP breakpoint governs the slope of both AOA and aptitude. If such a model provided a substantially better fit to the data than a model without a breakpoint for the aptitude slope *and* if the aptitude slope changes in the expected direction

<sup>12</sup>Given  $r$  and  $n$ , a  $t$ -statistic can be calculated using the formula  $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ . A two-tailed  $t$ -test reaches significance at  $\alpha = 0.05$  at around  $t = 2$ , though the precise figure varies as a function of  $n$ .

(i.e. a steeper slope for post-CP than for younger arrivals), only then would this particular prediction of the CPH be borne out.

## Discussion

Using data extracted from a paper reporting on two recent studies that purport to provide evidence in favour of the CPH and that, according to its authors, represent a major improvement over earlier studies (DK et al., p. 417), we found that neither of its two hypotheses were actually confirmed when using the proper statistical tools. As a matter of fact, we found that GJT scores continue to decline at essentially the same rate even beyond the end of the putative critical period. According to the paper’s lead author, such a finding represents a serious problem to his conceptualisation of the CPH (DeKeyser & Larson-Hall, 2005, p. 97). Moreover, although modelling a breakpoint representing the end of a CP at AOA 16 may improve the statistical model slightly in study on learners of English in North America, the study on learners of Hebrew in Israel fails to confirm this finding. In fact, even if we were to accept the optimal breakpoint computed for the Israel study, it lies at AOA 6 and is associated with a different geometrical pattern.

Diverging age trends in similar studies with participants with different L2s have also been reported by Birdsong and Molis (2001) and are at odds with an L2-independent CPH. One parsimonious explanation of such conflicting age trends may be that the overall, crosslinguistic age trend is in fact linear, but that fluctuations in the data (due to factors unaccounted for or randomness) may sometimes give rise to a ‘stretched L’-shaped pattern (Figure 1, left panel) and sometimes to a ‘stretched 7’-shaped pattern (Figure 1, middle panel; see also Birdsong, 2009, for a similar comment).

Importantly, the criticism that DeKeyser and Larson-Hall (2005) levy against two studies reporting findings similar to ours (Bialystok & Hakuta, 1999; Stevens, 1999), viz. that the data consisted of self-ratings of questionable validity, does not apply to the present data set. In addition, DK et al. did not exclude any outliers from their analyses, so we assume that DeKeyser and Larson-Hall’s (2005) criticism of Birdsong and Molis’s (2001) study, i.e. that the findings were due to the influence of outliers, is not applicable to the present data either. For good measure, we refitted our regression models without breakpoints after excluding one potentially problematic data point in both studies.<sup>13</sup> The resultant models were virtually identical to the original models, see supplementary materials. Furthermore, the AOA variable was sufficiently fine-grained and the AOA–GJT curve was not ‘presmoothed’ by the prior aggregation of GJT across parts of the AOA range (see Stevens, 2004). Lastly, seven of the nine “problems with supposed counter-evidence” to the CPH discussed by Long (2005) do not apply either, viz. (1) “[c]onfusion of rate and ultimate attainment”, (2) “[i]nappropriate choice of subjects”, (3) “[m]easurement of AO”, (4) “[l]eading instructions to raters”, (6) “[u]se of markedly non-native samples making near-native samples more likely to sound native to raters”, (7) “[u]nreliable or invalid measures”, and (8) “[i]nappropriate L1–L2 pairings”. Problem No. 5 (“Assessments based on limited samples and/or “language-like” behavior”) may be apropos given that only GJT data were used, leaving open the

<sup>13</sup>The following data points had standardised residuals larger than 2.5 in the original models and were excluded when refitting the models: the participant with AOA 17 and a GJT score of 125 in the North America study and the participant with AOA 12 and a GJT score of 117 in the Israel study.

theoretical possibility that other measures might have yielded a different outcome. Finally, problem No. 9 (“Faulty interpretation of statistical patterns”) is, of course, precisely what we have turned the spotlights on.

### Overall discussion and conclusion

The critical period hypothesis remains a hotly contested issue in the psycholinguistics of second-language acquisition. Discussions about the impact of empirical findings on the tenability of the CPH generally revolve around the reliability of the data gathered (e.g. Bialystok, 2002; DeKeyser, 2006, 2012; DeKeyser & Larson-Hall, 2005; Long, 2005; Rothman, 2008) and such methodological critiques are of course highly desirable. Furthermore, the debate often centres on the question of exactly what version of the CPH is being vindicated or debunked. These versions differ mainly in terms of its scope, specifically with regard to the relevant age span, setting and language area, and the testable predictions they make. But even when the CPH’s scope is clearly demarcated and its main prediction is spelt out lucidly, the issue remains to what extent the empirical findings can actually be marshalled in support of the relevant CPH version. As we have shown in this paper, empirical data have often been taken to support CPH versions predicting that the relationship between age of acquisition and ultimate attainment is not strictly linear, even though the statistical tools most commonly used (notably group mean and correlation coefficient comparisons) were, crudely put, irrelevant to this prediction. Methods that we consider valid, e.g. piecewise regression and scatterplot smoothing, have been used in some studies (Bialystok & Hakuta, 1999; Birdsong & Molis, 2001; Hakuta et al., 2003), but these studies have been criticised on other grounds. To our knowledge, however, such methods have never been used by scholars who explicitly subscribe to the CPH.

We suspect that what may be going on is a form of ‘confirmation bias’ (Nickerson, 1998), a cognitive bias at play in diverse branches of human knowledge seeking: Findings judged to be consistent with one’s own hypothesis are hardly questioned, whereas findings inconsistent with one’s own hypothesis are scrutinised much more strongly and criticised on all sorts of points (Firebaugh, 2007; Fugelsang, Stein, Green & Dunbar, 2004; Koehler, 1993; Mynatt, Doherty & Tweney, 1977). Our reanalysis of DeKeyser et al.’s (2010) recent paper may be a case in point. CPH exponents used correlation coefficients to address their prediction about the slope of a function, as had been done in a host of earlier studies. Finding a result that squared with their expectations, they did not question the technical validity of their results, or at least they did not report this.<sup>14</sup> That said, we are keen to point out that the statistical analyses in this particular paper, though suboptimal, are, as far as we could gather, reported correctly, i.e. the confirmation bias does not seem to result in the blatant misreportings found elsewhere (see Bakker & Wicherts, 2011, for empirical evidence and discussion). An additional point to these authors’ credit is that, apart from explicitly identifying their CPH version’s scope and making crystal-clear predictions, they present data descriptions that actually permit quantitative reassessments such as ours and

<sup>14</sup>Our reanalysis is actually a case in point in two respects: for an earlier draft of this paper we had computed the optimal position of the breakpoints incorrectly, resulting in an insignificant improvement of model fit for the North American data rather than a borderline significant one. Finding a result that squared with our expectations, *we* did not question the technical validity of *our* results—until this error was kindly pointed out to us by Martijn Wieling (University of Tübingen).

have a history of doing so (e.g. the appendix in DeKeyser, 2000). This leads us to believe that they analysed their data all in good conscience and to hope that they, too, will conclude that their own data do not, in fact, support their hypothesis.

We end this paper on an upbeat note. Even though we have argued that the analytical tools employed in CPH research generally leave much to be desired, the original data are, so we hope, still available. This provides researchers, CPH supporters and skeptics alike, with an exciting opportunity to reanalyse their data sets using the tools outlined in the present paper and publish their findings at minimal cost of time and resources. We therefore encourage scholars to engage their old data sets with an open mind and to communicate their analyses openly, e.g. by voluntarily publishing their data and computer code alongside their articles. Ideally, CPH supporters and skeptics would join forces to agree on a data collection and analysis protocol for a high-powered study in order to provide a truly convincing answer to a core issue in SLA.

### Supplementary materials

- `DeKeyser2010NorthAmerica.csv`: AOA and GJT data extracted from DeKeyser et al.'s (2010) North America study.
- `DeKeyser2010Israel.csv`: AOA and GJT data extracted from DeKeyser et al.'s (2010) Israel study.
- `ReanalysisDeKeyser2010.R`: Script with annotated R code used for the reanalysis and for generating Figures 3–9.

R (R Core Team, 2012) is an open source program and programming language for statistical computing. It can be downloaded freely from <http://www.r-project.org/>. The file `ReanalysisDeKeyser2010.R` can be opened within R or with any text editor. The `pwr.t2n.test()` function from the `pwr` package for R (Champely, 2009) was used for computing power levels in Abrahamsson and Hyltenstam's (2009) and Johnson and Newport's (1989) studies. For generating Figures 3 and 4, we used the `ggplot2` add-on package (Wickham, 2009). The `rlm()` function for fitting robust regression models is part of the `MASS` package (Ripley et al., 2012). All packages can be installed from within R.

## References

- Abrahamsson, N. & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249-306.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716-723.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Bakker, M. & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678.
- Bialystok, E. (2002). On the reliability of robustness: A reply to DeKeyser. *Studies in Second Language Acquisition*, *24*, 481-488.
- Bialystok, E. & Hakuta, K. (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (p. 161-181). Mahwah, NJ: Lawrence Erlbaum.
- Bialystok, E. & Miller, B. (1999). The problem of age in second-language acquisition: Influences from language, structure, and task. *Bilingualism: Language and Cognition*, *2*, 127-145.
- Birdsong, D. (Ed.). (1999). *Second language acquisition and the critical period hypothesis*. Mahwah, NJ: Lawrence Erlbaum.
- Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism. Psycholinguistic approaches* (p. 109-127). New York: Oxford University Press.
- Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning*, *56*, 9-49.
- Birdsong, D. (2009). Age and the end state of second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (p. 401-424). Bingley: Emerlad.
- Birdsong, D. & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language*, *44*, 235-249.
- Blalock, H. M., Jr. (1963). Correlated independent variables: The problem of multicollinearity. *Social Forces*, *42*, 233-237.
- Champely, S. (2009). *pwr: Basic functions for power analysis*. Retrieved from <http://cran.r-project.org/package=pwr> (R package, version 1.1.1)
- Christenfeld, N. J. S., Sloan, R. P., Carroll, D. & Greenland, S. (2004). Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine*, *66*, 868-875.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829-836.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.
- Cohen, J. (1994). The Earth is round ( $p < .05$ ). *American Psychologist*, *49*, 997-1003.

- Cook, V. J. (1992). Evidence for multicompetence. *Language Learning*, 42, 557-591.
- DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22, 499-533.
- DeKeyser, R. (2006). A critique of recent arguments against the critical period hypothesis. In C. Abello-Contesse, R. Chacón-Beltrán, M. D. López-Jiménez & M. M. Torreblanca-López (Eds.), *Age in L2 acquisition and teaching* (p. 49-58). Bern: Peter Lang.
- DeKeyser, R. (2012). Age effects in second language learning. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (p. 442-460). London: Routledge.
- DeKeyser, R., Alfi-Shabtay, I. & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, 31, 413-438.
- DeKeyser, R. & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism. Psycholinguistic approaches* (p. 88-108). New York: Oxford University Press.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.
- Firebaugh, G. (2007). Replication data sets and favored-hypothesis bias: Comment on Jeremy Freese (2007) and Gary King (2007). *Sociological Methods & Research*, 36, 200-209.
- Flege, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (p. 101-132). Mahwah, NJ: Lawrence Erlbaum.
- Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H. & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 34, 153-175.
- Flege, J. E., Yeni-Komshian, G. H. & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, 41, 78-104.
- Fox, J. (2002). *Robust regression. Appendix to An R and S-Plus Companion to Applied Regression*. Retrieved from <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix.html>
- Fugelsang, J. A., Stein, C. B., Green, A. E. & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, 58, 86-95.
- Gelman, A. & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60, 328-331.
- Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52, 647-674.
- Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, 36, 3-15.
- Hakuta, K., Bialystok, E. & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, 14, 31-38.
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, 3-7.
- Hyltenstam, K. & Abrahamsson, N. (2003). Maturation constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition*

- (p. 539-588). Malden, MA: Blackwell.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi: 10.1371/journal.pmed.0020124
- Johnson, J. S. (1992). Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence. *Language Learning*, *42*, 217-248.
- 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*, 60-99.
- Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28-55.
- Krashen, S. D., Long, M. H. & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly*, *13*, 573-582.
- Kroll, J. F. & De Groot, A. M. B. (Eds.). (2005). *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press.
- Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.
- Long, M. H. (2005). Problems with supposed counter-evidence to the critical period hypothesis. *International Review of Applied Linguistics in Language Teaching*, *43*, 287-317.
- Long, M. H. (2007). *Problems in SLA*. Mahwah, NJ: Lawrence Erlbaum.
- Marinova-Todd, S. H., Marshall, D. B. & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, *34*, 9-34.
- McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics*, *21*, 395-423.
- Muñoz, C. & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1-35.
- Mynatt, C. R., Doherty, M. E. & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, *29*, 85-95.
- Newport, E. L. (1991). Contrasting conceptions of the critical period for language. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (p. 111-130). Hillsdale, NJ: Lawrence Erlbaum.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175-220.
- Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105-1107.
- Olkin, I. & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155-164.
- Patkowski, M. S. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning*, *30*, 449-472.
- Penfield, W. & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton: Princeton University Press.
- R Core Team. (2012). *R: A language and environment for statistical computing*. Retrieved from <http://www.r-project.org/> (Software, version 2.15.2)
- Ripley, B., Hornik, K., Gebhardt, A. & Firth, D. (2012). *MASS: Support functions and datasets for Venables and Ripley's MASS*. Retrieved from <http://cran.r-project>



- .org/package=MASS (R package, version 7.3-17)
- Rothman, J. (2008). Why all counter-evidence to the critical period hypothesis in second language acquisition is not equal or problematic. *Language and Linguistics Compass*, 2, 1063-1088.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Ruben, R. J. (1997). A time frame of critical/sensitive periods of language development. *Acta Oto-laryngologica*, 117, 202-205.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.
- Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366. doi: 10.1177/0956797611417632
- Singleton, D. (2005). The critical period hypothesis: A coat of many colours. *International Review of Applied Linguistics in Language Teaching*, 43, 269-285.
- Singleton, D. (2007). The critical period hypothesis: Some problems. *Interlingüística*, 17, 48-56.
- Snow, C. E. & Hoefnagel-Höhle, M. (1977). Age differences in the pronunciation of foreign sounds. *Language and Speech*, 20, 357-365.
- Snow, C. E. & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, 49, 1114-1128.
- Stevens, G. (1999). Age of immigration and second language proficiency among foreign-born adults. *Language in Society*, 28, 555-578.
- Stevens, G. (2004). Using census data to test the critical-period hypothesis for second-language acquisition. *Psychological Science*, 15, 215-216.
- Weisstein, E. W. (n.d.). *Discontinuity*. From *MathWorld*—A Wolfram Web Resource. Retrieved from <http://mathworld.wolfram.com/Discontinuity.html> (Accessed March 23, 2012)
- Wetzels, R., Matzke, D., Lee, M. D., Rounder, J. N., Iverson, G. J. & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6, 291-298.
- White, L. & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research*, 12, 233-265.
- Wickham, H. (2009). *ggplot2: An implementation of the Grammar of Graphics*. Retrieved from <http://cran.r-project.org/package=ggplot2> (R package, version 0.8.9)