**PB230 Replication**

**Study Preregistration**

**Study Title:** *Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language* by Jacqueline S. Johnson & Elissa L. Newport (1989, *Cognitive Psychology*)

# I. Introduction

The critical period theory of language acquisition states that there exists a stage in human development during which, compared to other life stages, learning a language is the easiest. This hypothesis, first established by Penfield & Roberts (1959), is derived from prior developmental research into the cognitive development of children by Piaget (1936), who suggested childhood presents an optimal developmental phase for language acquisition time frame to learn a language. Further research claimed that age plays a key role in our ability to learn a language, with Lenneberg (1967) asserting that our language-learning ability is limited past puberty. This is a replication study of Johnson & Newport (1989), which seeks to find out whether there is an age-related effect on mastering the grammar of a second or non-native language.

Researchers found that the native (*n* = 23, *M* = 268.8, *SD* = 2.9) and 3-7 age group (*n* = 7, *M* = 269.3, *SD* = 2.8) had no significant difference in English-language ability, *t*(10.4) = 1.28, *p* > .05. The difference between the 3-7 and 8-10 age groups (*n* = 8, *M* = 256, *SD* = 6) was significant, *t*(10) = 5.59, *p* > .01. The differences were also significant between the 8-10 and 11-15 (*n* = 8, *M* = 235.9, *SD* = 13.6) age groups, *t*(9.7) = 3.83, *p* < .01, and between the 11-15 and 17-39 (*n* = 23, *M* = 210.3, *SD* = 22.8) age groups, *t*(21) = 3.78, *p* < .01 (Johnson & Newport, 1989). These findings indicate a negative relationship between age of exposure to a language and performance in that language, supporting the view that language acquisition gets harder as one gets older.

Aside from my interest in the topic of the original study, my decision to replicate this study stems the fact that this study was not preregistered, and its first t-test did not yield a strong power, which threatens that test's ability to detect a true effect. In addition, I found no previous replications for this study despite it having been cited 1465 times.

**Research Question:** What is the relationship between age of exposure to English as a second language and performance in English grammar tests?

# II. Methods and Measures

**Power Analysis**

The original study doesn't specify Cohen's d for any group so I have calculated it by hand and checked with http://psychometrica.de/effect\_size.html to calculate power analyses. The original effect size for the t-test that compared the native and 3-7 groups was Cohen’s *d* = .17, 95% CI -.67, 1.02. The original effect size of the second t-test comparing the 3-7 and 8-10 age groups was Cohen’s *d* = 2.773, 95% CI 1.35, 4.2. For the third t-test comparing the 8-10 and 11-15 age groups, Cohen’s *d* = 1.912, 95% CI -3.095, -0.729. For the final t-test comparing the 11-15 and 17-39 age groups, Cohen’s *d* = 1.222, 95% CI 0.362, 2.082.

A power analysis using the pwr R package (Champely et al., 2020) indicates that the first t-test in the original study had a low power of 0.067. The original study was not well powered to detect a Cohen’s d = .17 in this first test. A power analysis of the second t-test, comparing the 3-7 and 8-10 age groups, reveals a high power of 1. The power analysis of the third test finds a power of 0.944, and the power analysis of the fourth test finds a power of 0.82. Calculations for original study power are found in the power notebook of the research compendium.

**Planned Sample**

This replication will recreate the study's groups. They correspond to participants’ age when they arrived in the UK: a group for those aged 3-7 years old at their arrival, one for those aged 8-10, aged 11-15, aged 17-39, plus a group for native citizens. The planned data analysis requires the use of a data frame to export test scores to a csv file on R studio, which is the software that will be used for analysis. The original study has unequal sample sizes across four t-tests. Unfortunately, it is not possible to make a data frame in R out of data containing the scores of uneven samples. R returns the error “arguments imply differing number of rows” when trying to do so. This is because the data frame would not have the same number of rows for each column, as some columns would contain more data from a higher number of participants. To avoid this problem, I must use identical sample sizes across the four t-tests.

To achieve a high power, the first test requires a modification to its sample size. Using a power analysis, I found that the necessary sample size to achieve 80% power to detect the effect size is 545 participants in native group and in the 3-7 age group. Due to the need to have equal sample sizes across all groups, all groups will have 545 participants, with a total of 2725 participants. While this large number of participants will increase the study’s cost, it is necessary for the analysis in R. In addition, there is a tradeoff between study cost and power – and having a sufficient power is crucial to this replication. Therefore, I choose to increase the study’s cost for the sake of having sufficient power across the t-tests.

Calculations for the replication study’s power are also found in the power notebook of the research compendium. In the PBS lab, studies are run for one week at a time and data collection will stop at the end of one week. In this sample, participants will be student participants and will be compensated with course credit. The average LSE sample is typically between 18 and 22 years old and is primarily European and female.

**Materials**

The test used in the replication will be the same as the test employed in the original study, which is based on one used by Linebarger, Schwartz, and Saffran (1983). This test will examine participants’ abilities in 12 areas of English grammar. For instance, ‘Wh-questions’ are tested through such tasks as picking the correct phrase between “When will Sam fix his car?” and “When Sam will fix his car?”, pronominalisation is evaluated by asking participants to choose between “Susan is making some cookies for us.” and “Susan is making some cookies for we.”, and past tense with phrases such as “A bat flew into our attic last night.” versus “A bat flewed into our attic last night.”. The other areas of English grammar that are tested are: plural, third person singular, present progressive, determiners, particle movement, subcategorization, auxiliaries, yes/no questions and word order. The highest attainable score in this test is 276. The dependent variable in this replication is the total score across these questions, which will indirectly indicate participants' mastery of English grammar, which, in non-native learners, the original study took as a sign of their ability to learn a second language. Consistently, this replication will as well consider grammar ability to be telling of participants’ language acquisition.

**Procedure**

Just as in the original study, participants will be tested individually in the LSE lab after completing an informed consent form. They will be told that if a sentence is in any way wrong, it must be regarded as ungrammatical. Participants will be asked to circle Y or N on their answer sheet for each question to indicate whether it is grammatically correct, without using cellular devices or dictionaries. The test sentences will be pre-recorded and played aloud at a moderate speed. Experimenters’ faces will not be visible so as to avoid revealing expectations or allowing for experimenter bias. Participants will be made aware of the nature of the experiment but not of the hypotheses. Participants will be told they can ask for a break at any time, and will be given a break halfway through the test.

# III. Confirmatory Hypotheses & Analysis Plan

**Confirmatory Hypotheses**

In line with the original study, the confirmatory hypotheses are as follows:

H1: There will be a negative relationship between the age of exposure to English and the performance in the test, lasting up to adulthood

H0: There will be no relationship between the age of exposure to English and the performance in the test

**Analysis plan**

**Confirmatory analyses.** The original study did not mention how it handled data exclusion or outliers, so I will follow the convention for doing so: missing data will be removed listwise from the data. The descriptive statistics (means and standard deviations) will be calculated for each group’s scores. To address the hypothesis, the mean scores of each pair of adjacent age group will be compared, two at a time, using independent-sample t-tests, of which there will be four in total: first the native speakers will be compared with the group of participants arrived in the UK at 3-7 years of age, then the 3-7 and 8-10 age groups, the 8-10 with the 11-15 age groups, and finally the 11-15 with the 17-39 age groups. I will report the mean, t value and degrees of freedom, and Cohen’s d, alongside the p value for t. A successful replication will find evidence of support for the alternate hypothesis.

**Exploratory analyses.** I also plan on conducting an exploratory analysis. I believe a factor that may bias the results is the schooling the participants have received, as there has been some debate regarding the matter with some finding that private schools lead to better academic performance (Adeyemi, 2014), and others finding no such effect (Carbonaro, 2006). I will run a moderation analysis to find out whether the type of schooling, private or public, the participants have received moderates the effect of age on second language acquisition.

# IV. Differences from Original Study

There are several differences between the original study and this replication.

1. In the original study, participants were Chinese and Korean citizens living in the USA. The replication study will recruit any foreign participants from the LSE in London, UK, which is a different setting and context. The choice not to restrict the sample population to Chinese and Korean citizens alone is made because, if the alternate hypothesis finds confirmation in the replication’s results, this will support the argument that the critical period theory is not culture dependent. Indeed, a limitation of the original study is that by choosing testing Chinese and Korean immigrants (which the researchers justified by highlighting the big difference in grammar rules that may make Chinese and Korean speakers struggle to learn English), results may reflect a pattern of second language acquisition that only applies to Chinese and Korean learners. In addition, if the critical period theory is correct, the same results should be found regardless of the native language of the replication’s participants.
2. For reasons detailed above, the number of participants is changed. The replication will have an equal number of participants across all groups. This presents two benefits. Firstly, this will allow for high power, and therefore for a sufficiently high change of detecting the researched effect. Secondly, a larger sample size may better resemble the general population than a smaller one, which means the replication is more likely to produce results that are reflective of the general population.
3. It is not known how participants in the original study were compensated. The replication’s participants will be compensated with course credit.
4. I will add one non-graded question to the grammar test, which will ask participants whether they have received most of their education in a private or public school, to collect data for the exploratory analysis.

The replication of this study in a different location and with a broader sample will provide evidence of the effect’s robustness and of its cross-culture generalizability (Nosek & Errington, 2020).

# V. References

Adeyemi, S. B. (2014). Comparative Study of Pupils’ Academic Performance between Private and Public Primary Schools. *World Journal of Education, 4*(4), 55–60. Retrieved from <https://eric.ed.gov/?id=EJ1158541>.

Carbonaro, W. (2006). Public‐Private Differences in Achievement among Kindergarten Students: Differences in Learning Opportunities and Student Outcomes. *American Journal of Education*, *113*(1), 31–65. https://doi.org/10.1086%2F506493.

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. R package version 1.3-0, <https://CRAN.R-project.org/package=pwr>.

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. Retrieved from <https://www.sciencedirect.com/science/article/pii/0010028589900030?ref=cra_js_challenge&fr=RR-1>.

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

Linebarger, M. C., Schwartz, M. F., & Salfran, E. M. (1983). Sensitivity to grammatical structure in so-called a grammatic aphasics. *Cognition*, 13, 361-392.

Nosek, B. A., & Errington, T. M. (2020) What is replication?. *PLos biology, 18*(3), e3000691.

Penfield, W., & Roberts, L. (1959). *Speech and Brain Mechanisms*. Princeton University Press.

Piaget, J. (1936). *The Origin of Intelligence in the Child*. London: Routledge & Kegan Paul.