[Software Review]

A Review of Duolingo's In-House Research Reports

Charlie Taylor (15cgt1@queensu.ca)
National Taitung Senior High School, Taitung, Taiwan

Publisher	Duolingo, Inc.
Language(s)	Arabic, Catalan, Czech, Danish, Dutch, English, Esperanto, Finnish, French, German, Greek, Guarani, Hawaiian, Hebrew, Hindi, Indonesian, Irish, Italian, Japanese, Korean, Latin, Mandarin, Norwegian, Polish, Portuguese, Romanian, Russian, Scottish Gaelic, Spanish, Swahili, Swedish, Turkish, Ukrainian, Vietnamese, Welsh.
Products offered by Duolingo	Duolingo, Duolingo for Schools, Duolingo English Test, Duolingo ABC, Duolingo Events, Duolingo Math, Podcast, Stories, Duolingo for Business
Level	Beginner to early intermediate
Target Audience	Foreign/additional language learners
Operating System(s)	Windows, MacOS, as a mobile app for iOS or Android, or on a mobile web browser
Price	Basic: Free Plus: From $6.99 US per month

Duolingo (duolingo.com) is a language learning app that is used by millions of language learners around the world. It calls itself "the world's best way to learn a language", and it boasts that "research shows that it works". Duolingo presents three in-house research reports on its website to support these assertions. The studies, among other things, claim to demonstrate that beginner-level Duolingo courses are equivalent to four semesters at university (Jiang et al., 2020), and that seven units of Duolingo courses are comparable to five university semesters in reading and listening (Jiang et al., 2021a). However, are these claims warranted by the studies' results?

These three documents are Duolingo-produced white papers, and the researchers were almost entirely Duolingo employees, which could open them up to accusations of conflict of interest; however, at least one of the studies has withstood peer-review. The results of Jiang et al. (2020) were published in an almost identical paper in the Foreign Language Annals (Jiang et al., 2021c). This certainly gives the research some credibility; however, there are methodological problems with these studies which could cast doubts on the researchers' conclusions.

Table 1

Methodological Problems

Title of study (Publication details in References)	Selection bias	Lack of validity	Selective data use	Internal contradictions
Duolingo efficacy study: Beginning level courses equivalent to four university semesters (Jiang et al., 2020)	Yes	Yes		
Seven units of Duolingo courses comparable to 5 university semesters in reading and listening (Jiang et al., 2021a)	Yes	Yes		
How well does Duolingo teach speaking skills? (Jiang et al., 2021b)	Yes	Yes	Yes	Yes

Selection Bias

The first methodological problem is selection bias. In order to compare the two groups fairly, the control group and experimental group should be randomly selected from a single pool so that each group will be made up of participants whose average characteristics are roughly similar. However, in these studies, this was not the case.

The participants in the experimental groups for all three studies were not randomly selected; they were drawn from a pool of Duolingo users who had already demonstrated a high level of motivation by completing an entire Duolingo beginner's course of their own volition. After that, they had to volunteer to fill out a questionnaire and then, if selected, volunteer to participate in additional language tests. For example, in Jiang et al (2021c), out of presumably tens of thousands of users of the most popular language-learning app in the world, only 8367 users were motivated enough to complete the entire beginner Spanish course, and were thus contacted to fill out the survey. Of these, only 813 (less than ten percent) responded. Of the 499 who were deemed eligible, only 195 started the test. This initial selection already suggests a triple level of bias in favor of highly motivated language learners. This is problematic given that motivation has long been acknowledged to have a significant influence on acquisition success: "Motivation is one of the main determinants of second/foreign language (L2) learning achievement…" (Dörnyei, 1994, p. 273). As such, one would expect highly motivated language learners to outperform low-motivated learners regardless of the pedagogical method.

Furthermore, the prospective participants were informed in the recruitment email that one of the incentives for participating would be that they would receive a certificate from the testing company verifying their language level (Jiang, personal communication, February 4, 2022). This incentive would almost certainly have appealed more to higher-performing students since students who did not feel they had made much progress would hardly be seeking a certificate to attest to the fact. Given that only a small percentage of those contacted agreed to take the surveys (e.g. 10-26 percent in Jiang et al., 2021c), any selection bias skewing the sample in favor of successful learners could have a significant impact on the average results.

Finally, the vast majority of the Duolingo users in the studies had completed a university degree. In Jiang et al (2020), for example, 90 percent of the Duolingo group had completed at least an undergraduate degree, with 51 percent holding graduate degrees.

The control group in two of the studies (Jiang et al., 2020; Jiang et al. 2021a), on the other hand, consisted of first and second-year undergraduate students, most of whom were not voluntary language learners. For example, in Tschirner (2016), which was the main study used by the researchers as the control group, well over 80 percent of the participants came from Michigan State University, University of Utah, and University of Minnesota. All of these schools have mandatory foreign language requirements for undergraduate students, equivalent to four semesters of study. At the University of Minnesota and University of Utah, the minimum threshold for graduation in these courses is a C- average (University of Minnesota, n.d.; University of Utah, n.d.). In fact, these schools explicitly state on their websites that they do not expect their language students to become fluent (University of Minnesota, n.d.; University of Utah, n.d.), so it is fairly safe to say that these classes are not exclusively populated with overachieving language learners.

To their credit, the researchers themselves acknowledged the problem outlined above: “These differences may put into question the comparability of the learners and the learning that took place in these two very different settings (Jiang et al., 2021c, p. 993). However, they still believed that the comparison of general undergraduate students to an elite, well-educated, and mature group of highly motivated, self-confident learners was sufficient evidence to claim that beginning level Duolingo courses are the equivalent of four semesters of university (Jiang et al., 2020) and that seven Duolingo courses are equivalent to five semesters at university (Jiang et al., 2021a).

In Jiang et al (2021b), some members of the Duolingo group might have had further advantages over the American university control group, in that they were drawn from various countries. For example, five percent of Spanish learners were from Italy. Native Italian speakers would have a huge advantage over native English speakers when learning Spanish, a fact that was born out in the results of the study. Even within the Duolingo group, Italians unsurprisingly nearly doubled the average listening score of study participants from the US: 6.80 to 3.64.

Likewise, for whatever reason, the Dutch Duolingo users outperformed the American Duolingo users in French 5.4 to 4.0 in listening, while Indian Duolingo users underperformed the Americans with an average of 2.25 (Jiang et al., 2021b). As such, it is clear that different linguistic and cultural backgrounds have a significant impact on a student’s ability to learn a particular target language. Nonetheless, the researchers compared the language learning results of the diverse, multinational, multilingual Duolingo learners with the relatively linguistically homogenous group of largely native English-speaking American undergraduate students. (These universities offer exemptions or other considerations to students whose first language is not English (Michigan State University, n.d.; University of Minnesota, n.d.; University of Utah, n.d.); hence the assumption that most of the students in the control group were native English speakers.)

Lack of Validity

On top of the selection bias, the research suffered from a lack of validity in that it is not clear whether the linguistic proficiency measured by the tests was entirely attributable to Duolingo use. In all studies, the researchers claim that the participants used Duolingo as “their only learning tool” (Jiang et al., 2020, p. 1; Jiang et al., 2021a, p. 1; Jiang et al., 2021b, p. 1). However, this claim is not supported by the description of methods. While would-be participants were eliminated from the candidate pool if they reported taking formal classes or using any other apps to learn the target language, this is quite a narrow definition of “learning tool”. Potential participants were not screened for other language learning activities like extensive reading, extensive listening, conversations in the target language with native speakers or other language learners, for example.

Further extracurricular input is likely in the form of professional or social contacts, or travel, which were widely noted by the Duolingo participants as reasons for studying the language. In Jiang et al., (2021a), for example, the researchers noted that 30 percent of Duolingo users in the study reported learning the language for social purposes, and around 20 percent for job-related purposes. It is conceivable, and even likely, that many of these learners were using the target language socially or professionally concurrently with Duolingo, rather than waiting patiently until they completed their Duolingo courses. Well over half the participants reported learning the target language for travel, but they were not asked how often they visited the target destination while taking the Duolingo course. Any of these scenarios would affect the Duolingo group participants’ performance on the test, but instead all their gains were attributed to Duolingo use.

While the authors acknowledge these shortcomings in an endnote (Jiang et al., 2021c, p. 994), they argue that the university students might also have been getting additional input while studying; however, it is arguably more likely that highly motivated Duolingo users who are learning a language voluntarily would seek out additional sources of input than university students taking a required language credit as a condition of graduation.

The speaking study (Jiang et al., 2021b) did not compare the results of Duolingo users to those of university undergraduate language learners, but rather compared them to Duolingo’s targets. This study also had some elements that could impact validity. The researchers said that in order to avoid including participants who were immersed in the target language culture, they eliminated potential participants who lived in countries where Spanish or French were official languages or “widely spoken”. However, the list of countries where Spanish is widely spoken did not include the second-largest Spanish-speaking country in the world; the US is home to 42 million Spanish native speakers, accounting for 13.5 percent of the population (United States Census Bureau, 2019).

Selective Inclusion of Data

Out of the 173 users who started the French-speaking test in Jiang et al (2021b), only 102 (less than 60 percent) were included in the data analysis. Some 28 percent of the French test takers did not receive a score, presumably—according to the researchers themselves—because their pronunciation was so bad it could not be understood by the testing program. Rather than include these students in the results and attribute this incomprehensible language production to a shortcoming on the part of Duolingo, the researchers simply removed them from the study and calculated the average speaking performance based on those students who did manage to produce comprehensible output after completing the Duolingo course. The researchers acknowledged in the limitations portion of their paper that had these participants been included, this could indeed have lowered the outcome significantly. This is an understatement.

Internal Contradiction

The authors cite Krashen (2014) as saying Duolingo is "good" (Jiang et al., 2021a, p. 8; Jiang et al., 2021c, p. 990) for developing decontextualized linguistic knowledge. (He does not say this, but rather says Duolingo is "based on conscious learning" (Krashen, 2014, p. 14), and goes on to argue that conscious learning does not lead to true language competence.) However, the authors justify Duolingo's focus on decontextualized linguistic knowledge by citing DeKeyser's (2015) claim that—according to Skill Acquisition Theory—"practice and repetition can lead to proceduralization of explicit knowledge" (Jiang et al, 2021c, p. 991). However, in Jiang et al (2021b), the researchers contradict this claim by blaming the relatively low vocabulary scores in the speaking tests on the fact that the Pearson test which was used requires a "high level of automaticity in speech production" (p. 8). Given that the vast majority of Duolingo users, according to the researchers' studies, want to use the target language for communicative purposes (socially, professionally, or for travel), it seems determining whether the explicit knowledge gained from Duolingo is transferable into implicit, usable knowledge is a question of crucial importance. If the low results on the vocabulary sections of the speaking tests do indicate a lack of automaticity development on the part of Duolingo users, this should not be shrugged off and must be seen as a critical failure of Duolingo to effectively achieve the language acquisition goals of most of its users.

Research Strengths

The above criticisms of the Duolingo reports are not intended to suggest that the research is entirely without merit. In fact, the findings are of value for a number of reasons. Firstly, the research topic is particularly timely. Millions of global users of language learning apps deserve to know whether they are using their time effectively, and practitioners need to know whether such apps are a useful tool. The Duolingo studies address questions which are worthy of a more robust examination than they have previously been given.

Second, while there are shortcomings in the research, many of these are acknowledged by the researchers themselves. For example, they note that the high number of French speaking test scores which were eliminated from the study could cast doubt on the results (Jiang et al., 2021b). Furthermore, there is no lack of transparency. The corresponding author was very forthcoming when contacted by the reviewer for further information about the study, and the researchers go to great lengths to include all relevant data, even that which could be used to call their conclusions into question.

Finally, there is no indication that the researchers set out to find a result that was favorable to their employer. Throughout all studies they upheld a degree of academic rigor. The sample sizes of are significant, demographic data were carefully collected and reported, and much of the methodology is sound. For example, the researchers use standardized tests that are given by a third party to eliminate any possible bias in the test design or administration. Furthermore, behaviors flagged as suspicious during the testing were grounds for elimination from the study (Jiang et al., 2021b).

Conclusion

Given that the Duolingo users in all three of the studies may have been passively or actively getting comprehensible input from other sources, and/or using the target language socially, professionally, or for travel, it is not clear how much of the language proficiency demonstrated by the Duolingo users can be attributed to their Duolingo use. Furthermore, in the two studies that compared Duolingo users to a control group of American undergraduate students, the differences between the two groups in terms of motivation, education, life experience, etc., preclude any accurate comparison, and certainly the claims that Duolingo courses are equivalent to a specific number of university semesters is not supported by these studies. What is clear from the studies is that the Duolingo users did make progress in their language learning, and it is very likely—though not certain—that at least some of the progress can be attributed to their Duolingo use; however, it is just as unlikely that all of it can. As such, all that can be recommended is further research, and until there is more robust evidence supporting the language acquisition benefits of Duolingo, language learners should hedge their bets by not relying solely on the self-proclaimed "best" way to learn a language.

References

DeKeyser, R. M. (2015). Skill acquisition theory. In J. Williams, & B. VanPatten (Eds.), Theories in second language acquisition: An introduction (2nd ed., pp. 94–112). Routledge.

Dörnyei, Z. (1994). Motivating in the foreign language classroom. *The Modern Language Journal, 78*(3), 273-284. https://doi.org/10.2307/330107

Jiang, X., Chen, H., Portnoff, L., Gustafson, E., Rollinson, J., Plonsky, L., & Pajak, B. (2021a). *Seven units of Duolingo courses comparable to 5 university semesters in reading and listening* [White paper]. https://duolingo-papers.s3.amazonaws.com/reports/duolingo-intermediate-efficacy-whitepaper.pdf

Jiang, X., Rollinson, J., Chen, H., Reuveni, B., Gustafson, E., Plonsky, L., & Pajak, B. (2021b). *How well does Duolingo teach speaking skills?* [White paper]. Duolingo https://duolingo-papers.s3.amazonaws.com/reports/duolingo-speaking-whitepaper.pdf

Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., & Pajak, B. (2021c). Evaluating the reading and listening outcomes of beginning-level Duolingo courses. *Foreign Language Annals, 54*, 974–1002. https://doi.org/10.1111/flan.12600

Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). *Duolingo efficacy study: Beginning level courses equivalent to four university semesters* [White paper]. Duolingo http://duolingo-papers.s3.amazonaws.com/reports/duolingo-efficacy-whitepaper.pdf

Krashen, S. (2014). Does Duolingo "trump" university-level language learning? *The International Journal of Foreign Language Teaching, 9*(1), 13–15. http://sdkrashen.com/content/articles/krashen-does-duolingo-trump.pdf

Michigan State University. (n.d.) Chapter IV: Graduation requirements. https://acadgov.msu.edu/print/1120#:~:text=For%20admission%2C%20Michigan%20State%20University,non%2Dnative%20speakers%20of%20English

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*, 201–223. https://doi.org/10.1111/flan.12198

United States Census Bureau. (2019). American Community Survey. https://data.census.gov/cedsci/table?g=0100000US&tid=ACSST1Y2019.S1601

University of Minnesota. (n.d.) Second language frequently asked questions. https://cla.umn.edu/undergraduate-students/requirements-policies/second-language/second-language-frequently-asked

University of Utah. (n.d.) BA language requirement (BA) criteria. https://languages.utah.edu/language-requirements/ba-language-requirements.php#:~:text=Student%20Requirements,the%20Bachelor%20of%20Arts%20degree.