Age Effects in Second Language Learning: Stepping Stones Toward Better Understanding

Robert M. DeKeyser
University of Maryland

The effect of age of acquisition on ultimate attainment in second language learning has been a controversial topic for years. After providing a very brief overview of the ideas that are at the core of the controversy, I discuss the two main reasons why these issues are so controversial: conceptual misunderstandings and methodological difficulties. The main part of the article then makes suggestions for improvement in subject selection, data collection, and instrumentation, in the hope that both sides of the debate will be able to agree on them. More sophisticated research in this area is of the utmost importance given how crucial understanding age effects is for educational policy and curriculum design. Where foreign language learning rather than second language learning is concerned, directly relevant research, carried out with classroom foreign language learners, is even more sorely needed.

Keywords age effects; critical period; immigrants; bilingualism; research methodology

Introduction

Even though just about every layman knows and just about any psychologist or linguist would agree that somehow children are better at learning languages than adults, the exact extent, cause, and nature of this phenomenon have been controversial for decades. In this article I want to concentrate on what have been the biggest points of contention in the last ten years or so—for a state-of-the-art overview of the findings and a discussion of possible interpretations, see, for example, Birdsong (2006), Herschensohn (2007), Hyltenstam and Abrahamsson (2003), DeKeyser (2012a), DeKeyser and Larson-Hall (2005), and Muñoz (2008b).

Correspondence concerning this article should be addressed to Robert DeKeyser, School of Languages, Literatures and Cultures, University of Maryland, 3125 Jimenez Hall, College Park, MD 20742. E-mail: rdk@umd.edu
The following three questions have been the most central in these debates. First, what is the exact shape of the age of acquisition (AoA)—ultimate attainment (UA) function? Both proponents (e.g., DeKeyser, 2012a; Hyltenstam & Abrahamsson, 2003) and critics (e.g., Birdsong, 2005; Hakuta, Bialystok, & Wiley, 2003; Munro & Mann, 2005) of the critical period hypothesis (CPH) agree that there should be a breaking point in that function; one would therefore expect a flattening of the AoA-UA function somewhere in childhood or adolescence—and perhaps an increased decline in early childhood after a few years of (apparent) stability (cf. Hyltenstam & Abrahamsson, 2003).

Second, assuming there is such a break in the AoA-UA function, to what extent does this reflect a maturational phenomenon as opposed to a confound with individual variables such as motivation, attitude, and identity; and contextual variables such as quantity and quality of input, and extent of schooling in the second language (L2)? Even if there is a discontinuity in the AoA-UA function, there is no ‘critical period’1 if the function is due to such confounds. Again, both proponents and critics of the CPH agree on this point (e.g., DeKeyser & Larson-Hall, 2005; Flege, Yeni-Komshian, & Liu, 1999; Hakuta et al., 2003).

Finally, assuming a maturational phenomenon is at issue, what is its nature, that is to say, what kinds of differences develop in the crucial/critical learning processes as a function of AoA? These could be hard-wired neuro-anatomical changes or more subtle changes in cognitive functioning due to previous experience (although critics of the CPH have often equated it with a specific neurological interpretation, in particular the link that Lenneberg [1967] made with the development of hemisphere dominance).

Why It Is Important to Resolve This Controversy

It is very important to understand the true nature of these age effects, for a number of reasons. For the psycholinguist, the age questions unavoidably lead to reflection on the basic nature of (possible) learning processes. Various researchers have connected age differences with a wide variety of issues, from the nature of inferring structure from input (Goldowsky & Newport, 1993; Monner, Vatz, Morini, Hwang, & DeKeyser, 2012; Newport, 1990; Saffran, Johnson, Aslin, & Newport, 1999; Ullman, 2001) to the mechanisms of historical change (Hudson Kam & Newport, 2005, 2009; Kerswill, 1996). For the cognitive psychologist, these questions lead to broader ones about the nature of learning and development (Janacsek, Fiser, & Nemeth, 2012; Saffran, Pollak, Seibel, & Shkolnik, 2006; Thomas et al., 2004).
For the educational psychologist and the second language teacher, these questions raise important concerns about all aspects of curriculum development and its adaptation to different ages. In the case of immigrant children who are not native speakers of the dominant language, it is important that researchers understand better what L2 structures are problematic under what circumstances, and what the educational system can do about this. It is often assumed that as long as immigrant children arrive before the age of approximately 15, they will acquire the L2 perfectly from natural exposure, but even immigrants who arrived at a younger age often have gaps in their linguistic competence, particularly with the more abstract and formal aspects of language that are essential for academic and professional success (Abrahamsson & Hyltenstam, 2009; Bigelow & Tarone, 2004; Christie, 2012; Schleppegrell, 2004; Tarone, Bigelow, & Hansen, 2009). In the case of native speakers of the dominant language, it is important to understand what the age effects that have been documented so often with immigrants imply for the teaching of foreign languages: not simply starting at a younger age, but perhaps more importantly providing different instructional processes at different ages (Cameron, 2001; Muñoz, 2007). For parents, administrators, and educational authorities, the answer to these questions determines how they will invest their children’s time and their education budgets (Curtain, 1998; García Mayo & García Lecumberri, 2003; Larson-Hall, 2008).

**Conceptual Misunderstandings**

Before I discuss the difficulties that characterize research on these issues, let me point out a few common misunderstandings about age effects in second language acquisition (SLA). First and foremost, it is often said that children learn a language faster than adults; there is no evidence for that. Immigrant children’s learning is slower than adults’ (Krashen, Long, & Scarcella, 1979), but they keep going until they have reached the level of a native speaker their age, while adults all show some form of arrested development (albeit at a very high level for some; cf. Abrahamsson & Hyltenstam, 2009; Abrahamsson, 2012).

While the speed/ultimate attainment distinction was made decades ago, the implicit/explicit distinction is only beginning to be explored in the literature on age effects, and ignoring it leads to misunderstandings with expensive consequences. The difference that everybody can observe within one and the same immigrant family, where the children soon overtake their parents, reflects implicit acquisition processes only; adolescents and adults do not have any
more problems than children with the kind of learning that is typical of most foreign language learning, on the contrary. This can be seen from the role of age in classroom foreign language learning, where younger children have been shown to be slower than older children and adolescents (García-Mayo & García Lecumberri, 2003; Muñoz, 2006), and from the important role that aptitude for explicit learning plays in those adults who are relatively successful (DeKeyser, 2000; DeKeyser, Alfi-Shabtay, & Ravid, 2010; Paradis, 2009).

Still less obvious to most people, psycholinguists included, is that acknowledging there are age effects or even referring to them with terms such as “critical period” does not necessarily imply broad neurological changes like the ones that have been suggested in the literature over the decades, such as changes in hemisphere dominance (Lenneberg, 1967), myelin(iz)ation (Aslin & Schlaggar, 2006; Long, 1990, 2007; Pulvermüller & Schumann, 1994), or the development of the amygdala (Pulvermüller & Schumann, 1994). The neurological correlates could be much more subtle, and could be completely dependent on accumulated learning experience rather than independently dictated by biological development.

Nor does the term critical period, from my point of view, necessarily refer to representational issues as opposed to processing. The age effects observed may or may not characterize representation (Hopp, 2009, 2010; Juffs & Harrington, 1995; McDonald, 2006); the “mere processing” interpretation, however, does not imply there is no critical period.

Last but not least, whatever studies find out about age effects, they may not have simple educational implications; the interpretation along the lines of “just teach earlier” that seems to have inspired starting foreign language education at very early ages in many countries is almost certainly wrong, because the learning contexts for the immigrant are totally different from those of the foreign language learner (DeKeyser & Larson-Hall, 2005; Grotjahn, 2003; Hoff-Ginsberg, 1998; Hyltenstam & Abrahamsson, 2001; Patkowski, 1994; Muñoz, 2006, 2008a).

Methodological Difficulties

With this background in mind, I will now proceed to a discussion of methodological issues in research on age effects in SLA. Some of the methodological difficulties involved have been discussed to some extent in the literature; others barely ever get mentioned, even though they could be used as obvious arguments in the sometimes acrimonious debate about age effects.
Most important, from my point of view, is an issue that is almost always ignored: Researchers need to gather data from a sample of immigrants that is much more representative of the immigrant population than the convenience samples typically used (which are strongly biased toward the more educated and hence tend to minimize the age effects). At the same time, it is important to obtain data from immigrants who have had a fair chance to learn the language. Convenience samples are often taken from local communities that have many speakers of a specific first language (L1), which almost automatically means that they spend a lot of time using the L1, which in turn must have at least some negative influence on their L2 development; this causes a bias in the opposite direction of the previous point. Hence, there is a need for learners who are largely isolated from other speakers of the same L1.2

A more obvious point is that researchers need to recruit a larger number of participants in each age group than is usually the case so that inferential statistical analysis is possible for each group separately (recall that it is theoretically important to establish different correlations/regression lines for age ranges during as opposed to after the critical period, so each of the age ranges needs enough participants and enough spread to establish these patterns). A global correlation or regression analysis for the entire lifespan cannot answer the relevant questions (it will tend to provide a continuous function, even if separate age ranges show different rates of decline, something known as the Simpson paradox; see Blyth, 1972), and analyzing the data for each group separately, potentially with several predictors or covariates, requires a sizeable number of participants in each subgroup. Combining these two points makes it clear that recruitment is a difficult issue: Researchers need large samples of largely isolated individuals, which means collecting data over very wide geographical areas.

Besides just documenting age and proficiency, one also needs to document more precisely than in the past what (kinds of) structures are increasingly problematic as a function of age, and for what kinds of learners. All too often, age research does not distinguish between different learning problems or only looks at one very narrow learning problem. At the same time, it needs to be investigated how specific aptitudes and other individual characteristics such as various affective and conative variables (Dörnyei & Ushioda, 2009; Moyer, 1999, 2004) as well as identity issues (Caldas & Caron-Caldas, 2002; Cheung, Chudek, & Heine, 2011) mitigate the age factor, and what impact these interactions between age and individual characteristics have under different conditions of performance (cf. Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; DeKeyser et al., 2010; Granena, 2012).
The difference in learning processes that the previous point implies means that researchers need both tests for multiple predictors and outcome tests in multiple formats, so that different forms of knowledge can be teased apart, whether they be called implicit/explicit (DeKeyser, 2000; Paradis, 2009), conscious/unconscious (Abrahamsson, 2012), declarative/procedural (Ullman, 2001, 2005; see also Paradis, 2009), or representation/processing (Hopp, 2009, 2010; Juffs & Harrington, 1995; McDonald, 2006).

Finally, as already hinted at above, if studies aim to make any educational pronouncements, research needs to be conducted in the relevant educational contexts, for instance in foreign language classes if that is the sphere of interest for policy or practice applications (cf. Hyltenstam & Abrahamsson, 2001; Muñoz, 2006, 2007, 2008a).

This admittedly incomplete list of difficulties should be sufficient to illustrate how problematic it is to conduct good research on a question that every layperson knows (and often thinks to have the answer to). In the following section, I present some more specific suggestions for how to deal with these issues.

**Toward Better Sampling**

Sampling is one of the most neglected problems in research on age effects in SLA. While the research methodology in this area, as elsewhere in SLA (cf. Klein, 1998; Ortega, 2013) has become increasingly diversified and sophisticated, there has hardly been any attempt at systematic sampling from a relevant population. Two kinds of critical period research need to be distinguished here. On the one hand, there is the paradigm that identifies exceptionally good learners and examines what they can do as a function of age; the most systematic sampling in this paradigm was carried out by Abrahamsson and Hyltenstam (2009). On the other hand, most studies that purport to test the CPH do not preselect learners on the basis of proficiency and simply study correlations between AoA and proficiency for the whole range of both variables. Important points to be taken into account for future research of the latter type are:

1. All participants in a given sample should be native speakers of the same language, to avoid introducing variables due to the L1; of course one study could have several samples representing different L1s, chosen on a principled basis, and compared with each other.
2. The L1(s) should be distant from the L2, so that the number of learning problems is great and that one does not need to test rather subtle aspects of the grammar, for which there is often variation among native
speakers too, depending on regional dialect, level of education, and so on. This variation is often underestimated in acquisition research, which still tends to compare learners to the ideal native speaker (cf. Dąbrowska, 2012). When there is substantial variation among native speakers, then the native and nonnative ranges of variation are almost bound to overlap substantially, whether for the same reasons or not, yielding trivially predictable results. Comparisons with native speakers have been criticized in general (Grosjean, 1998; Ortega, 2013), and making such comparisons where native speakers vary substantially only exacerbates the issue.

3) All participants, from the time of immigration, should have spent most of their time communicating in the L2; clearly, if they have had insufficient input and practice, that is not a fair test of the age effect. It is, of course, impossible to tell exactly what percentage of L1 use is acceptable from this point of view; for design purposes one should err on the side of caution and limit the sample to those that have used their L1 very little since immigrating, in order to avoid the obvious criticism that any imperfections in the L2 are due to continued strong L1 interference.

4) Participants should be spread over a variety of socioeconomic levels to avoid the usual bias toward more highly educated people, whose results are not generalizable to the population of immigrants (e.g., in the sense of the relative importance of implicit and explicit learning). This point too only applies to research aiming at representing the whole range of learners, not for the type that zeros in on the best learners to establish how far they can go; the latter type of study will inevitably have more educated learners.

5) Length of residence (LOR) should preferably be at least 10 years, so that ultimate attainment can be assessed, not speed of learning or a mixture of the two. For the basic morphosyntax, LOR does not play much of a role after that (cf., e.g., DeKeyser, 2000; DeKeyser et al., 2010). For certain aspects of the target language, however, further development can be expected for a much longer period of time, for instance in the case of vocabulary, collocations (Abrahamsson & Hyltenstam, 2009; Flege et al., 1999), phonetic detail (Flege & Liu, 2001), or perhaps degrees of automaticity. Therefore, there should be sufficient variability in LOR to assess the importance (if any) of input and practice beyond the first 10 years for certain aspects of the L2. Moreover, LOR in itself does not mean much without sufficient quantity and quality of input (cf., e.g., Flege, 2009); this variable needs to be either controlled through participant selection or documented and used as an independent variable.
(6) Age at testing should not go beyond middle age, to avoid effects of cognitive aging on testing results, which constitute a serious risk (DeKeyser et al., 2010), and which are a different research issue altogether, having nothing to do with arguments for or against a critical period.

(7) The sample size should be sufficient to allow for statistical analysis, not just for the sample as a whole, but for each age-defined subgroup (e.g., 6–12, 12–18, 18–24, 24–30), which means at least about 20 participants per subgroup (assuming there are no covariates in the analysis).

(8) Participants should be spread fairly homogeneously over the age ranges within each group to allow for statistical analysis that is not handicapped by severely nonnormal distribution.

**Toward Better Instrumentation**

Given the various kinds of knowledge that learners can have about a second language (cf. DeKeyser, 2009) and how different testing formats can bring out these forms of knowledge to different extents (cf., e.g., Ellis, 2005; Jiang, 2012), it is important to assess the L2 learners’ knowledge in appropriate ways to elicit the kind of knowledge that matters to the theory. Because of what was said above about age effects being different for implicit and explicit learning and knowledge, and because age differences may affect representation differently from processing, testing formats are required that elicit and measure these various forms of knowledge separately. Time pressure may be enough to elicit information on processing rather than representation, and it may contribute to a reduced reliance on explicit knowledge (R. Ellis, 2005), but it is not enough to force the use of implicit knowledge only (cf., e.g., Paradis, 2009). Therefore other testing formats are required such as word monitoring tasks, eye-tracking during reading, the visual world paradigm, the measurement of event-related potentials, or neuro-imaging to examine the areas of the brain and therefore the kinds of knowledge involved (for two literature overviews of neurological research on L2 age effects, coming to rather different conclusions, see Birdsong, 2006; DeKeyser, 2012a). Which of these (combinations of) methods is best suited to the study will depend, of course, on the structures being elicited and the characteristics of the participants.

Other requirements for instrumentation are of a more general nature in the sense that they are important for any kind of psycholinguistic research. The structures tested need to be at the right level of difficulty to allow for substantial spread in the dependent variable, given the target populations (specifically with regards to L1 and level of education); there needs to be a variety of structures so
that generalization to (at least a well-defined subset of) the L2 grammar can be made; and there should be enough test items per structure to allow for reliable results for each structure, not just for “grammar” on average—needless to say this requirement is hard to satisfy at the same time as the previous one.

Moreover, as testing the various forms of knowledge in their purest form is virtually impossible for theoretical and practical reasons, a number of cognitive variables besides L2 proficiency may have to be assessed. As there is growing evidence of an interaction between age and aptitude (Abrahamsson & Hyltenstam, 2008; Bylund, Abrahamsson, & Hyltenstam, 2010; DeKeyser, 2000; DeKeyser et al., 2010), but as aptitude has been defined and measured in a variety of ways, and as aptitude is really an almost metonymical term for a whole set of aptitudes (cf., e.g., DeKeyser, 2012b; DeKeyser & Koeth, 2011), it is important to test a wider variety of aptitudes in interaction with AoA. This may mean casting the aptitude net much more widely than before and including aspects of aptitude that are rarely addressed in the verbal domain (such as aptitude for implicit learning; cf. Doughty et al., 2010). Given the multifaceted nature of L2 knowledge as described above, it is equally important to assess these multiple aspects of knowledge through multiple testing formats, in order to see the interaction on the outcome side too, that is, how AoA affects different forms of knowledge and their uses differentially. All of this is necessary for a full accounting of age-related qualitative differences in learning processes.

**Alternative Approaches**

Besides research with immigrants or with classroom learners, other forms of research on age effects are possible. At one extreme is computational modeling (Hurford & Kirby, 1999; Monner et al., 2012), which has the advantage that variables can be manipulated completely independently, eliminating the confounding of variables that is otherwise nearly impossible to avoid in research on age effects. Disadvantages, of course, are: (a) that it is hard to imagine at this point building a computational model of learning with the linguistic complexity of a full-fledged human language and, additionally, all the social science variables of interest and (b) that whatever can be shown to work in the computational model does not mean that this is necessarily what happens in human learners.

At the other extreme is completely naturalistic research, on the acquisition of a second language that is not taught anywhere, and in which no education is provided, at least not at the research site. That would eliminate education as a
variable and also seriously limit the influence of explicit learning. One could imagine, for instance, researching the acquisition of Quechua by speakers of Aymara or small Amazonian languages on the eastern slopes of the Andes. That phenomenon is disappearing rapidly, however, because native speakers of different indigenous languages now learn Spanish and use that as a lingua franca. Similar switches to international languages as the only L2 are occurring in other parts of the world. Moreover, if it were shown that learners past a certain age do poorly, it could always be blamed on lack of motivation to become a full-fledged speaker of a minority language.

What the entire SLA community probably can agree on, however, is that whether one does research with immigrants, or with classroom learners, or with indigenous languages as the L2, it would be preferable to have more longitudinal research. Such research is desirable in SLA/applied linguistics generally for a variety of reasons (see, e.g., Ortega & Byrnes, 2008; Ortega & Iberri-Shea, 2005) and would be particularly useful in this area as a means of disentangling variables by separating them in time and by avoiding having to rely on questionnaires administered many years after the facts, with the inevitable consequences for reliability and validity. On the other hand, it is, of course, extremely difficult to combine the need for longitudinal research with the need for large numbers of subjects tested on a large set of variables. An interesting attempt at longitudinal research in this area is Jia and Fuse (2007), albeit with a very limited number of participants.

**Conclusion**

It is of paramount importance for L2 instruction to learners of different ages, and in particular for the educational integration of immigrant children with limited L2 proficiency, not just to document quantitative and qualitative effects of age of learning, but also the nature of the learning processes involved and how disadvantages of later learning can be mitigated by stimulating positive interaction between available aptitudes and available forms of input/teaching for specific areas of grammar. At this point researchers are still far from reaching that goal. There is little research on age effects that meets very high methodological standards, no research whatsoever that meets all the standards outlined here, and almost no evidence that is clearly of educational relevance.

The challenges outlined above are daunting, in particular where data collection efforts are concerned. They may require some sort of a consortium between different institutions to facilitate data collection from individuals that
meet the criteria for participation throughout the country. Such a consortium, if it managed to include proponents of both sides of the critical period debate, would also contribute to sharpening the methodological focus by making the researchers agree on very specific theoretically motivated hypotheses and on methods for data collection and analysis that are the best possible fit for testing these hypotheses.

Revised version accepted 20 September 2012

Notes

1 I am using single quotations marks with the term ‘critical period’ here to indicate that I am aware of the fact that many researchers find this term unfortunate for various reasons, especially because it sounds too strong (giving the impression language acquisition is no longer possible after a certain age, or something along those lines). Other terms such as “sensitive period” do not necessarily resolve the problem, however. Those who do not accept the maturational interpretation of the widely documented age effects often object to that term too, so I am simply adhering to the most common terminology here, without any implications for how strong the age effect is or what the causes may be.

2 As one participant at the *Currents in Language Learning* conference pointed out, such individuals may suffer from some L1 attrition after many years in an (almost) exclusively L2 environment. That does not make them less valid L2 learners, however. In other words, I accept the point of view that people whose exposure to the L2 has been limited cannot be considered to be representative of what the L2 acquisition process can accomplish (taking age into account or not); I do not accept the point of view that when there is a certain degree of L1 attrition, then the learners are not representative of L2 acquisition processes. All that can be said is that they are not fully bilingual, and that they cannot answer any questions about the degree to which complete balanced bilingualism is attainable (again, taking age into account or not). Incidentally, the extent of L1 attrition because of limited exposure is also known to be a function of the ‘critical period’ (cf., e.g., Bylund, 2009; Bylund et al., 2010; DeKeyser, 2012a; Moon, 2012).

References


