Commentary on "Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs"

Daniel J. Levitin

Musicae Scientiae 2013 17: 350
DOI: 10.1177/1029864913490633

MUSICÆ SCIENTIÆ

Musicae Scientiae
17(3) 350–355

DOI: 10.1177/1029864913490633
msx.sagepub.com
SAGE

Commentary on "Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs"

Daniel J. Levitin
McGill University, Canada

Ask the average person to describe how science works and it becomes clear that they subscribe to two pervasive myths. The first is that science is neat and tidy, that scientists never disagree about anything. The second is that a single experiment tells us all we need to know about a phenomenon, that science moves forward in leaps and bounds after every experiment is published. It is important to dispel these myths by doing just what Frieler et al. (2013) have done here – to take findings from the literature, attempt to replicate them, and engage in an open dialog about the nature of experimental findings.

My interpretation of the data they present here is that they replicated the original Levitin (1994) finding in their Frankfurt sample, with a nearly identical effect size. Their other five samples replicated it with a much more modest effect size. This raises the interesting scientific question of "what's going on?"

A proper replication should precisely repeat every aspect of the procedure – except that at least twice as many participants should be tested as in the original study (Tversky & Kahneman, 1971). If the new results are substantially similar to the original study, the interesting work begins of changing variables one a time to see which ones can "break" the effect. If the new results are substantially different from the original study, this is even more interesting as we try to determine whether some latent variable was driving the effect, or instead whether the original effect might have been a statistical anomaly.

As Frieler et al. state, their six-laboratory study replicates the basic finding of my 1994 paper, but with what amounts to a smaller combined effect size. This raises four possibilities:

(1) The 1994 sample showed a larger effect size than the true effect size that exists in the population;
(2) Five of Frieler et al.'s six samples showed smaller effect sizes than the true effect size that exist in the population;
(3) The 1994 sample provided a good estimate of the true effect size, and Frieler et al.'s methods are divergent from mine in at least one critical factor, accounting for the observed smaller effect size in their study;

Corresponding author:
Daniel J. Levitin, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC H3A 1B1, Canada.
Email: daniel.levitin@mcgill.ca

(4) Despite taking great care to employ identical methods, somehow the Frankfurt group (whose effect size was similar to Levitin, 1994) ended up employing methods that were more similar to the 1994 study, and that differed in some critical way from the other five laboratories.

Comparing carefully their report and the 1994 report, a number of key differences in methodology could contribute to different effect sizes and slightly divergent results. Here, I present them one at a time, organized by the mental system or methodological issue that may be responsible for the disparities.

Methodological differences between the replication and the original study

Mental imagery

An important component of the 1994 experiment was preparation time, a period during which participants were explicitly instructed to take time to form a mental auditory image of the song they intended to sing. Imaging times were typically 10 seconds.

A mental image or representation has two major components: a deep representation that draws on information in long-term memory, and a surface image that depicts the object in long-term memory (Kosslyn, 1981). Surface images are presumably held in an auditory buffer where they "are transient and begin to decay as soon as they are activated" with a capacity "defined by the speed with which parts can be generated and the speed with which they fade" (p. 50). The surface image is likely to be lower in resolution than long-term memory, what Kosslyn calls graininess. Individuals may well have an accurate long-term memory representation that is not accessed if the surface image is improperly formed. In short, the surface image is subject to decay, distortion, and interference effects that do not affect the long-term memory trace.

Thus the preparatory image formation period in the 1994 may have been crucial for participants to form a stable and accurate mental trace to match with their voices. Frieler et al. do not report on this, and so it appears that their participants were not given this same opportunity; we therefore don't know if the participants were attempting to match pitches to a rich and detailed, stable surface image.

Memory: Competing traces

Given the lability of the surface image, another issue centers on the specific songs that qualify as stimuli. Many popular songs are performed in different keys, either by the original artists who may alter the key for live performance, or by subsequent artists who alter the key to better fit their vocal range. Such cases present two experimental difficulties. First, the participant may experience competition between two or more memory traces and be unsure as to which key, and hence which pitches, constitute the target to be produced. Second, the experimenter can't be sure which memory trace the participant has accessed, and therefore can't be certain what pitches to compare to the participant's production.

The 1994 protocol called for the exclusion of any songs that existed in recordings in more than one key. Examples of such songs would be "Yesterday" by The Beatles, or "Just the Way You Are", by Billy Joel, songs that have been recorded by multiple artists in multiple keys. Although the original artists' renderings are certainly canonical ones, the neuroscience of pitch perception strongly suggests that competing versions of the songs in different keys could

easily cause interference in the memory trace, given that pitch information is carried throughout every stage of the auditory system (Kolb & Whishaw, 1990). Frieler et al. evidently did not prune the stimulus set in this way.

Memory: Familiarity with stimulus materials

According to multiple trace memory theory (MTMT; Goldinger, 1998; Hintzman, 1986), each time we hear a song, it lays down a memory trace in the brain. MTMT states that these memory traces contain veridical perceptual information. We are not able to recall every perception we've ever experienced because a single memory trace is only weakly activated in neural networks, whereas multiple repetitions of a stimulus strengthen the activation of its trace. A prediction of MTMT is that a greater number of stimulus repetitions leads to a greater probability of accurate recall of a given stimulus. (The issue here has to do with the signal-to-noise ratio of the memory trace being accessed. Familiarity increases the strength of the signal, compared to the noise of competing, irrelevant memory traces that are co-activated during an act of retrieval.)

Given this, an important point is that at the outset participants in the 1994 study were asked to sing their *favorite* song, and, presumably, they've heard their favorite song more times than their second favorite song (and, moreover, attach more emotional weight to it, further strengthening the representation). Thus, a plausible explanation for why participants in the 1994 experiment performed less well on Trial 2 than Trial 1 is that they simply knew the song in Trial 1 better. The claim of the 1994 paper was not that participants could reproduce every song they had ever heard with absolute pitch memory but, rather, that they could do so for a song they knew very well (the "learned melodies" of the paper's title). Participants in the Frieler et al. study were not asked to produce their favorite song; rather, they were asked to produce a song they were "very familiar with". It is safe to assume that less familiar songs are remembered less accurately, and it is difficult to judge, in absolute terms, how familiar the participants in the replication study were with the songs they sang.

Stimulus selection

In the replication, the authors provided a list of songs that were hits over the last 50 years and asked the participants to select from the list. Although it seems *prima facie* reasonable, this procedure assumes that their participant population – university students, and many of them music majors – has similar tastes to the public at large.

The 1994 protocol made no assumptions about what songs would be well known by the participants. A separate norming study was conducted using the same population from which experimental participants would eventually be drawn. These participants answered a questionnaire containing the names of 50 well known popular songs used in previous musical memory research, and in addition used a free response format to list 10 additional songs they knew "well enough to hear them playing in your head". From this list, the 75 songs that received the highest count were selected. Songs that were readily available in different keys (see *Memory: Competing traces* above) were eliminated, as were songs with tight vocal harmonies that would have created an ambiguity about which musical part the experimental participant was attempting to sing. This careful culling of the original list eliminated 17 songs, resulting in 58. The compact discs containing those 58 songs were used for the study, resulting in well over 600 songs potentially available to participants (because each compact disc contained 10–15 songs).

Short-term memory biases

Mental images tend to be delicate and easily overpowered by veridical perceptual input. Listening to a piece of music, or any tone with definite pitch (such as telephone ring, hum of a refrigerator or fan, etc.), can easily set up a pitch reference. When asked to sing a song from memory, participants might use that reference as an anchor similar to the anchoring and adjustment procedure described by Tversky (1974). The 1994 protocol was careful to ensure that participants remained isolated from any sounds for 30 minutes prior to participating in the study. (This detail was left out of the original report at the request of the journal's editor, due to space limitations.)

Stress

Finally, amateurs find singing out loud in front of someone to be a very stressful experience. To do so in a laboratory while being recorded is especially stressful. Recognizing this, the 1994 protocol required that the microphone used for recording be hidden underneath the table in the testing room; participant permission for recording was obtained only after the experimental session was over (this was approved by the human participants review committee). This subtle but important aspect of preparing the experimental environment may have contributed to the 1994 participants feeling relaxed. It is further possible that subtle features in the way the experimenter interacts with participants – body language, overall demeanor – contribute to the participant feeling relaxed or tense. Participants are unlikely to perform their best if feeling stressed. It would be interesting to administer an instrument that indexes stress, such as the State Trait Anxiety Inventory (Spielberger et al., 1983) to use as a covariate in analyzing performance mistakes.

Pitch analysis

There is always some degree of muscle memory necessary for the vocal generation of pitch (Cook, 1991; Ward & Burns, 1978). Yet muscle memory itself is insufficient to allow accurate pitch production. Even trained singers typically miss the starting pitch of a note and then glide into it after they receive auditory feedback and correct their mistake (Campbell & Heller, 1979; Murry, 1990).

The original 1994 experiment omitted the first 100 ms of the participants' vocal production from the analysis, based on the work of Murry (1990), who found that trained singers are typically off by as much as 2.5 semitones during the first 100 ms of a vocal production.

This methodological difference alone could account for the difference in results between the 1994 paper and the current replication. Singers in the Frieler et al. cohort may have done what many singers do, missing the correct tone initially and then self-correcting. It is unknown how the different pitch detection methods employed by Frieler might have coded such sound files, but this could account for the reduced effect size by adding measurement error.

Quantifying the strength of the replication

Finally, it is worth asking whether or not Frieler et al.'s replications show evidence that the effect originally reported in the 1994 paper exists. All six of the Frieler et al. laboratories found an effect in the same direction as the original 1994 report: individuals unselected for musical ability tend to be able to produce from memory (in a free recall task) the pitches of familiar songs.

If the effect originally reported did not exist, we would not expect to find it replicated in six independent laboratories. By the sign test, the probability of six laboratories showing an effect in the same direction is $1/2^6$ or $p < .02$.

Looking at this another way, Edgington (1972) specifies a method for combining the results of independent studies to obtain an omnibus p value (incorporated into Rosenthal's 1978 method). When applied to the results of the six European laboratories, this yields $p < .001$. If the results of the 1994 study are included, this yields an even more significant $p <. 0001$.

Discussion

There exists a pressing need for replications in psychology, the publishing of failures to replicate, and negative results. Based on the statistical analysis in the previous section, I consider this a successful replication of the original 1994 study.

There remains the question of why relatively large variability in effect sizes existed across the six laboratories involved in the replication. I've reviewed seven possible explanations, each addressing a divergence from the methods originally employed. These concern mental imagery, two kinds of memory, stimulus selection, short-term biases, stress and pitch analysis. Future work might attempt to replicate the methods of the original study more strictly (holding these seven variables constant) in order to determine which variables are critical for the underlying effect.

I am grateful to Frieler et al. for renewing interest in the 1994 paper, and for the time and care they put into conducting a study that raises a number of interesting scientific questions.

Acknowledgements

I am grateful to Perry Cook, Stephen Kosslyn, Michael Posner, Lewis Goldberg and Heather Bortfeld for their comments on a draft of this article. I remain grateful to Roger Shepard for supervising the collection and analysis of the data for the original 1994 paper, a version of which was submitted as my undergraduate honors thesis at Stanford University.

Funding

Preparation of this report was made possible by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Campbell, W. C., & Heller, J. (1979). Convergence procedures for investigating music listening tasks. *Bulletin of the Council for Research in Music Education, 59*, 18–23.

Cook, P. R. (1991). Identification of control parameters in an articulator vocal tract model, with applications to the synthesis of singing (Doctoral dissertation, Stanford University). *Dissertation Abstracts International, 52*, 419B. (University Microfilms No. 91-15, 756).

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *Journal of Psychology: Interdisciplinary and Applied, 80*(2), 351–363.

Frieler, K., et al. (2013). Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs. *Musicae Scientiae 17*(3).

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428.

Kolb, B., & Whishaw, I. (1990). *Fundamentals of human neuropsychology* (3rd edition). New York: W. H. Freeman.

Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. *Psychological Review*, *88*(1), 46–66.
Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, *56*(4), 414–423.
Murry, T. (1990). Pitch-matching accuracy in singers and nonsingers. *Journal of Voice*, *4*(4), 317–321.
Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85*(1), 185–193.
Spielberger, C. D., Gorssuch, R. L., Lushene, P. R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press, Inc.
Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society. Series B*, *36*(2), 148–159.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.
Ward, W. D., & Burns, E. M. (1978). Singing without auditory feedback. *Journal of Research in Singing & Applied Vocal Pedagogy*, *1*, 24–44.