DOI:10.1037/14939-010

Chapter 9

Age of second language acquisition: Critical periods and social concerns

David Birdsong
University of Texas at Austin

Jan Vanhove
University of Fribourg

In E. Nicoladis and S. Montanari (Eds.), 2016, *Bilingualism across the lifespan: Factors moderating language proficiency*, 163-182.

APA and De Gruyter

David Birdsong, Department of French and Italian, University of Texas at Austin, 201 W. 21st Street, STOP B7600, Austin, TX 78712 USA, birdsong@austin.utexas.edu.

Jan Vanhove, Department of Multilingualism, University of Fribourg, Rue de Rome 1, CH-1700 Fribourg, Switzerland, jan.vanhove@unifr.ch.

Age of Second Language Acquisition: Critical Periods and Social Concerns

A classic topic in research on bilingualism across the lifespan is the relationship between the age at which learners start to acquire a second language (L2) and their ultimate level of proficiency in that language. Learning of an L2 that begins in infancy is typically associated with fluent speech, effortless language processing, and native accent. In contrast, late L2 learners tend to diverge from monolingual natives on measures of grammatical and lexical knowledge, processing speed, and acoustic properties of speech. Various classes of explanations for age effects in L2 acquisition – attitudinal, neurobiological, experiential, psycho-social, and cognitive – have been proposed in the literature. It is not the purpose of this chapter to examine these various accounts; for an overview, see Muñoz and Singleton (2011). Similarly, with the exception of a brief discussion in the section on social concerns, this chapter does not consider research on the efficiency of early foreign language teaching in schools; for an overview, see Lambelet and Berthele (2015). Rather, we are concerned with the hypothesis that L2 learning in naturalistic contexts is constrained by a critical period (CP). With its roots in the seminal works of Penfield and Roberts (1959) and Lenneberg (1967), the Critical Period Hypothesis for L2 acquisition (CPH/L2A) posits that nativelike attainment in the L2 from mere exposure is possible if learning begins within, but not after, a limited developmental span.

Several recent overviews comprehensively summarize studies inspired by the CPH/L2A (references below), and we do not intend to rehash these surveys. Instead, we aim to provide readers with a technical toolkit to critically evaluate research on the divisive issue that is the CPH/L2A. To this end, we first discuss the CPH/L2A's prediction that nativelikeness among learners with post-CP AoAs is impossible and highlight epistemological difficulties with this prediction. We then turn to the nature of the function that relates AoA to L2 attainment. The logic here is that a discontinuous function, but not a straight-line function, properly reflects the

workings of a critical period. Here we illustrate how seemingly minor technical (statistical) caveats, which often risk being brushed under the rug as nit-picking, can fundamentally affect the conclusions of a study. In the final section, we expand our scope by considering the relevance of L2 acquisition theory to the social context of L2 learners and users. Specifically, we examine three issues relating to the notion that late L2 acquisition is inferior to L1 acquisition: early instruction of foreign languages in schools, the emphasis on deficits versus capacities, and societal prejudices against non-nativelikeness.

The Critical Period Hypothesis: Predictions, Evidence and Criticisms

While often referred to as *the* critical period hypothesis for L2 acquisition, the CPH/L2A is actually a conglomerate of partly overlapping, partly contradictory hypotheses (Long, 2005; Singleton, 2005). This variation is evident in the different posited time windows of the critical periods (see Muñoz & Singleton, 2011, pp. 8-11) as well in the language domains presumed to be affected by them (see Birdsong, 2006; Long, 2007; Scovel, 2006). In broad terms, however, empirical evaluations of CP accounts of L2 acquisition revolve around two key predictions. The first prediction is that nativelikeness is possible in young L2 acquirers but impossible in older (post-CP) acquirers. The second is that the function linking age of onset of acquisition (AoA) to the learners' ultimate attainment in the L2 is different before and after the closure of the hypothesized critical period. In what follows we present a brief overview of behavioral evidence pertaining to these predictions and highlight epistemological and technical caveats concerning this evidence.

The 'Non-Nativelikeness' Prediction

Several CPH/L2As predict that L2 learners starting the acquisition process after the closure of the critical period cannot achieve nativelike competence in the target language (e.g. DeKeyser & Larson-Hall, 2005; Hyltenstam & Abrahamsson, 2003; Long, 1990). On the face of

it, this 'non-nativelikeness' prediction is falsifiable: finding a single nativelike post-CP L2 learner would prompt a re-evaluation of the critical period concept (see Long, 1990). Studies examining this prediction typically involve L2 learners and L1 controls participating in a linguistic task. The L1 speakers' scores are then used to establish an interval of nativelike performance on the task in question. Depending on the study, this interval is constructed as within one, two or three standard deviations around the mean of the native speakers' scores, or as the range of the native speakers' scores. L2 learners with scores within this interval are considered to perform to nativelike standards.

However, studies claiming to have observed nativelike performance by post-CP learners (e.g. Bongaerts, 1999; White & Genesee, 1996) have met with strong critiques. One recurring critique is that nativelike performance on a small battery of L2 tasks is not enough to establish across-the-board nativelikeness in real communicative situations and under close examination: had these learners been tested more thoroughly, their non-nativeness might have been discerned (see Long, 2005, 2007; DeKeyser, 2006; DeKeyser & Larson-Hall, 2005, for this and similar issues, but see Rothman, 2008, for counter-arguments). For reasons of space, we will not address these studies in detail. Instead, we ask whether the concept of nativelikeness can fruitfully be used for addressing CPH/L2As.

Hyltenstam and Abrahamsson (2003) argued that researchers should conduct highly detailed linguistic analyses on learners' performance. From their review they conclude that not only is there no convincing evidence that any claim of post-CP nativelikeness stands up to scrutiny, but that it is doubtful that *any* L2 learner, pre-CP learners included, can fully acquire all the subtleties of true nativeness. In an empirical follow-up to this idea, Abrahamsson and Hyltenstam (2008, 2009) found that L2 learners who pass for natives in an accent judgment task with naïve native judges fare substantially worse when subjected to a battery of L2 tasks, with

only a handful of early learners and no late learners passing the nativeness mark on all tasks.

While Hyltenstam and Abrahamsson's emphasis on analytical rigor should be applauded, no constraints on its scope of application have been proposed. Thus, when an L2 speaker meets the nativelikeness criterion on one task or measure, s/he may be subjected to unlimited additional tasks and more microscopic measures until a non-nativelike feature is detected (Birdsong, 2005a; Davies, 2004). Operationalizing and implementing the native standard, whether 'scrutinized' or merely 'perceived', is therefore difficult and bound to be controversial. The issues involved range from which tasks to consider (e.g. Are tasks on which L1 speakers show substantial variability at all suited?), whom to use as controls, to how to construct the interval of nativelike performance (i.e., Is the criterion too stringent or too lenient?). For discussion of some of these issues, see Abrahamsson and Hyltenstam (2009), Andringa (2014), Birdsong and Gertken (2013), and Long (2005). Additionally, some researchers maintain that nativelike performance is observed only among extraordinarily gifted language learners, and that the performance of such individuals should not be taken as falsifying evidence for the CPH/L2A (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000).

An additional consideration relates to the nature of bilingualism. Among active users of more than one language, no matter what the age of acquisition of the L2, the L2 influences a variety of features of the L1 just as the L1 influences features of the L2 (e.g. Cook, 1999; Grosjean, 1989; see also Gathercole, this volume). (Note, though, that proponents of this position have never claimed that bilingualism effects alone can explain all departures from monolingual-likeness.) Since departures from monolingual-likenesss (in both the L1 and the L2) are inevitable in early and late bilingualism alike, it is not reasonable to stipulate across-the-board nativelikeness as a criterion for falsifying the CPH/L2A (e.g. Cook, 1999; Ortega, 2009).

Responding to this notion, a reviewer comments: "Why are such differences inevitable?

The finding from the many such comparisons in the SLA literature is precisely that some young starters are *not* distinguishable from native speakers, whereas older starters always are" (original emphasis). This comment highlights the difficulties involved in defining what constitutes nativelikeness: if reports of nativelikeness in late learners can be critiqued for not stemming from sufficiently excruciating scrutiny across all linguistic domains, the evidence that "some young starters are *not* distinguishable from native speakers" would have to come from similarly meticulous and exhaustive scrutiny of these young learners, too. Thus, the absence of evidence for non-nativelikeness in early learners does not amount to evidence for the absence of such nativelikeness.

Under the nature-of-bilingualism account, there should be no bilinguals, of any L2 AoA (including bilinguals from birth), who, under scrutiny and across all measurable domains, are identical to monolinguals. This account is supported by the fundamental observation that, among bilinguals of any L2 AoA — and even those with low L2 proficiency — the lexicons of both languages are activated in processing either language (e.g. Dijkstra & Van Heuven, 2002; Schwartz & Kroll, 2006). By definition, bilinguals do not process their lexicons like monolinguals, precisely because monolinguals do not have two languages. Additional evidence for the nature-of-bilingualism account comes from studies of Voice Onset Time (VOT; e.g. Fowler, Sramko, Ostry, Rowland, & Hallé, 2008; Macleod & Stoel-Gammon, 2005; Sundara, Polka, & Baum, 2006). In brief, these studies reveal that, among bilinguals from birth and early bilinguals, some VOT values in both the L1 and L2 of adult bilinguals differ from those of adult monolinguals.

To summarize, the CPH/L2A and the nature-of-bilingualism argument make the same prediction of inevitable non-monolingual nativelikeness among late L2 learners. However, on the nature-of-bilingualism view, non-monolingual nativelikeness should be observed in the L2 of all

active users of two languages, no matter what the AoA of the L2. For the CPH/L2A to prevail over the bilingualism account, absolute *absence* of bilingualism effects in some early L2 learners would have to be demonstrated. One might expect, however, that such endeavors would be unsatisfactory for the same reason that reports of nativelikeness in late L2 learners have been critiqued in the past: absence of evidence is not evidence of absence. Given the conceptual and empirical difficulties associated with the non-nativelikeness prediction, the critical period concept can perhaps be more fruitfully addressed via a second common prediction, which we turn to next.

The 'Discontinuity' Prediction

The second strong prediction made by a broad class of CPH/L2As is that the relationship between age of onset of acquisition (AoA) and ultimate L2 attainment (UA) will not be linear. Instead, these CPH/L2As predict that the relationship between AoA and UA will be characterized by a bend in the function following the closure of the critical period (often referred to as a 'discontinuity'). The 'discontinuity' prediction is spelled out by DeKeyser and Larson-Hall (2005; see also Long, 2013, p. 7):

> [T]he critical period concept implies a break in the AoA-proficiency function, i.e., an age … 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)

The logic behind the 'discontinuity' prediction is that, if heightened sensitivity to L2 input is confined to a temporally constrained critical period, then there should be no age-related differences due to maturational effects after the closure of the critical period. Note that CP accounts do not necessarily predict the absence of age-related differences after the closure of the CP for reasons other than maturation: Age-related decreases in memory performance and test taking skills as well as changing socialization patterns may well result in a negative AoA-UA

relationship in adult learners (DeKeyser, 2006, 2013). Such a pattern would not, however, suffice to falsify critical period accounts. In practice, then, most CPH/L2As still allow for some (less pronounced) post-CP declines in UA (e.g., Long, 2013, p. 4).

Both DeKeyser's (2012) and Long's (2013) prediction of a non-linear relationship between AoA and UA correspond to the left and middle panels of Figure 1 (see Birdsong, 2005b). In the left panel, UA declines steeply up to the closure of the critical period, from which point onwards the AoA-UA function levels off ('stretched L'). In the middle pattern, maximum levels of L2 attainment are still possible during the first few years in the critical period but UA again drops steeply towards the closure of the CP ('stretched Z'). These patterns have in common a function that levels off after the closure of the critical period. A third possible pattern — inconsistent with both DeKeyser's and Long's predictions but sometimes encountered in CPH/L2A-related discussions (e.g., Hyltenstam & Abrahamsson, 2003; Yeni-Komshian, Flege, & Liu, 2000) — is the 'stretched 7' pattern plotted in the right-hand panel. According to this conceptualization, UA remains at its peak level during the entire critical period and starts to decrease linearly thereafter. In the remainder of this section, we will discuss findings pertaining to predictions that the AoA-UA function levels off after the closure of the hypothesized critical period – i.e., the common feature of the 'stretched L' and 'stretched Z' discontinuities.

[FIGURE 1 ABOUT HERE]

Census studies. One class of evidence addressing the 'discontinuity' prediction comes from large-scale census studies. Stevens (1999), for instance, extracted self-reported English proficiency data of 8,046 foreign-born adults residing in the U.S. as well as the approximate ages at which they arrived in the U.S. from the 1990 U.S. Census. She found that the probability with which these immigrants considered that they spoke English 'very well' decreased non-linearly as a function of their AoA, but noted that the data suggested no "abruptly defined 'critical' or

sensitive period in L2 learning" (Stevens, 1999, p. 569). However, both non-linear trends and the absence of abrupt discontinuities follow naturally from the ordinal logistic regression model that was used to analyze these data (i.e., under this model Stevens could not *not* have found non-linear continuous trends; see Vanhove, 2013). In other words, Stevens's results can be taken neither in support of a critical period account nor as evidence against it.

In a similar study, Bialystok and Hakuta (1999) made use of a regression model that is able to capture abrupt changes in the function's slope ('locally weighted scatterplot smoothing') to model the relationship between self-reported English proficiency data from nearly 64,000 Chinese- and Spanish-speaking immigrants to the U.S. and their approximate ages of arrival. They found that this relationship was captured by a linearly declining function. Comparable results were obtained in even larger studies (Hakuta, Bialystok & Wiley, 2003; Chiswick & Miller, 2008).

In summary, these large-scale census studies did not confirm the 'discontinuity' prediction. Instead, they point to a continual linear decrease of L2 proficiency levels with increasing AoA. These results have, however, been criticized on the grounds that the underlying data are not suited for testing the CPH/L2As (see DeKeyser, 2006; DeKeyser & Larson-Hall, 2005; Long, 2013; Stevens, 2004). Specifically, UA levels were extracted using coarse self-evaluations on a four- or five-level scale, and the AoA variable consisted of rough approximations. According to the critics, both of these procedures could have obscured a critical period effect in the data (but see Wiley, Bialystok & Hakuta, 2005). Given the debated validity of these large-scale census studies, we turn our attention to studies designed specifically to address the 'discontinuity' prediction next.

Task-based studies. Comprehensive reviews of task-based studies are provided by Birdsong (2005b), DeKeyser (2012) and DeKeyser and Larson-Hall (2005). We will not rehash

these surveys but instead discuss four studies in some detail in order to illustrate common designs, touch on recurring points of criticism and raise some less obvious but nonetheless important technical caveats about these studies.

Johnson and Newport (1989). Johnson and Newport (1989) recruited 46 Chinese- or Korean-speaking immigrants to the U.S. whose ages of arrival ranged from 3 to 39 years. 23 of these participants had arrived to the U.S. before age 15 ('early arrivals') and 23 had arrived after age 17 ('late arrivals'). These participants had been living in the U.S. for 3 to 26 years at the time of testing.[1] Participants' proficiency in English was tested with an aurally presented 276-item grammaticality judgment task (GJT).

Johnson and Newport report a global correlation between the participants' ages of arrival and their GJT scores of $r = -.77$. Additionally, they also computed separate correlations for the early and for the late learners. For the early learners, GJT performance was strongly correlated with age ($r = -.87$), whereas for the late learners, the correlation was appreciably weaker ($r = -.16$). Johnson and Newport conclude from this finding that the AoA–UA function flattens off from puberty onwards, suggesting the workings of maturational constraints in L2 acquisition.[2] Since Johnson and Newport, several more studies have been conducted in similar vein, including those we turn to next.

Birdsong and Molis (2001). Birdsong and Molis (2001) presented the same grammaticality judgment task as Johnson and Newport's (1989) to a sample of 61 Spanish-speaking immigrants to the U.S. Their ages of arrival ranged from 3 to 44 years. 29 were 'early arrivals' ($\text{AoA} \leq 16$); 32 were 'late arrivals' ($\text{AoA} \geq 17$).

Like Johnson and Newport, Birdsong and Molis found a global correlation of $r = -.77$ between the participants' ages of arrival and their GJT scores. Contrary to Johnson and Newport, however, Birdsong and Molis found that the subgroup correlation was stronger in the late ($r =$

-.69) than in the young arrivals (r = -.24). As we will illustrate below, however, differences between subgroup correlation coefficients do not necessarily indicate discontinuities in the AoA–UA function. Such discontinuities can more appropriately be identified using regression techniques. Birdsong and Molis therefore carried out regression analyses on Johnson and Newport's as well as on their own data, the main results of which are presented in Figure 2. Note that while the Johnson and Newport data seem to be captured by a CPH/L2A-consistent 'stretched L' shape, the Birdsong and Molis data form the inverse pattern ('stretched 7').

[FIGURE 2 ABOUT HERE]

In sum, Birdsong and Molis failed to reproduce the results that Johnson and Newport (1989) had taken in support of a critical period account of L2 acquisition and, in fact, ended up with results that form the mirror image of Johnson and Newport's findings.

Long (2007) critiqued Birdsong and Molis's study for having used too easy a test, which he claims gave rise to ceiling effects (i.e. the early plateau in Figure 2). With Birdsong and Molis (2001, p. 245), however, we note the need for criteria other than a post-hoc interpretation of the results for accepting an instrument in the original study and rejecting it in the replication. DeKeyser and Larson-Hall (2005) additionally suggest that Birdsong and Molis's results were tainted by outliers, notably the participant with AoA 35 and a test score of 166 (Note 2, p. 104). Reanalyzing their results without the three late arrivals with the lowest scores, however, Birdsong and Molis (2001, Note 2, p. 240) still found a negative correlation. Furthermore, note again the need for criteria other than post-hoc interpretations for rejecting data points: Johnson and Newport's study features a participant with AoA 18 and a test score of 163 who arguably constitutes a larger outlier (see Figure 2). Without external criteria, it seems difficult to defend critiquing Birdsong and Molis's finding without similarly critiquing Johnson and Newport's results.

DeKeyser, Alfi-Shabtay, and Ravid (2010). DeKeyser et al. (2010) set out to test the CPH/L2A's 'discontinuity' prediction in two parallel studies. The first sample consisted of 76 Russian-speaking immigrants to the U.S. and Canada; the second of 62 Russian-speaking immigrants to Israel. The participants' ages of arrival ranged from 4 to 71 years, and the lengths of residence ranged from 8 to 28 years. UA levels were assessed using a 200-item grammaticality judgment task (an adaptation of the Johnson & Newport's, 1989, instrument).

Global correlations computed for the relationship between AoA and UA closely resembled those reported by Johnson and Newport (1989) and by Birdsong and Molis (2001) in both studies (r = -.80 and -.79 for the North America and Israel studies, respectively). However, the correlation was stronger for early arrivals (AoA < 18; r = -.69 and -.48) than for middle-aged arrivals (18 ≤ AoA ≤ 40; r = -.45 and -.37). DeKeyser et al. interpret these differences as indicative of a discontinuity in the AoA–UA relationship and hence as cross-linguistic confirmation of the CPH/L2A's 'discontinuity' prediction. DeKeyser et al. also report correlations between AoA and UA in late arrivals (AoA > 40; r = -.27 and -.53), which they attribute to the late arrivals' old age at the time of testing.

DeKeyser et al.'s conclusions have been called into question on statistical grounds by Vanhove (2013). In CPH/L2A-related discussions, differences between subgroup correlation coefficients are often taken as evidence of changes in the slope of the AoA–UA function (e.g., DeKeyser, 2012, p. 448). Such a conclusion conflates correlation and regression. Briefly, correlation coefficients express how closely the data map onto a line, whereas regression analysis, seeks to estimate the functional characteristics of this line. It is possible for two relationships to be characterized by the same correlation coefficient but to have different functional forms, and vice versa (see, e.g., Vanhove, 2013, Fig. 2). As CPH/L2As are currently formulated, however, the 'discontinuity' prediction is concerned with the functional

characteristics of the AoA-UA relationship, more specifically with changes in its slope.

In the Johnson and Newport ‘s (1989) and Birdsong and Molis’s (2001) studies discussed earlier, correlational and regressional analyses sketch broadly the same picture (as illustrated by Figure 2). However, to illustrate how conflating correlation coefficients with regression coefficients can lead to diametrically opposed conclusions about the CPH/L2A, Vanhove (2013) carried out a reanalysis of DeKeyser et al.’s (2010) data. Specifically, the data were modeled using piecewise regression models. These are comparable to ordinary regression models with the difference that the slope of the function relating the predictor to the outcome is allowed to change at a specified inflection point. Figure 3 graphically summarizes the piecewise regression models with inflection points at AoA = 18 for DeKeyser et al.’s North America and Israel studies and also shows simple regression fits as a reference. As can be gleaned from these plots, piecewise regression models did not substantially improve the fit in either the North America or Israel study (gain in $R^2 \leq .015$).

[FIGURE 3 ABOUT HERE]

It may be argued that the cut-off AoA of 18 years was decided on for partly arbitrary reasons (see DeKeyser et al., 2010, p. 418) and that a cut-off at the true (but undefined) closure of the CP would have yielded different results. Vanhove therefore computed a series of piecewise regression models for which the cut-off AoAs varied from 5 to 19 years. For the North America study, the best-fitting model had a cut-off at 16 years, which marginally improved the fit over a linear model (gain in $R^2 = .019$, $p = .05$);[3] for the Israel study, the best-fitting model had an inflection point at 6 years, which did not appreciably improve the fit (gain in $R^2 = .002$). In conclusion, then, the DeKeyser et al. data do not provide cross-linguistic evidence in support of a critical period account but instead point to a continual linear decrease in ultimate L2 attainment as a function of the age of onset of L2 acquisition.

Despite some earlier precedents (e.g., Birdsong & Molis, 2001; Hakuta et al., 2003), reservations about the use of piecewise regression to address CPH/L2A predictions have been expressed in the SLA literature (Stevens, 2004; Long, 2013, p. 10). Unfortunately, spatial restrictions prevent us from considering these here.

Granena and Long (2013). A recent study by Granena and Long (2013) addresses the 'discontinuity' prediction in three linguistic domains: morphosyntax, pronunciation, and lexis. The sample consisted of 65 speakers of Chinese who had moved to Spain. In contrast to the studies reviewed thus far, these speakers were screened for their advanced L2 proficiency, i.e., the participants did not represent the full spectrum of L2 proficiencies. The participants' ages of acquisition ranged from 3 to 29 years and their lengths of residence from 8 to 31 years. The participants completed a battery of linguistic tasks testing their pronunciation, lexical and collocational knowledge, and morphosyntactic attainment.

Granena and Long reported global correlations between AoA and UA in line with those in the studies discussed above: r = -.81 for pronunciation, -.79 for lexical and collocational knowledge and -.73 for morphosyntax. Additionally, they fitted piecewise regression models with two inflection points: one at AoA 6 and one at AoA 15 years.[4] The increase in variance explained compared to simple linear models varied between 4 and 5 percentage points. In a further analysis, Granena and Long pooled the pronunciation, lexis and morphosyntax data into one dataset, which was then analyzed using a repeated-measures (RM) ANOVA. Of specific interest was the question whether linguistic domain interacted with AoA-defined subgroup status (i.e., AoA < 6, AoA 7-15, AoA > 15). The authors found a significant interaction between linguistic domain and AoA-subgroup such that the earliest arrivals had a higher mean score on the pronunciation measures than on the lexical and morphosyntax measures and the later arrivals had a higher mean score on the morphosyntax measures.

Granena and Long interpret these findings as evidence of critical periods with different timings according to the linguistic domain affected: a closure for pronunciation in early childhood, followed by lexical and collocational knowledge and lastly a closure of a critical period for morphosyntax around age 15. However, the repeated-measures (RM) ANOVA that was carried out in order to bolster this claim makes the fundamental assumption that the dependent variable consists of repeated measures *of the same variable*. The three types of outcomes observed in Granena and Long's study are not the same variable measured under three different conditions (i.e. a pronunciation test condition, a lexical test condition, and a morphosyntactic test condition); rather, they represent three different, if correlated, qualities. Neither the fact that these outcomes can all be scaled to percentages nor that they could be construed as tapping into the same latent variable (i.e. L2 proficiency) changes this: the dependent variable has to be the same across all conditions. An inspection of the scatterplots presented in Figure 5 in Granena and Long (2013, p. 328) reveals that the spread in the morphosyntax data was visibly lower than that in the pronunciation and lexis-collocation data. As a result, the difference in the mean scores between early arrivals and late arrivals was smaller for the morphosyntax data, resulting in – or at least contributing to – a significant interaction term. However, this interaction can be explained without having to invoke multiple critical periods by simply recognizing that differences in the variability of scores on different tests are also caused by differences in the design and rating of the tests. In brief, we maintain that the results that Granena and Long (2013) marshal in support of a multiple CPs account may be artifacts of an unwarranted analysis.

Summary: The AoA-UA function controversy. The studies discussed above sketch a picture of how the 'discontinuity' prediction has been assessed. As is obvious from this selective overview, the issue is highly controversial with researchers on either side of the debate arguing

forcefully that evidence garnered on the other side of the fence rests on faulty interpretations and/or design (e.g., Bialystok, 2002; Long, 2005, 2007; DeKeyser, 2006; DeKeyser & Larson-Hall, 2005; Stevens, 2004; see also the points raised in this chapter). But taking a bird's-eye view, what general conclusions can be drawn from the current state of the art?

Firstly, a recurrent finding is that ultimate L2 attainment levels in the domains of pronunciation, morphosyntax and lexical and collocational knowledge are negatively associated with the learners' ages of onset of acquisition over a wide AoA range (see also, among others, DeKeyser, 2000; Flege, Yeni-Komshian & Liu, 1999). This finding is widely accepted and is not a source of controversy.

Secondly, researchers subscribing to the CPH/L2A have argued that the AoA–UA function shows a distinct flattening that marks the closure of the (or 'a') critical period, typically in the mid-teens. While these conclusions typically rest on generous interpretations of group mean and correlation coefficient comparisons (Vanhove, 2013; see also above), a handful of studies have indeed produced evidence to such discontinuities, notably Johnson and Newport (1989) according to Birdsong and Molis's (2001) reanalysis and, more recently, Granena and Long (2013; but see the caveat in Note 4). Other results contradict a critical period account, however. On the one hand, Birdsong and Molis (2001) found a steeper rather than a flattened slope in later arrivals. On the other hand, comparable linear decreases were found before and after the closure of the critical period under consideration (e.g., Flege et al., 1999). Such a continued linear decrease was observed even in Vanhove's (2013) reanalysis of data from two studies that had been taken as evidence confirming the 'discontinuity' prediction. We have previously proposed that these conflicting findings could be accounted for by assuming that UA levels decrease linearly across a wide AoA range (Birdsong, 2009; Vanhove, 2013), i.e., without positing a temporally delineated period of heightened sensitivity to linguistic input or critical

period. Depending on the specifics of the study (including floor and ceiling effects) as well as random variability, 'L' or '7' shaped deviations from linearity can arise. A general linear trend could also explain why incidences of L2 learners performing within a native control sample's range decrease with advancing AoA (see Birdsong, 2005b, Fig. 6.8). Interestingly, the suggestion that conflicting findings in CPH/L2A-related research can be reconciled by assuming a continual linear decline was also put forward by Hyltenstam and Abrahamsson (2003), who proposed a maturation-based account of L2 attainment without temporally constrained 'windows of opportunity'.

Lastly, we hope that our discussion has illustrated that seemingly minor technical details (e.g., the difference between correlation and regression and the specification of statistical models) can dramatically affect the inferences drawn from the data.

Second Language Outcomes in the Social Context

In this section we consider the issue of age of L2 acquisition in its societal context. In the form of brief précis, we highlight three popular manifestations of the notion that (late) L2 acquisition is inferior relative to L1 acquisition: the idea that "earlier is better" in foreign language education, the emphasis on linguistic deficits among bilinguals as compared to monolinguals, and socially-ingrained prejudices against non-nativelikeness.

The CPH/L2A and the foreign language classroom. Questions about the optimal age for beginning L2 classroom learning are at the center of ongoing debates among educators and parents about the timing, amount, and usefulness of classroom teaching of foreign languages (e.g. Gürsoy, 2011). CPH/L2A research has figured prominently in these debates. However, most studies relating to the CPH/L2A involve ultimate attainment data gathered from bilinguals who have acquired, and who use, the L2 in naturalistic contexts. The application of findings under the CPH/L2A is thus of limited utility in the classroom setting, which concerns neither

ultimate attainment nor immersion.

Still, the notion that "earlier is better" that is associated with the CPH/L2A is embraced by the general public and by stakeholders in elementary education. This notion has recently been called into question by Huang (2015), who reviews 40 empirical studies of the age of beginning foreign language instruction. Huang finds that, while there are non-linguistic benefits of early foreign language instruction (e.g. in terms of cognitive development, attitudes toward languages other than the L1, and general academic achievement), there is no decisive evidence that "earlier is better" with respect to indices of either short-term or long-term L2 attainment in the classroom setting.

Deficits and capacities. The literature on age and L2 acquisition, with its emphasis on non-nativelikeness,[5] often obscures the high levels of L2 ultimate attainment and communicative competence enjoyed by millions of users of multiple languages of varying AoA. Commenting on a 2014 special issue of *Applied Linguistics* "Complexities and Interactions of Age in Second Language Learning: Broadening the Research Agenda" (*Applied Linguistics, 35, 4*), Yates and Kozar (2015) suggest that researchers' preoccupation with L2 deficits as a function of AoA "is socially out of kilter with trends toward the re-evaluation of age-related expectations" which are redirecting modern societal attention to human cognitive and physical capacities over the lifespan.

Similarly, while comparisons of L2 users with a monolingual standard have yielded important insights into L2 development over varying AoA, this methodology is insensitive to changes in the L1 among bilinguals over the lifespan, which cannot be ascribed to deficient language learning mechanisms. Given that more than half the world's population is bilingual (Grosjean, 2010), it is important that societies understand that monolingual-likeness in both languages of speakers is an unrealistic expectation.

***Societal evaluation of L2 outcomes*.** The non-nativelike pronunciation that is typical of immigrants is known to trigger hostile reactions among native speakers. For example, human rights' cases have examined employment-related discrimination and harassment based on speech accent (Munro, 2003, p. 38). In contrast, positive attitudes are exhibited toward speech that is deemed standard by a majority of members of a speech community (Fuertes, Gottdeiner, Martin, Gilbert, & Giles, 2011).

Gluszek and Dovidio (2010) look at social impressions of both non-native accents and regional native accents. They find that non-native accents, more so than regional accents, cue negative assessments of speakers' overall linguistic competence. The researchers argue that stigmatization of non-native accents extends beyond the immediate speech context to negatively impact future conversational interactions. On a more positive note, Weyant (2007) finds that negative stereotyping of non-native speakers can be reduced in experimental settings by perspective-taking activities (Galinsky & Moskowitz, 2000).

Conclusions

Scientific theories should make predictions that, in principle, permit a falsification of the theory (Popper, 1959). The CPH/L2A is associated with two key predictions: 'non-nativelikeness' and 'discontinuity'. As we have seen, however, defining what constitutes sufficient evidence for nativelikeness is difficult if not impossible on epistemological grounds: the absence of evidence for non-nativelikeness does not equal evidence for nativelikeness. The 'discontinuity' prediction can more readily be assessed, but in assessing the data available, several caveats have to be kept in mind, some of which we have discussed. We tentatively conclude that there is no convincing evidence for a CPH/L2A-compatible 'bend' in the AoA-UA function. Such evidence may, however, still come in the form of highly-powered, carefully analyzed and cross-linguistically replicated studies.

The scientific study of age and L2 attainment does not take place in a vacuum, disconnected from societal concerns. For example, we pointed out that what learners achieve in classrooms may not reflect an advantage for early foreign language learning, and thus is inconsistent with results of ultimate attainment studies in immersion contexts. In addition, we underscored the risks of comparing L2 users with monolingual natives, and suggested that it is in the interest of bilingual societies to recognize that monolingual-bilingual differences are normal, not deficient, outcomes in both the L1 and the L2. Finally, we noted that prejudices against the non-nativelikeness associated with late L2 learning are widespread, come in many forms, and can be corrosive cultural forces. In these various respects, the complete picture of age and L2 acquisition both incorporates and transcends theory.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30,* 481-509. DOI: 10.1017/S027226310808073X

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*, 249-306. DOI: 10.1111/j.1467-9922.2009.00507.x

Andringa, S. (2014). The use of native speaker norms in critical period research. *Studies in Second Language Acquisition, 36,* 565-596. DOI: 10.1017/S0272263113000600

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

Bialystok, E. (2002). On the reliability of robustness. *Studies in Second Language Acquisition, 24*, 481-488. DOI: 10.1017/S0272263102003054

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* (pp. 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. DOI: 10.1017/S1366728999000231

Birdsong, D. (2005a). Nativelikeness and non-nativelikeness in L2A research. *International Review of Applied Linguistics in Language Teaching, 56*, 9-49. DOI: 10.1515/iral.2005.43.4.319

Birdsong, D. (2005b). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109-127). Oxford: Oxford University Press.

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*, 9-49. DOI: 10.1111/j.1467-9922.2006.00353.x

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* (2nd ed.) (pp. 401-424). Bingley: Emerald.

Birdsong, D., & Gertken, L. (2013). In faint praise of folly: A critical review of native/non-native speaker comparisons, with examples from native and bilingual processing of French complex syntax. *Language, Interaction and Acquisition, 4,* 107-133. DOI: 10.1075/lia.4.2

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language, 44*, 235-249. DOI:10.1006/jmla.2000.2750

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. M.

Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41-68). Cambridge: Cambridge University Press.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the Critical Period Hypothesis* (pp. 133-159). Mahwah, NJ: Lawrence Erlbaum.

Chiswick, B. R., & Miller, P. W. (2008). A test of the critical period hypothesis for language learning. *Journal of Multilingual and Multicultural Development, 29*, 16-29. DOI: 10.2167/jmmd555.0

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33,* 185-209. DOI: 10.2307/3587717

Davies, A. (2004). The native speaker in applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431-450). Malden, MA: Blackwell.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499-533. DOI: 10.1017/S027226310808073X

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* (pp. 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* (pp. 442-460). London: Routledge.

DeKeyser, R. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning, 63(Suppl. 1)*, 52-67. DOI: 10.1111/j.1467-9922.2012.00737.x

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. DOI: 10.1017/S0142716410000056

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* (pp. 88-108). New York: Oxford University Press.

Dijkstra, A., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5,* 175-197. DOI: 10.1017/S1366728902003012

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41*, 78-104. DOI: 10.1006/jmla.1999.2638

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., & Hallé, P. (2008). Cross language phonetic influences on the speech of French-English bilinguals. *Journal of Phonetics, 36,* 649-663. DOI: 10.1016/j.wocn.2008.04.001

Fuertes, J. N., Gottdiener, W., Martin, H., Gilbert, T. C., & Giles, H. (2009). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology, 42,* 120-133. DOI: 10.1002/ejsp.862

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78,* 708–724. DOI: 10.1037/0022-3514.78.4.708

Gluszek, A., & Dovidio, J. F. (2010). The way *they* speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review, 14,* 214-237. DOI: 10.1177/1088868309359288

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*, 311-

343. DOI: 10.1177/0267658312461497

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*, 3-15. DOI: 10.1016/0093-934X(89)90048-5

Grosjean, F. (2010). *Bilingual.* Cambridge, MA: Harvard University Press.

Gürsoy, E. (2011). The Critical Period Hypothesis revisited: The implications for current foreign language teaching to young learners. *Journal of Language Teaching and Research, 2,* 757-762. DOI: 10.4304/jltr.2.4.757-762

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. DOI: 10.1111/1467-9280.01415

Huang, B. H. (2015). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism*. Published online 24 July 2015. DOI:10.1080/14790718.2015.1066792

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539-588). Malden, MA: Blackwell.

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. DOI: 10.1016/0010-0285(89)90003-0

Lambelet, A., & Berthele, R. (2015). *Age and foreign language learning in school*. Basingstoke: Palgrave Macmillan.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition, 12,* 251-285. DOI: 10.1017/S0272263100009165

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. DOI: 10.1515/iral.2005.43.4.287

Long, M. H. (2007). *Problems in SLA*. Mahwah, NJ: Lawrence Erlbaum.

Long, M. H. (2013). Maturational constraints on child and adult SLA. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 3-41). Amsterdam: John Benjamins.

Macleod, A. N., & Stoel-Gammon, C. (2005). Are bilinguals different? What VOT tells us about simultaneous bilinguals. *Journal of Multicultural Communication Disorders, 3,* 118-127. DOI: 10.1080/14769670500066313

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching, 44*, 1-35. DOI: 10.1017/S0261444810000327

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal/Revue TESL du Canada, 20,* 38-51.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder.

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton: Princeton University Press.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson & Co.

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. DOI: 10.1111/j.1749-818X.2008.00098.x

Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language*, *55*, 197-212. DOI:10.1016/j.jml.2006.03.004

Scovel, T. (2006). Age, acquisition, and accent. In C. Abello-Contesse, R. Chacón-Beltrán, M. D.

López-Jiménez, & M. M. Torreblanca-López (Eds.), *Age in L2 acquisition and teaching* (pp. 31-48). Bern: Peter Lang.

Singleton, D. (2005). The critical period hypothesis: A coat of many colours. *International Review of Applied Linguistics in Language Teaching, 43*, 269-285. DOI: 10.1515/iral.2005.43.4.269

Stevens, G. (1999). Age of immigration and second language proficiency among foreign-born adults. *Language in Society, 28*, 555-578. DOI: 10.1017/S0047404599004030

Stevens, G. (2004). Using census data to test the critical-period hypothesis for second-language acquisition. *Psychological Science, 15*, 215-216. DOI: 10.1111/j.0956-7976.2004.01503012.x

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneous bilingual adults. *Bilingualism: Language and Cognition, 9,* 97-114. DOI: 10.1017/S1366728905002403

Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLoS ONE, 8*, e69172. DOI: 10.1371/journal.pone.0102922

Weyant, J. M. (2007). Perspective taking as a means of reducing negative stereotyping of individuals who speech English as a second language. *Journal of Applied Social Psychology, 37,* 738-748. DOI: 10.1111/j.1559-1816.2007.00181.x

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. DOI: 10.1177/026765839601200301

Yates, L., & Kozar, O. (2015). Expanding the horizons of age-related research: A response to the Special Issue 'Complexities and Interactions of Age in Second Language Learning: Broadening the Research Agenda'. *Applied Linguistics, 36*. DOI: 10.1093/applin/amv031

Yeni-Komshian, G. H., Flege, J. E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Bilingualism: Language and Cognition, 3*, 131-149. DOI: 10.1017/S1366728900000225

Notes

[1]Note that other studies featuring learners with short lengths of residence (e.g. Bialystok & Miller, 1999) have been dismissed by reviewers (e.g. DeKeyser & Larson-Hall, 2005; Long, 2013) on the grounds that their results do not speak to ultimate attainment but rather to rate of learning. DeKeyser (2013) recommends that length of residence be at least 10 years to ensure that learners are at asymptote.

[2]Johnson and Newport (1989) additionally computed a series of *t*-tests between AoA-defined subgroups in order to determine whether the AoA–UA relationship exhibits discontinuities. For a critique of the use of subgroup mean comparisons to identify discontinuities see Vanhove (2013).

[3]The reported *p*-values were not corrected for multiple testing. Looping through the data in search of the ideal cut-off age inflates the Type-I error rate. We estimate that for the North America study, the *p*-value would have to be recalibrated to about $p = 0.10$ to take this Type-I error inflation into account. For details concerning this estimation, see http://dx.doi.org/10.6084/m9.figshare.1206291.

[4]"A restricted one-slope model with AO [age of onset of acquisition] as a single predictor was compared against a full model that included interaction terms between the predictor and dummy-coded AO group variables" (Granena & Long, 2013, p. 16). While this is the method for fitting piecewise regression models recommended by Baayen (2008, p. 214; also cited in Long's, 2013, discussion of this study) and used by Vanhove (2013) in his reanalysis of DeKeyser et al. (2010), it only yields sensible regression fits when modeling one inflection point. Fitting a

piecewise regression model with multiple inflection points, though possible by other means, results in disjointed segments rather than a single connected regression fit.

[5]As mentioned earlier, many factors in addition to AoA have been examined in studies of (non-)nativelike attainment. These factors include degree of entrenchment of the L1, years of schooling, linguistic training, literacy, length of residence, L2 exposure/input/use, goals of learning, social-psychological integration and identity, and motivation to become indistinguishable from natives. As Flege, Yeni-Komshian and Liu (1999) demonstrate, different linguistic features of the L2 (e.g. morphosyntax, lexical knowledge, pronunciation) are not equally affected by these factors.

Figures

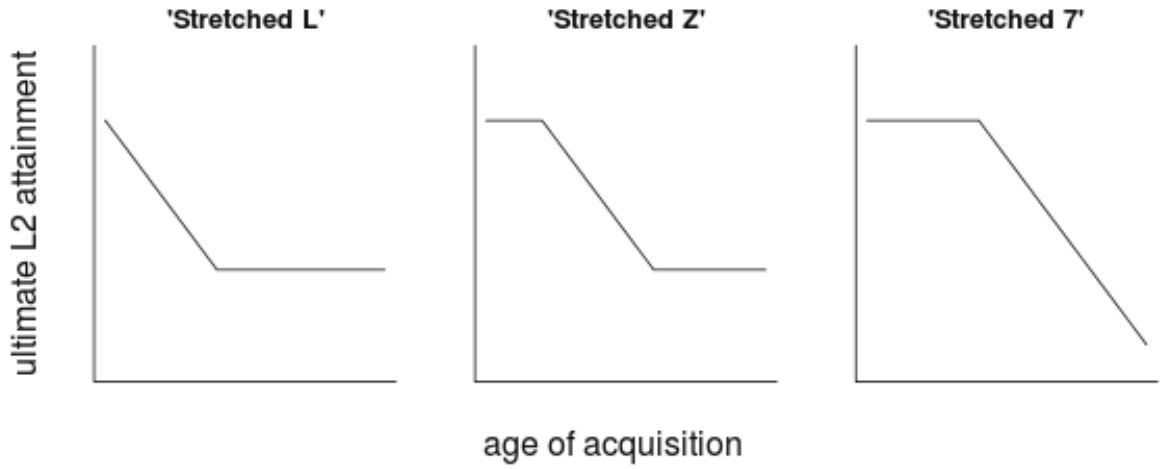

Figure 1. Three possible non-linear relationships between age of acquisition and ultimate L2 attainment proposed in the literature (see Birdsong, 2006).

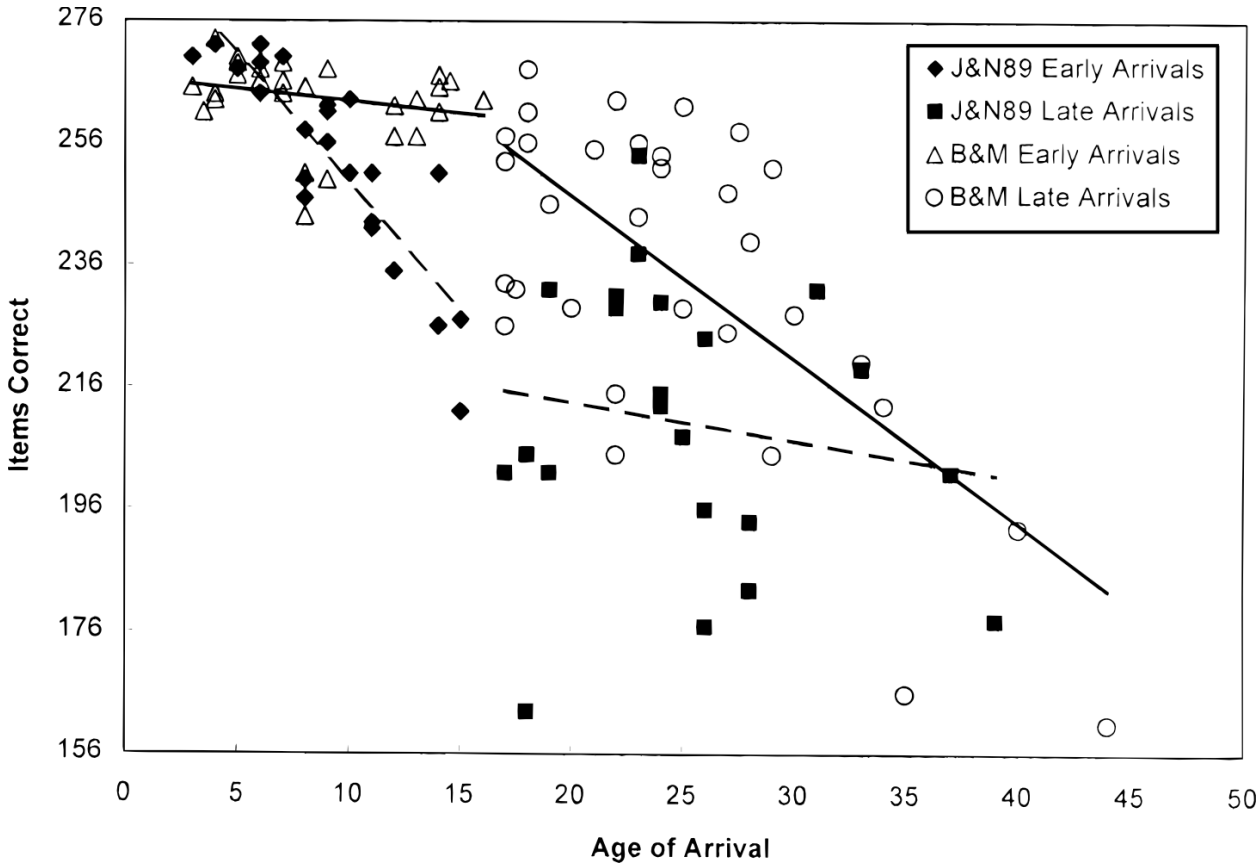

Figure 2. Scatterplot showing the relationship between age of arrival and GJT scores for both Johnson and Newport's (1989; dotted lines) and Birdsong and Molis's (2001; solid lines) studies. The regression lines were computed separately for early (AoA ≤ 16) and late (AoA ≥ 17) arrivals. (source: Birdsong & Molis, 2001, Fig. 3)

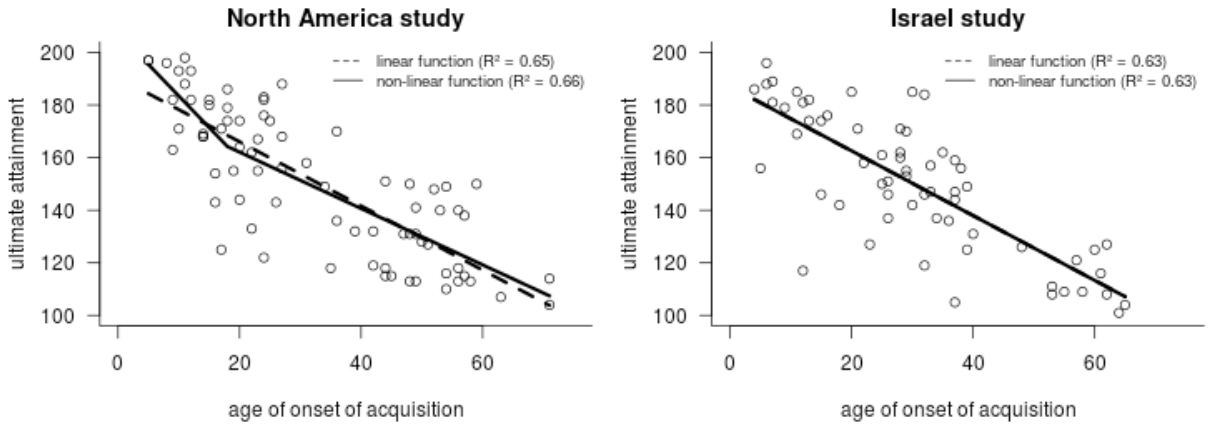

Figure 3. Graphical summary of the reanalysis of DeKeyser et al.'s (2010) North America and Israel studies by Vanhove (2013). The dotted lines show the fitted values derived from a simple linear regression model; the solid lines were derived from piecewise regression models featuring a breakpoint at age 18. (The dotted line is not absent from the right-hand panel but is hardly visible due to its near-complete overlap with the solid line.)