How Long Does It Take for a Voice to Become Familiar?

A replication by Adam Maghout & Polina Revina

Research questions

The study under scrutiny [1] investigates the development of voice familiarity over time and its impact on speech intelligibility under different conditions. To investigate these research questions, participants were exposed to different voices in different tasks for varying durations, in noisy or quiet environments.

Research questions:

- How much training on a voice is required to derive the maximum intelligibility benefit?
- How does the type of training influence how voices are learned?
- How do training conditions (voices presented alone or in the presence of multi-talker babble) influence voice recognition?

Methods in Holmes et al.

Four **Planned Mixed ANOVAS** were conducted in the study:

- **Two-way mixed ANOVA for training scores** where the between-subject DV was training group (quiet, babble) and the within-subject DV was familiarity (most familiar, moderately familiar, least familiar).
- **Two-way mixed ANOVA for d′ values in the explicit recognition test** where the between-subject DV was training group (quiet, babble) and the within-subject DV was familiarity (most familiar, moderately familiar, least familiar).
- **Three-way mixed ANOVA for Percentage of correct responses on the speech intelligibility task** where the between-subject DV was training group (quiet, babble) and the within-subject DVs were familiarity (most familiar, moderately familiar, least familiar, unfamiliar) and Target-to-Masker Ratio (TMR: −6 dB, +3 dB).
- **Two-way within-subject ANOVA for comparing performances across tasks using z scores** where the within-subject DVs were familiarity (most familiar, moderately familiar, least familiar, unfamiliar) and task (speech intelligibility, voice recognition).

For each ANOVA effect, effect sizes and associated confidence intervals were calculated. T-tests were then performed to determine differences between variable levels. Finally, an exploratory unplanned two-way within-subjects ANOVA to look for voice learning over the course of the study was conducted. No correction was used to take into account multiple testing.

Replication target

One of the study's main findings is that **a short ten-minute training session is sufficient to improve intelligibility and voice recognition**. This conclusion is driven by a significant effect of familiarity on speech intelligibility and a voice recognition higher than chance demonstrated by participants in the study. Furthermore, higher familiarity, created by a longer exposure to a voice, lead to increased speech intelligibility in the study, although this beneficial effect was not observed for voice recognition. The aim of this replication was to **validate these reported effects by replicating the statistics from the four main planned ANOVAs of the study**. This is relevant as it allows the confirmation of the links drawn between each task.

The target p-values and F-statistics for each individual and interaction term are presented in table 1. Effect sizes are also provided in the article but were not targeted here, in part because the method for obtaining them was left undefined. Finally, in absence of clear summary statistics for each task, the figures 1 and 2, labelled 'Fig. 2' and 'Fig. 3' in the original paper, were replicated to illustrate the mean and standard deviation between observations.

Replication results

We replicated all key findings except for effect sizes because authors did not describe the methods they used in detail. However, the key findings of the paper still hold up. Most of the statistics were easily replicable and, where there were slight discrepancies, these did not affect the conclusions. These may have arisen due to software differences or minor oversights on the authors' part. Nevertheless, observed p-values are consistent with the degrees of freedom that they reported, so this was unlikely malicious. Graphs and a table that are presented below summarize findings in the original paper and our replication.

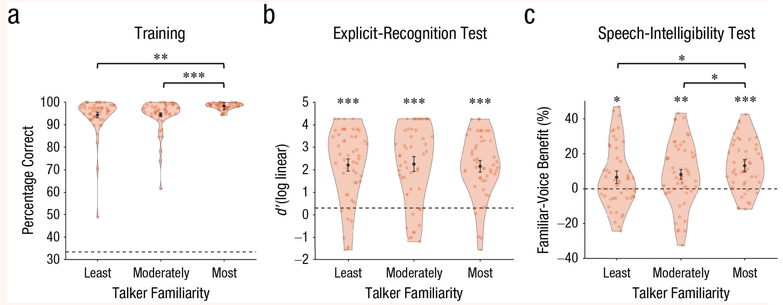

Figure 1 (a): Task performance in each of the three familiarity condition, Holmes et al.

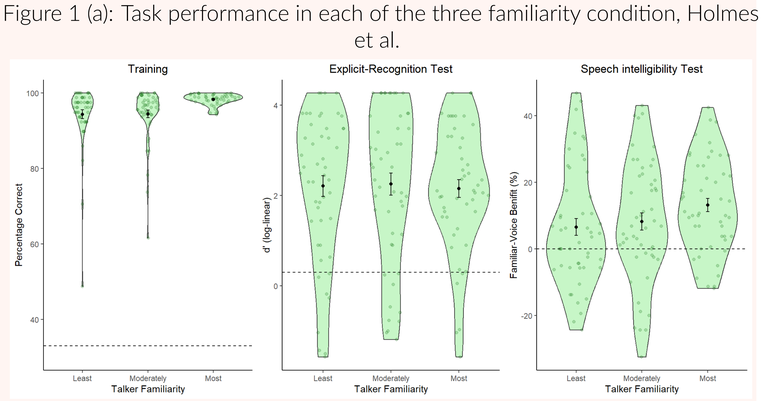

Figure 1 (b): Task performance in each of the three familiarity condition, replication

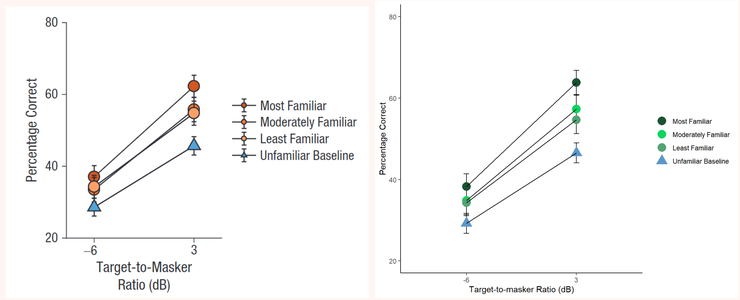

Figure 2: Percentage of correct responses on the speech-intelligibility test as a function of target-to-masker ratio and familiarity condition, Holmes et al. (left) and replication (right)

Verification process

What was difficult to replicate?

- In the dataset for the training task, there was a missing column for training condition and no participant ID numbers to make matching observations across datasets possible. However, since a pattern for training condition was otherwise present in all the other data frames (25 first rows – quiet, 25 last rows – babble), we assumed that the participants were ordered identically. The results are coherent with those provided so the assumption probably holds.
- Methods used in the paper were not described in detail, and some steps needed clearer explanations. For example, they did not describe what method they used for effect size estimations, so we could not replicate the results.

What made the replication process easy?

- A clear description of data frames and variables. The description was listed in one text document, it was easy to navigate.
- The authors provided lots of statistics for every result, which was useful for comparison with replication.

ANOVA	Effect	Reported F	Replicated F	Reported p	Replicated p
1	G	.36	.36	.55	.55
	F	12.64	12.64	< .001	< .001
	G x F	.21	.21	.81	.73*
2	G	< .01	< .01	.95	0.95
	F	.12	.12	.89	.89
	G x F	2.58	2.58	.08	.08
3	G	.20	.20	.66	.66
	F	10.49	10.49	< .001	< .001
	T	140.96	140.96	< .001	< .001
	G x F	.17	.17	.91	.89*
	G x T	1.19	1.19	.28	.28
	F x T	7.47	7.47	< .001	< .001
	G x F x T	.15	.15	.93	.93
4	F x Ta	4.55	4.55	.017	.017

Table 1. Comparison of values found in the four planned ANOVAs to those found by replication, where G = Group, F = Familiarity, T = TMR and Ta = Task. * means the value is different than in the paper. ** means the value is different and has an impact on the acceptance/rejection. Not all values are provided for the fourth ANOVA.

Conclusion

- We replicated the main findings of the paper but as the authors did not describe their methods in detail, we failed to replicate effect sizes. However, the authors' conclusion still holds up.
- Description of data structure and variables helped us to conduct replication without any troubles. However, there were some problems with one dataset that had a missing column. It did not affect the final results but was not good for the replication process.

References

[1] Emma Holmes, Grace To, and Ingrid S Johnsrude.
How long does it take for a voice to become familiar? speech intelligibility and voice recognition are differentially sensitive to voice training.
Psychological Science, 32(6):903–915, 2021.

Github: https://github.com/Adlieben/Replication