SSLA, **24**, 481–488. Printed in the United States of America.
DOI: 10.1017.S0272263102003054

ON THE RELIABILITY OF ROBUSTNESS

A Reply to DeKeyser

Ellen Bialystok
York University

DeKeyser (2000) reports a study in which he examines three hypotheses concerning the existence of a critical period for second language acquisition. He concludes that his data support all three predictions and that the notion of a critical period is the best account of the data. However, there are problems in both his interpretation of the data and the issues raised in his discussion that undermine that conclusion. The present paper examines the evidence for the three hypotheses proposed by DeKeyser and argues that the data do not provide the necessary support for the interpretation that a critical period has influenced the results.

In a recent article, DeKeyser (2000) offers a set of data as evidence in support of the critical period hypothesis for second language acquisition and proposes counterarguments against some of the detractors of this position. DeKeyser's study is a partial replication of the methodology used by Johnson and Newport (1989) in their classic study and an extension of the argument developed by them in their traditional formulation of the problem. His primary innovation is to predict a differential influence of verbal ability on learning outcomes inside and outside the critical period and test that prediction by including a standardized measure of language learning aptitude based on the Carroll and Sapon (1959) Modern Language Aptitude Test (MLAT). He proposes three hypotheses: (a) there will be a negative correlation between age of arrival and performance, but some adult learners will score within the range found for

Address correspondence to: Ellen Bialystok, Department of Psychology, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada; e-mail: ellenb@yorku.ca.

© 2002 Cambridge University Press 0272-2631/02 $9.50

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

child learners; (b) those adult learners scoring within the range of children's success will have high scores on the verbal aptitude test; and (c) different age effects will be found for different grammatical elements. He concludes that all three hypotheses were supported and that the critical period hypothesis is vindicated. Although these hypotheses are largely confirmed in the data he reports, the logic that connects these results to his preferred conclusions is flawed. Moreover, his analysis includes misrepresentations of aspects of the data, misinterpretations of studies that do not report critical period effects, and unconventional theorizing about linguistic phenomena.

DeKeyser's first hypothesis was that that there would be a strong negative correlation between age of acquisition and scores on the grammaticality judgment test, a relation that was obtained in his data, $r = -.63$, $p < .001$. However, this linear relation is not decisive to the critical period hypothesis. The essence of a critical period is that there is discontinuity in learning outcomes that corresponds to a maturational point in the animal's development (Bornstein, 1989; Columbo, 1982). The discontinuity can be established either by the cessation of learning (strong CP) or a change in the slope of the learning curve (weak CP) after the close of the critical period. In either case, an overall correlation between age and learning across the boundary of the critical period, or demonstration of similar learning curves before and after the close of the critical period, is contrary to the conclusion that there is a critical period in learning.

A negative correlation between age of arrival and proficiency across the whole range of tested ages has been reported by many investigators; those in search of critical period effects (e.g., Johnson & Newport, 1989; Oyama, 1976), as well as those who deny such influences on acquisition (e.g., Bialystok & Hakuta, 1999; Bialystok & Miller, 1999; Birdsong & Flege, 2000; Birdsong & Molis, 2001; Flege, 1999; Flege, Munro, & MacKay, 1995; Flege, Yeni-Komshian, & Liu, 1999). Evidence for a lifelong decline in language learning ability is *not* evidence for a critical period but indicates a gradual change in some mechanism responsible for that learning domain. Johnson and Newport understood this and were clear in both predicting and reporting their results: "There should be a consistent decline in performance over age for those exposed to the language before puberty, but no systematic relationship to age of exposure . . . among those exposed to the language after puberty" (p. 79). Elsewhere, Newport (1990) continued to insist that no relation should be found between age and proficiency for older learners: "Once the organism is fully mature (that is, during adulthood), there should no longer be a systematic relationship between age of arrival and performance" (p. 20).

Johnson and Newport (1989) paid careful attention to these different patterns of correlation with age that they obtained for their older (17+ years old at arrival) and younger (0–15 years old at arrival) learners. They found that the relation between age of arrival and performance for the early learners was strongly negative ($r = -.87$), but the relation between these variables for the older learners was not significant ($r = -.16$). This pattern is essential to the

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

argument. In contrast, DeKeyser's data report no significant relation between age and proficiency when the correlations are calculated separately for the two age groups. Proficiency scores are generally flat across the early learners and possibly at ceiling for some portion of them. Therefore, these data both replicate an effect frequently found by those who reject the argument for a critical period and at the same time fail to replicate the pattern that is necessary to isolate the critical period as the cause of that effect. DeKeyser's discussion of this point is contradictory: He both claims to have replicated the data reported by Johnson and Newport and, aware of the crucially diverging pattern, suggests that the Johnson and Newport data include an artifact that overestimates the true value of the correlation.

The second hypothesis was that verbal aptitude would be influential in determining the ultimate proficiency level of late but not early learners. The hypothesis is surprising because it is offered as a test of Bley-Vroman's (1988) Fundamental Difference Hypothesis, yet this prediction neither follows from that theory nor provides critical evidence with which to evaluate it. The evidence offered in support of this hypothesis is that five of the six late learners who achieved high proficiency scores also recorded high verbal-aptitude scores and that correlation coefficients indicate a relation between aptitude and proficiency for late ($r = .33$, $p < .05$) but not early ($r = .07$, n.s.) learners. Both interpretations are flawed.

The first point regarding the quantitative count of late learners who obtained high grammaticality scores is suggestive at best but essentially indefensible as a statistical argument. Concluding anything from the data profiles of six participants is precarious, but the leap is particularly bold in this case because it is based on arbitrary criteria not included in the scoring system. DeKeyser, however, appears determined to demonstrate that these six participants who achieved proficiency levels usually reserved for younger learners possess some intellectual advantage that overrides their putative linguistic disabilities that befall older learners. Hence, to rationalize the single exception in which one high-proficiency participant did not score highly on the aptitude test, he points out that this individual "was doing postdoctoral studies in the natural sciences; this suggests that he must be of above-average analytic ability and that his aptitude test score is not indicative of his analytic abilities" (p. 514). It is dangerous enough to attribute levels of intellectual or analytic ability to individuals when standardized test scores *are* available; it is astonishing to presume to comment on such faculties in the absence of any supporting psychometric information. The argument is further compromised by noting that Appendix A includes the data from several late learners who achieved high aptitude scores but did relatively poorly on the grammaticality judgment test.[1] There is simply no basis for concluding a relation between these factors.

The second point concerns the statistical correlations. DeKeyser argues forcefully that different correlation patterns between proficiency and MLAT scores for early and late learners are evidence of different operational factors

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

in their language acquisition. However, the evidence is inconclusive. The premise of correlational analysis is that both the scores and the variances for the two variables are normally distributed, but the truncated distribution of scores for the early learners makes a significant correlation less likely than is the case for the older learners. It seems eminently reasonable that individuals vary in verbal aptitude—indeed, this is precisely the premise on which the verbal aptitude test was based. It would be surprising if people who distinguished themselves as more verbally gifted were not more successful at both acquiring high levels of competence in a second language and achieving high scores on a test measuring sensitivity to linguistic structure. Indeed, the predicted (and observed) correlation is arguably a tautology. The failure to demonstrate this general relation for the younger group is a statistical by-product of the distribution of scores for those learners. Because there was little variance in scores for the younger group, there cannot be a statistical correlation between the two scores.

It is not surprising that some proportion of the late learners in the sample achieved high levels of competence in the second language. Research has repeatedly identified late learners who master the task of language learning with nativelike success (Birdsong, 1992; Bongaerts, 1999; Cranshaw, 1997; Ioup, Boustagui, El Tigi, & Moselle, 1994; van Wuijtswinkel, 1994; White & Genesee, 1996). The combination of these cases with the reported evidence of a correlation between age of acquisition and proficiency that extends across the lifespan indicates that the most plausible explanation is that language-learning ability continues to decline throughout life and, like all cognitive abilities, is mediated by special ability in that domain. In fact, a specific type of cognitive processing model based on such factors as verbal ability is not necessary to explain the data. Elman et al. (1996) demonstrated a single linear function between age of acquisition and proficiency and explained that function using a connectionist network. There is nothing in these data that require a critical period in their account.

As a final comment on the evidence for DeKeyser's second hypothesis, the data are prima facie questionable in that they are based on scores from the Carroll and Sapon (1959) MLAT. Aside from being almost 50 years old, this test investigates a narrow and almost parochial definition of language aptitude. Its gradual disappearance as a research instrument reflects a lack of confidence in its psychometric properties. The study used only one subtest of this battery, which further undermines the validity of its interpretation. The translation of this instrument into Hungarian would do little to appease the concerns of a rigorous psychometrician concerned with the reliability and validity properties of the test instrument.

DeKeyser's third hypothesis predicted differential effects of age for the acquisition of different elements of grammar. Taking the notion even further, DeKeyser calculated the degree of relation with age by computing the correlation for each individual stimulus item rather than summing those items across categorical structures. Using arbitrary statistical boundaries for strength of re-

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

lation, the items were then grouped into three categories, and the grammatical elements frequently appearing in each were extracted as evidence for the vulnerability of differential effects of age on linguistic structures.

This procedure raises both statistical and theoretical concerns. The probability that an effect has occurred by chance is expressed in a statistical test by the value for *p*. The information about that probability is used by the researcher to decide categorically if the effect indicates a chance distribution of the dependent variable or evidence of a systematic factor. Although *p* values vary continuously along a dimension of probability, decisions do not. If the researcher decides that the probability of a correlation indicates that there is a systematic relation between two variables, then they are related. Comparisons of the strength of relation (in terms, for example, of being more significant) are technically not permissible inferences from these data, although such inferences are frequently made. In the case of the critical period hypothesis being tested here, it is particularly important to respect the categorical nature of the interpretation of the results. The point of the study is to determine if there is or is not a critical period influencing the results, so it is difficult to see how that factor would change its influence with different grammatical structures.

The theoretical problem is more worrying. Johnson and Newport (1989) also reported that the age effect observed for their various grammatical structures was different, but they offered neither an explanation for this effect nor an indication why a critical period hypothesis would predict it. Other studies, too, have found that the effect of age on learners' competence in making grammaticality judgments varies with the type of grammatical violation. For example, age-related differences in learning have been shown for the distinction between regular and irregular inflectional morphology, an observation not easily handled by the critical period hypothesis that awards native language mastery to young learners irrespective of linguistic structure (Birdsong & Flege, 2000; Flege et al., 1999). Bialystok and Miller (1999) reported significant differences in proficiency between structures that were constructed the same in the bilingual's two languages and those that were constructed differently. This difference based on structural similarity applied to learners at all ages of acquisition. It is clear that linguistic structure is a crucial factor in developing a complete model of language acquisition; it is not clear what the critical period hypothesis contributes to this issue.

DeKeyser's explanation for why some structures demonstrate age of acquisition effects and other do not is that the structures have different degrees of salience. Although ad hoc explanations clearly have a role in interpreting data, such explanations must still meet the scientific standards of evidence that apply to hypothesized effects. The explanation based on salience is wholly lacking in definition and verification. Not only is salience—the primary explanatory variable—not defined either operationally or theoretically, but also it is used in contexts that are decidedly nontechnical and unusual. DeKeyser states: "Pronoun gender errors are so irritating to native speakers

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

that they will almost always correct them when their nonnative interlocutors make such mistakes, even though overt correction of grammar errors is otherwise rare in adult native-nonnative interaction" (p. 516). No evidence is cited for this extraordinary claim. No hierarchy of irritation obviously presents itself (at least to this native speaker of English), but if forced to construct one, it is unlikely it would bear much resemblance to DeKeyser's intuitions.

The extension of this interpretation to an equation between perceptual salience and explicit learning is ungrounded in any evidence or logical explanation. The claim for such an equation is serious because the dichotomy between implicit and explicit learning has been the subject of considerable research, much of that based on well-conceived theory (see, e.g., the collection in Ellis, 1994). Invoking such constructs as part of an ad hoc rationalization is dismissive of a serious empirical and theoretical literature that has investigated these issues.

On the basis of the evidence presented around these three hypotheses, DeKeyser concludes that he has confirmed that a critical period functions to limit the L2 learning of older learners, although he refrains from attributing cause or mechanism to that critical period. Furthermore, he proposes that his data support the notions specified by Bley-Vroman's (1988) Fundamental Difference Hypothesis. In contrast, the present argument has been that none of the data offered in support of his three hypotheses provides the crucial evidence required for those conclusions. In fact, all three types of evidence can be used to support the contrary view—namely, that learning is not governed by a maturationally defined critical period that qualitatively changes the possibility for ultimate achievement past a designated boundary.

Negative correlations between age of acquisition and proficiency across the lifespan indicate only that there exists a gradual change in the learning mechanism with age (see Bialystok & Hakuta, 1999). A critical period requires a discontinuity in that function to signal a fundamental change in learning potential. This discontinuity can be expressed either through a change in the slope of the learning curve or the end of access to successful learning. The existence of older learners who achieve high (and possibly nativelike) competence is not new (see Birdsong, 1992), and it is not surprising that individuals who are more verbally talented are likely to both score well on standardized tests of language ability and achieve high levels of competence in L2 acquisition. Finally, different patterns of age effects for different elements of grammar (or for different languages, as reported by Birdsong & Molis, 2001) undermine the conclusion that a critical period constrains acquisition. It is difficult to imagine a compelling explanation for why a biologically driven mechanism would differentially exert its effect on different parts of speech. In contrast, differences between competence in mastering individual elements of grammar have been traced to the similarity between structures in two languages, making the likelihood that a particular structure will be mastered a problem for general learning mechanisms.

DeKeyser is correct that there remain many unanswered questions regard-

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

ing issues of age effects on the ability to master a second language. These questions, however, will be answered only when careful research is examined from the lens of coherent theory using a defensible methodology.

(*Received 14 August 2001*)

Note

1. I am grateful to an anonymous SSLA reviewer for pointing this out.

References

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: Erlbaum.

Bialystok, E., & Miller, B. (1999). The problem of age in second language acquisition: Influences from language, task, and structure. *Bilingualism: Language and Cognition*, *2*, 127–145.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language*, *68*, 706–755.

Birdsong, D., & Flege, J. E. (2000, November). *Regular-irregular dissociations in the acquisition of English as a second language*. Paper presented at the 25th Boston University Conference on Language Development, Boston, MA.

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

Bley-Vroman, R. (1988). The fundamental character of foreign language learning. In W. Rutherford & M. Sharwood Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 19–30). Rowley, MA: Newbury House.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133–159). Mahwah, NJ: Erlbaum.

Bornstein, M. H. (1989). Sensitive periods in development: Structural characteristics and causal interpretations. *Psychological Bulletin*, *105*, 179–197.

Carroll, J. B., & Sapon, S. (1959). *Modern Language Aptitude Test: Form A*. New York: The Psychological Corporation.

Colombo, J. (1982). The critical period concept: Research, methodological, and theoretical issues. *Psychological Bulletin*, *91*, 260–275.

Cranshaw, A. (1997). *A study of Anglophone native and near-native linguistic and metalinguistic performance*. Unpublished doctoral dissertation, Université de Montréal, Quebec, Canada.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

Ellis, N. C. (Ed.) (1994). *Implicit and explicit learning of languages*. San Diego, CA: Academic Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Flege, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 101–131). Mahwah, NJ: Erlbaum.

Flege, J. E., Munro, M., & MacKay, I. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*, 3125–3134.

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

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

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.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261–283.
van Wuijtswinkel, K. (1994). *Critical period effects on the acquisition of grammatical competence in a second language*. Unpublished master's thesis, Katholieke Universiteit, Nijmegen, the Netherlands.
White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research*, *12*, 238–265.

https://doi.org/10.1017/S0272263102003054 Published online by Cambridge University Press