sempre:

Society for Education, Music and Psychology Research

Article

Psychology of Music
2019, Vol. 47(4) 504–520
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0305735618765302
journals.sagepub.com/home/pom
SAGE

The time will come: Evidence for an eye-audiation span in silent music reading

Susana Silva and São Luís Castro

Abstract

Musical literacy allows one to "hear music from the page". What can we say about this internal music if we follow the reader's eyes? Do readers hear a given fragment while they are looking at it? Or do they hear it later, when they are already gazing at the following fragment? We hypothesized that the second possibility is more likely, since it allows the reader to start processing one fragment while the previous one is being heard, and thus to keep the musical rhythm going. We refer to this as the eye-audiation span hypothesis, which we tested with an innovative eye-tracking paradigm. We found convergent evidence of an eye-audiation span: first, temporal representations (the internal rhythms) are not concurrent with gaze; second, they emerge later than gaze (gaze-lagged representations). Evidence of lagged temporal representations was stronger in non-experts compared to experts, suggesting either that experts are more efficient in parallel processing, or that their representations are more amodal. Our approach to the relation between gaze and internal rhythm paves the way to mind-reading silent music readers, and provides cues for understanding mechanisms in extra-musical domains, such as implicit prosody in text reading.

Keywords

silent music reading, eye-tracking, audiation, rhythm, musical expertise

Silent music reading requires the ability to audiate (Gordon, 1993), or generate musical representations inside the mind. It differs from music sight-reading, in that the latter engages motor activity and the actual production of sound. Compared to sight-reading, silent music reading remains poorly understood, due to the methodological challenge of accessing the online musical representations of the reader (Rayner & Pollatsek, 1997).[1] In this study, we take a first step

Center for Psychology at University of Porto (CPUP), Faculty of Psychology and Educational Sciences, University of Porto, Porto, Portugal

Corresponding authors:
Susana Silva, Faculty of Psychology and Educational Sciences, University of Porto, Rua Alfredo Allen s/n, Porto, 4200-135, Portugal.
Email: susanamsilva@fpce.up.pt

towards this challenge by examining how the timing of eye movements relates to the timing of internal representations in silent music reading: are musical representations as fast as the eye (are they concurrent to gaze), or do they lag behind the eye?

In sight-reading, the hand that plays the instrument lags behind the eye. This is known as the *eye-hand span* (e.g. Goolsby, 1994; Penttinen, Huovinen, & Ylitalo, 2015; Rosemann, Altenmüller, & Fahle, 2015; Truitt, Clifton, Pollatsek, & Rayner, 1997). Something similar happens when reading text aloud: when a word *n* is being articulated, the eye is often gazing at *n*+1 or *n*+2 (the following words), processing ahead of the motor output. This is named the *eye-voice span* (Buswell, 1921; De Luca, Pontillo, Primativo, Spinelli, & Zoccolotti, 2013; Inhoff, Solomon, Radach, & Seymour, 2011; Laubrock & Kliegl, 2015; Pan, Yan, Laubrock, Shu, & Kliegl, 2013; Silva, Reis, Casaca, Petersson, & Faísca, 2016). The benefit of these gaze-lagged motor outputs (hand or voice) is to afford fluency. The fact that the eye is doing the initial processing of *n*+1 (Silva et al., 2016) while the hand/voice is still playing/articulating *n*, prevents any motor interruption between *n* and *n*+1. The motor output is fluent and there are no bursts.

Should there be an analogue of the eye-hand/voice span in silent music reading? If we consider that the point of the eye-hand/voice span is to preserve the fluency of the motor output, and if there is no motor output in silent music, then this analogue should not be necessary. But if readers need to preserve a different kind of fluency – the fluency of music representations – they will need time to do the visual decoding of *n*+1 before its musical representation is created, and thus they may need an analogue of the eye-hand/voice span. And do silent music readers need fluent internal representations? Silent text readers do not seem to need them to understand the text, but silent music readers probably do. Unlike language, fluency is mandatory in music: one can extract a message from language even if it comes in bursts, but one cannot do the same in music, because music *is rhythm*. The point is that losing control over the rhythm of musical events would mean losing music itself, which is in line with the notion that rhythm decoding is the keystone of music reading skills (Boyle, 1970; Elliott, 1982; Fourie, 2004; Gromko, 2004; McPherson, 1994). Therefore, music readers should yield fluency even when reading in silence. To achieve that, they should be able to do the visual processing of a music pattern before listening to it internally. Their eyes should be ahead of their music representations. They should have an eye-audiation span.

The possibility of an eye-audiation span has not been investigated so far. We did this in the present study, where we used a novel eye-tracking paradigm to determine whether music representations are gaze-lagged, rather than gaze-concurrent. To this end, we tested two complementary predictions: first, the temporal representation of a rhythmic pattern n *is not concurrent with gaze* on *n*; second, the temporal representation *is concurrent with gaze on the following pattern* – the pattern to the right (*n*+1).

We tested our first prediction – that temporal representations are not concurrent with gaze – with a priming manipulation, inspired by the duality between orthographic (visual) and phonological (auditory) representations that is found in language research paradigms (e.g. Frisson, Koole, Hughes, Olson, & Wheeldon, 2014). We generated four-bar rhythmic phrases with three levels of similarity between the second bar (prime) and the third bar (target, see Figure 1): visuo-temporal (condition Same), temporal (condition Homophones), and no similarity (condition Different). The similarity between bars 2 and 3 is expected to facilitate the processing of bar 3, reflecting into shorter viewing times. By comparing the viewing times across these three conditions, we can determine whether there was visual priming, temporal priming, or both. If it is true that temporal representations are not concurrent with gaze, as we hypothesized, we should see *no temporal priming*. In case we see neither temporal nor visual priming, we will not be sure whether the absence of temporal priming was simply due to the lack of sensitivity of

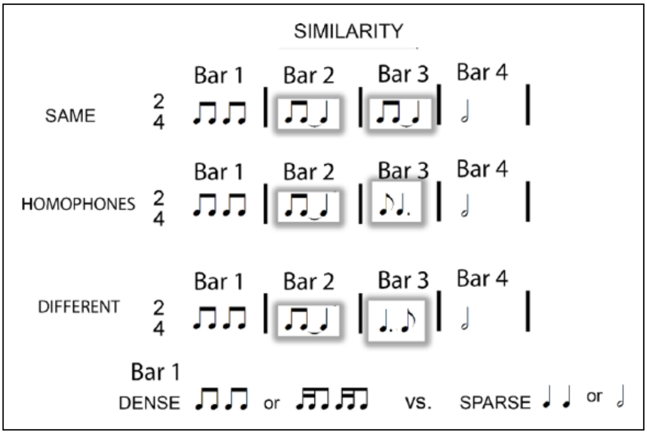


Figure 1. Stimulus structure. One example stimulus is presented, under the three Similarity conditions. In this example, the tied pattern in bar 2 becomes dotted in the Homophones version. The four-bar sequence could have either a sparse or a dense pattern at bar 1 (dense in the example).

our paradigm to capture priming effects. However, if we see visual, but not temporal priming, we will have evidence in support of our hypothesis.

In order to test our second prediction – that the temporal representations of one pattern are *concurrent with gaze on the following one* – we analyzed the effect of a bar's density (the number of notes) on the processing time of the following bar. If temporal representations lag behind gaze, bar *n* should load the processing time of bar *n*+1 with the emergence of rhythmic pattern *n*, and thus high-density patterns should load the following bar more heavily than low-density ones. Conversely, high- and low-density patterns should have the same impact on the following bar if temporal representations are concurrent with gaze.

The general hypothesis that music temporal representations lag behind gaze is based on the assumption that score-reading triggers something like an internal unfolding-rhythm that requires time to exist as an internal representation (the length of the rhythm itself). This view of *timely* temporal representations matches the common-sense idea of "hearing from the page,"[2] the formal concept of audiation proposed by Gordon (1993), and it is supported by experimental findings as well: for instance, some studies showed that the visuospatial patterns in the score are transferred to auditory codes (Hoppe et al., 2014; Simoens & Tervaniemi, 2013) as well as to kinaesthetic ones (Brodsky, Henik, Rubinstein, & Zorman, 2003; Brodsky, Kessler, Rubinstein, Ginsborg, & Henik, 2008; Hubbard, 2013), and both these codes afford timely representations, or internal unfolding-rhythms akin to temporal imagery (Schaeffer, 2014). The literature on temporal imagery (Hubbard, 2010; Jakubowski, Farrugia, Halpern, Sankarpandi, & Stewart, 2015) indicates that temporal representations extracted from scores may even keep a predefined, externally given tempo. Therefore, temporal representations may be so timely that they even show temporal reliability relative to the sensory input. Despite this, the view that temporal representations are timely is not consensual. An alternative view is that readers – mostly expert musicians – may "think from the page" rather than "hear" from it (Drai-Zerbib & Baccino, 2014; Drai-Zerbib, Baccino, & Bigand, 2012). The representations of expert musicians may be amodal and conceptual, rather than perceptual and linked to auditory imagery. If this is true, expert musicians may not show evidence of gaze-lagged temporal representations as strongly as non-experts, and therefore we split our sample into subgroups of experts and non-experts. We compared experts with non-experts for our main hypothesis of gaze-lagged temporal representations, as well as for additional expertise effects (on the total viewing-time of the trial) that might validate the difference between the two groups (see Arthur, Blom, & Khuu, 2016; Penttinen & Huovinen, 2011; Penttinen, Huovinen, & Ylitalo, 2013).

Finally, we chose to use unpitched (rhythm-only) music notation, so as to implement fluency demands at the simplest level: if keeping the rhythm is what presses music readers to be fluent and avoid interruptions, then nothing other than rhythm is required to induce an eye-audiation span. Therefore, among music representations, we focused on temporal music representations, keeping pitch aside.

Method

Participants

Thirty-three participants volunteered to take part in the experiment, but three were excluded due to excessive signal loss in the eye-tracking data. The remaining 29 participants were assigned to one of two groups: experts (n = 16) and non-experts (n = 13). Experts were either musicians, music teachers or music students (eight women; Age, M ± SD = 32 ± 11, Schooling = 18 ± 2). They were currently engaged in music reading at least once a week, and they had been reading it for 21 years on average (SD = 7.8). Non-experts (10 women; Age = 27 ± 10, Schooling = 17 ± 3) were not professionally engaged in music. At the time of the experiment, they read music less often than once a week, and they had done it on average for 10 years (SD = 4).

Stimulus materials

The stimuli consisted of four-bar rhythmic sequences (time signature 2/4), presented under three *Similarity* conditions (Figure 1): visual and temporal, condition Same; temporal only, condition Homophones; no similarity, condition Different. Bars 2 and 3 were the critical ones for testing priming effects. Condition Same repeated the pattern of the second bar in the third bar without any changes in notation style (tied or dotted). Condition Homophones repeated the second bar pattern with a different notation style (tied of second bar became dotted in third; dotted of second became tied in third). Condition Different kept the third pattern of condition Homophones and switched the second bar to the symmetric pattern. These manipulations of Similarity were made to demonstrate the absence of temporal priming (first prediction), and therefore the absence of gaze-concurrent temporal representations.

In order to test for lagged temporal representations (second prediction), we focused on the representation of bar 1 while gazing at bar 2. We manipulated the *Density of bar 1* by creating two types of patterns: a sparse pattern (1–2 events) and a dense pattern (4–6 events). We ruled out the alternative of considering the lagged representation of bar 2 during the viewing of bar 3 because it could involve contamination from priming effects, and that of considering the lagged representation of bar 3 because it could be disturbed by wrap-up effects at the end of the sequence.

We generated the four-bar stimuli starting from four different patterns at bar 2 (Appendix 1).We presented these four patterns at bar 2 under two styles of notation (tied vs. dotted), which gave rise to eight different patterns at bar 2 (4 × 2). These eight patterns appeared twice (4 × 2 × 2 = 16), combined either with a sparse or a dense pattern at bar 1. The last bar of every stimulus was always a half note (Figure 1).

The 48 sequences (16 × 3) were registered in a MIDI sequencer and played with a single drum sound to serve as audio probes in the experimental task (see Procedure, below). The duration of each quarter note (the beat) was 650 ms, and that of each 2/4 bar was 1300 ms. The whole sequence of four bars lasted 5200 ms.

Procedure and eye-tracking recordings

Participants were instructed to "listen internally" to each visually presented rhythmic sequence, and then compare their auditory-imagined rhythm with an audio probe sequence. The question was "Was this what you heard inside your mind?". At the beginning of each trial, participants fixated a cross for 500 ms. At the onset of each visual sequence, they heard two beats (650 ms inter-Onset Interval) that cued them on the tempo used in the audio probes. Each rhythmic sequence was visually presented for 10 s, allowing for the duration of approximately two sequences (5200 ms each) if one follows the given tempo. After the audio probe sequence, participants used the mouse to respond YES or NO. Half the audio probes matched the visual sequence (YES trials), and the other half did not (NO trials). The 24 NO trials were designed to elicit different levels of difficulty in the judgement task (see Appendix 1). Participants were given three practice trials before the experimental session. At the end of the session, they filled in a questionnaire where they were asked about their impression of the stimuli and task, and about any strategies they might have used.

Eye movements were recorded while participants viewed the 48 rhythmic sequences in a 46 × 30 cm monitor. Recordings were made with an SMI RED eye-tracking system (www.smivision.com) at 120 Hz sampling rate. Participants sat 80 cm away from the eye-tracker. At this distance, each bar (9.96 cm) corresponded to 7.12° of the visual field, and so parafoveal preview (< 5°, see Schotter, Angele, & Rayner, 2011) of the following bar was impeded.

Preprocessing

We defined four rectangular areas of interest (AOIs) around each bar in the sequence. Events (saccades and fixations) were obtained for each of the four areas, using a low-speed algorithm to fit the low sampling frequency of the eye-tracker (120 Hz). Trials were inspected visually, and we excluded those with blinks and other tracking losses exceeding 30% of trial time. Using this threshold, we could be certain that all our trials were free from artefacts for at least 70% of viewing time. This would grant signal quality for a time window longer than the critical first-pass viewing period (~50% of trial time). Event statistics were computed in Begaze analysis-software (www.smivision.com), namely the duration of the first fixation on the bar, first-pass viewing time (time spent in each bar before leaving it) and total viewing time (first-pass viewing time + second-pass, or time spent revisiting the bar). In addition, we extracted the amount of regressions into bar 2 after the first transition between bar 2 and bar 3.

Statistical analysis

In order to know whether the behavioural task had been accomplished successfully, we used one-sample *t*-tests to check the response accuracy and the d-prime of non-experts vs. experts against chance levels (50% for accuracy and 0 for d-prime). Since we expected experts to show enhanced performance, we also did direct comparisons between the two groups using independent-sample *t*-tests.

In the eye-tracking data analysis, we started by checking the spatio-temporal trajectory of participants (first-pass scanpath), so as to make sure that the left-to-right order was being followed in all conditions and expertise levels. We then tested the data for priming effects, for lagged temporal representations, and for additional expertise effects that might strengthen possible findings of different priming effects and/or lagged temporal representations for our expert vs. non-expert subgroups. In all analyses, we used linear mixed effects models as implemented in the lme4 package (Bates et al., 2015, lmerTest package used for significance values) for R (R Core Team, 2013). We considered Similarity (Same vs. Homophones vs. Different), Bar (bar 2 vs. others), Expertise (Non-experts vs. Experts), Density (dense vs. sparse bar 1) as fixed factors, and Subjects and Items as random factors.

The analysis of *priming effects* relied on Similarity × Bar interactions. We considered the changes in viewing time from bar 2 (prime) over bar 3 (target), which could consist either of decreases or increases (Bar effects). In order to capture priming effects, we focused on how the similarity between bars 2 and 3 modulated these changes (Similarity × Bar interactions). *Visual priming* was defined as a larger decrease or a smaller increase in viewing time from bar 2 over bar 3 when the visual pattern was repeated (condition Same) than when it was not repeated (Homophones and Different). Correlatively, *temporal priming* was indexed by a larger decrease/smaller increase in viewing time from bar 2 over bar 3 when the temporal pattern was repeated (condition Same and Homophones) than when it was not (condition Different). We focused on two types of measures, indicating two time-scales, possibly related to two processing stages: first-fixation duration and first-pass viewing time. We excluded second-pass reading (return to target areas) from this analysis because it seemed likely that priming effects would require a continuous (non-interrupted) process of confrontation with the visual pattern to emerge.

In addition to Similarity × Bar interactions, priming effects were tapped with the effects of Similarity on the subject-level proportion of regressions into bar 2 after the first transition between bar 2 and bar 3 (again, we excluded second-pass regressions). It is known that regressions follow processing obstacles (Rayner, Chace, Slattery, & Ashby, 2006; Staub, 2010). If visual priming (fewer obstacles after same visual pattern) occurred, a smaller amount of regressions should be expected in condition Same compared to the other two. If temporal priming (fewer obstacles after same temporal pattern) occurred, a smaller amount of regressions should be expected in conditions Same and Homophones compared to condition Different. Similarity and Expertise were used as fixed factors, and Subjects as random factors.

In order to test for *lagged temporal representations* of bar 1 while gazing at bar 2, we examined whether the density of bar 1 (Density) affected the first-fixation duration and the first-pass viewing-time of bar 2 in the two groups (Density × Expertise). If temporal representations were lagged, readers should look longer at bar 2 when it was preceded by dense bar 1 patterns, compared to sparse ones. In order to make sure that the density manipulation was effective, we first analyzed the effects of Density on the first-fixation duration and the first-pass viewing time of bar 1. We assumed that dense patterns would affect both visual processing and temporal processing. If our density manipulation was effective, bar 1 should highlight the effects of density, at least on visual processing.

Additional expertise effects were tapped with Bar × Expertise interactions on the total viewing time. Subjects and Items were used as random factors.

Results

Behavioral results

Accuracy for the whole group was significantly above chance ($M \pm SD = 91.03 \pm 9.02$, $t(28) = 24.49$, $p < .001$), and d-prime values were significantly different from zero (3.70 ± 1.34, $t(28)$

Table 1. Predictors of first-pass viewing time.

Fixed effects	Estimate	SE	*t*	*p*
Similarity				Ns
Bar				
2–3	−107.99	37.61	−2.87	.004
2–4	−137.76	39.63	−3.47	<.001
2–1				Ns
Expertise (NE-E)				Ns
Similarity × *Bar*				
Same-Homophones (2–3)	140.68	55.78	2.52	0.012
Same-Different (2–3)	143.26	54.44	2.63	0.008
Similarity × Bar (2–4/1)				Ns
Similarity × *Expertise*				Ns
Bar × *Expertise*				
2–3 NE-E	103.06	50.72	2.032	.042
2–4/1 NE-E				Ns
Similarity × *Bar* × *Expertise*				Ns
Random effects		*Variance*	*SD*	
Item	Intercept	3294	157	
Subject	Intercept	46520	215	
Residuals		151362	389	

Number of observations: 5043; Items: 48; Subjects: 29. NE: Non-expert; E: expert.

= 14.86, $p < .001$). Experts showed increased accuracy, $t(12.88) = 2.98$, $p = .011$, and discrimination, d-prime: $t(19.27) = 2.68$, $p = .015$, compared to non-experts, but both groups performed above chance (experts: 95.31 ± 2.41, $t(15) = 75.07$, $p < .001$, non-experts: 85.76 ± 11.32, $t(12) = 11.39$, $p < .001$) and showed d-prime values different from zero (experts: 4.26 ± 0.92, $t(15) = 18.48$, $p < .001$, non-experts: 3.01 ± 1.47, $t(12) = 7.33$, $p < .001$). Therefore, experts showed enhanced performance compared to non-experts, but both performed the task successfully. The questionnaires showed no evidence of strategies other than internal listening (no finger tapping or vocalizations). Some participants – mostly experts – reported memorization strategies, which included the focus on bars 2 and 3.

Given the high accuracy rates, the eye-tracking analyses were run on all trials, correct and incorrect. We cross-checked these results with the analysis of correct trials only, and found the exact same pattern.

Priming effects: first-fixation duration vs. first-pass viewing time

For first-fixation duration, there was no evidence of priming effects, since neither Similarity × Bar (2–3) nor Similarity x Bar (2–3) × Expertise interactions were significant ($p > .28$). For first-pass viewing time, Similarity × Bar (2-3) interactions were significant (Table 1 and Figure 2). Increases in processing time from bar 2 over bar 3 were larger for conditions Homophones and Different than for condition Same, indicating *visual priming* effects, which did not depend on expertise (Table1, Similarity × Bar × Expertise). The comparison between Homophones and Different showed non-significant Similarity × Bar (2–3) interactions ($p > .94$), ruling out temporal priming effects. There were no significant Similarity × Bar × Expertise interactions ($p > .60$).

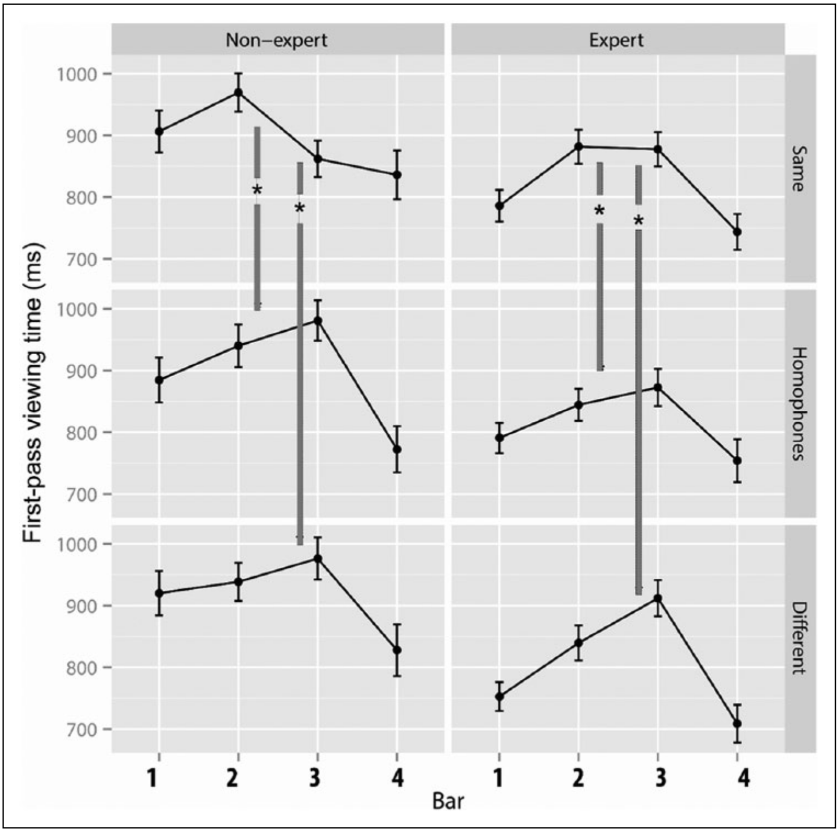


Figure 2. First-pass viewing time on the four areas of interest (AOI) across Similarity and Expertise levels. Changes in viewing time from bar 2 over bar 3 differed across Similarity levels (Similarity × Bar interaction): viewing time decreased/maintained in Same but increased in Homophones and Different. There were no differences between Homophones and Different.

There were no main effects of expertise either on first fixation duration ($p > .72$), or on first-pass viewing time (Table 1), but there was a Bar × Expertise interaction on first-pass viewing time: it increased from bar 2 over bar 3 for experts (averaging similarity conditions), while it decreased for non-experts (Table 1 and Figure 2).

Priming effects: regressions into bar 2

After the first transition from bar 2 to bar 3, participants regressed into bar 2 in ~22% of the trials in condition Same, and in ~30% of the trials in conditions Homophones and Different (Figure 3a). Differences between Same and Homophones were significant (Beta = 0.08, SE = 0.03, t = 2.83, p = .006) and so were those between Same and Different (Beta = 0.09, SE = 0.03, t = 3.19, p = .002), indicating *visual priming*. Homophones did not differ from Different (Beta = 0.01, SE = 0.03, t = 0.37, p = .71), not providing any evidence of temporal priming.

When expertise was added to the model (Figure 3b), the effects of similarity lost significance (Homophones: Beta = 0.07, SE = 0.044, t = 1.55, p = .12; Different: Beta = 0.07, SE = 0.044, t = 1.72, p = .09). There was no interaction between expertise and similarity (Homophones: Beta = 0.03, SE = 0.059, t = 0.47, p = .64; Different: Beta = 0.03, SE = 0.058, t = 0.57, p = .57).

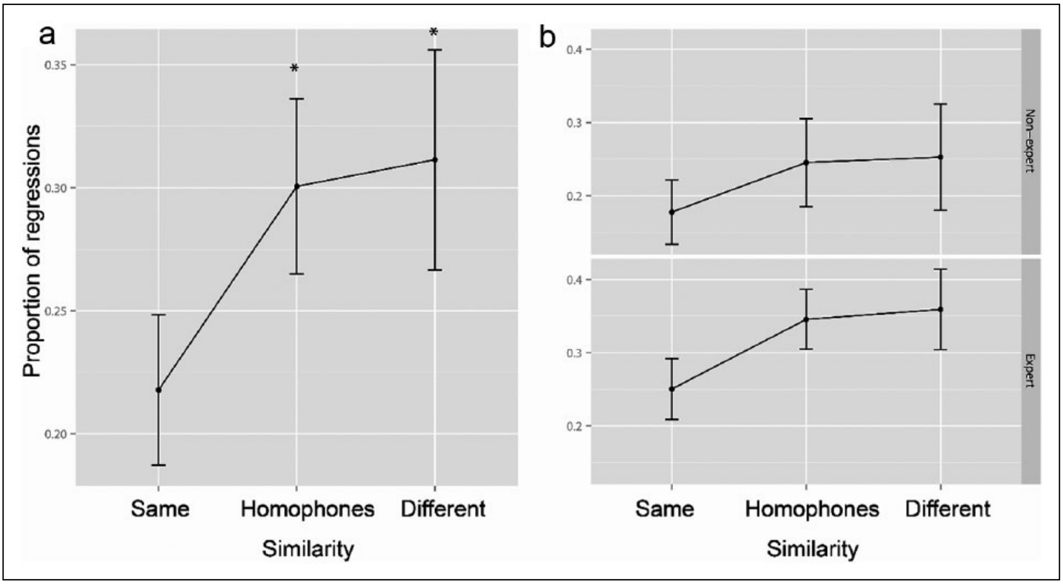


Figure 3. Proportion of regressions into bar 2 after the first transition between bar 2 and bar 3 (a = whole sample; b = non-experts vs. experts). Asterisks indicate significant effects of Similarity (more regressions in Homophones compared to Same, and in Different compared to Same).

Lagged temporal representations: first-fixation duration and first-pass viewing time

As shown in Figure 4a, the density of bar 1 affected the time spent on it: although dense patterns had no effect on first-fixation duration (Density 1: $p > .72$; Density 1 × Expertise: $p > .86$), they increased first-pass viewing time (Beta = 303.23, SE = 36.56, t = 8.29, $p < .001$) for both groups (non-experts: Beta = 302.25, SE = 41.86, t = 7.22, $p <.001$; experts: Beta = 164.54, SE = 30.62, t = 5.37, $p < .001$). Therefore, the density manipulation was effective.

Critical to the hypothesis of lagged temporal representations, the density of bar 1 increased the first-fixation duration and the first-pass viewing time of bar 2 (first fixation: Beta = 43.91, SE = 10.77, t = 4.07, $p < .001$; first-pass: Beta = 110.20, SE = 33.01, t = 3.33, $p = .001$, Figure 4b). In both cases, there were interactions with Expertise (first fixation: Beta = 32.68, SE = 14.13, t = 2.31, $p = .02$; first-pass: Beta = 82.65, SE = 36.56, t = 1.98, $p = .042$), such that the effect was significant for non-experts (first fixation: Beta = 43.79, SE = 12.34, t = 3.54, $p < .001$; first-pass: Beta = 110.10, SE = 32.53, t = 3.38, $p < .001$) but not for experts (first fixation: $p > .21$; first-pass: $p > .32$). Therefore, at least for non-experts, the density of previously viewed materials seems to affect the processing of currently viewed ones, which is consistent with the idea of lagged temporal representations.

Expertise effects on total viewing time

Expertise effects on total viewing-time concerned the time allocated to each bar. Experts differed from non-experts in all comparisons across bars: although the profile of the two groups was qualitatively similar (bar 2 received the longest gaze time, Figure 5), experts showed smaller differences (smaller decline) than non-experts between bar 2 and bar 3 (Beta = 281.01, SE = 104.27, t = 2,68, $p = .007$), but larger differences between bar 2 and bar 1 (Beta = 413.51, SE

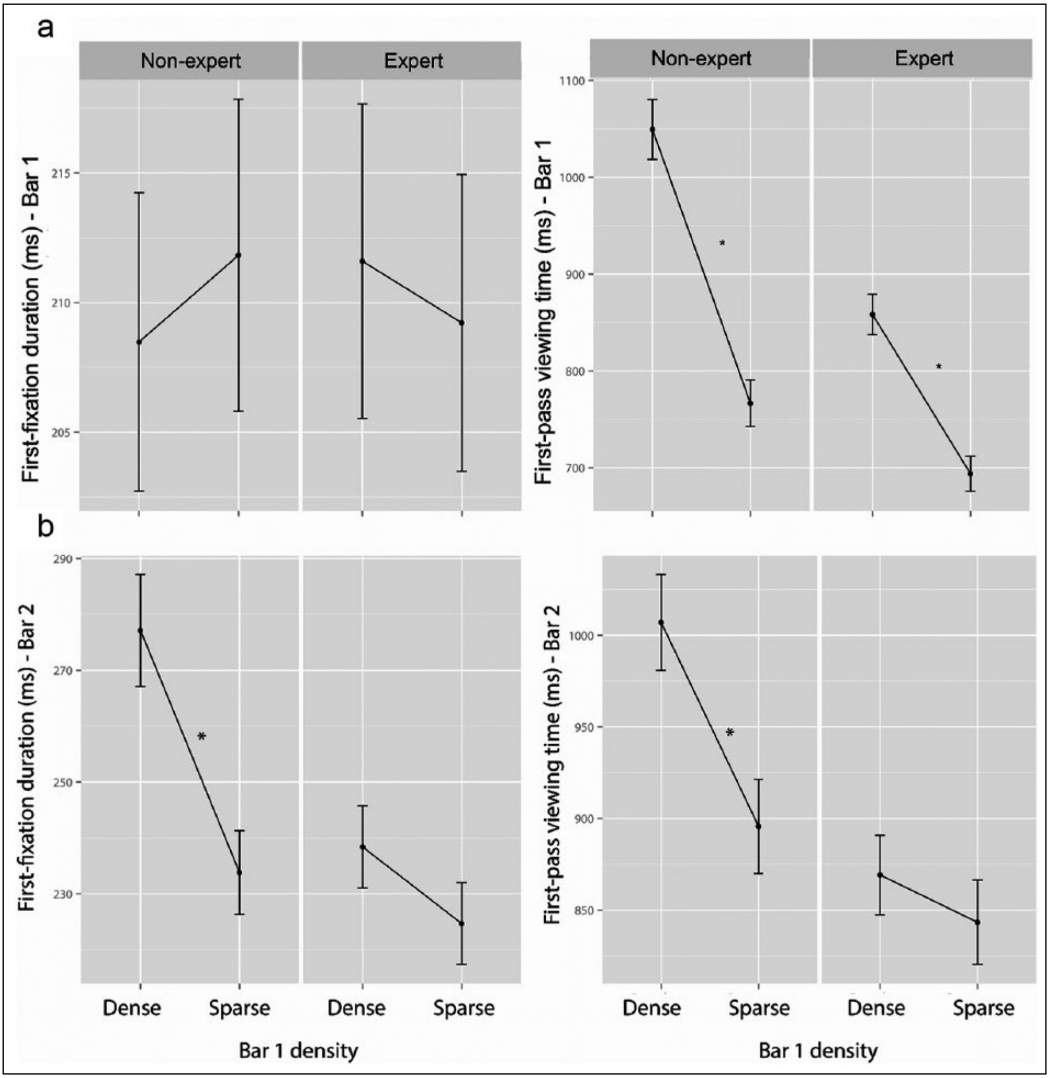


Figure 4. Effects of Bar I density on first-fixation duration and first-pass viewing time of bar I (a, control analysis) and for bar 2 (b, test of lagged temporal representations).

= 105.02, $t = 3.94$, $p = <.001$) as well as between bar 2 and bar 4 (Beta = 283.66, SE = 104.27, $t = 2.72$, $p = .007$). Therefore, experts showed increased focus on bars 2 and 3 (the critical ones) compared to non-experts.

Discussion

We wanted to know whether temporal representations extracted during silent music reading lag behind gaze on the target music symbols (eye-audiation span hypothesis), and whether these gaze-lagged representations are more apparent in non-experts than in expert readers.

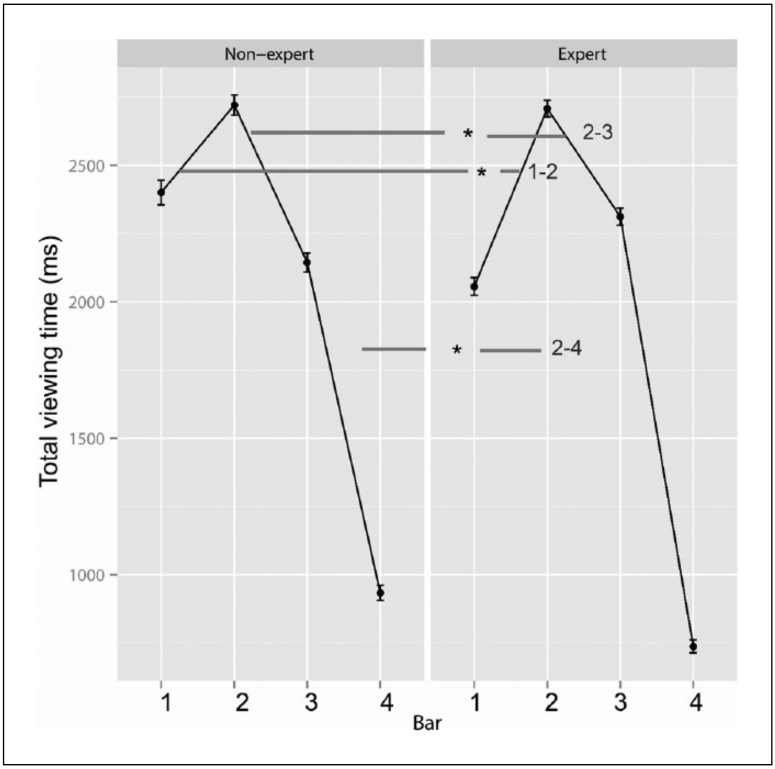


Figure 5. Total viewing time on the four AOIs across Expertise levels. Changes in viewing time from bar 1 over 2, 2 over 3, and 2 over 4 differed across Expertise levels (Similarity × Expertise interaction): experts assigned longer viewing times to bar 3, and shorter viewing times to bars 1 and 4.

To that end, we designed a novel eye-tracking silent-reading paradigm that we used to test non-expert and expert musicians. The two groups differed in behavioral and eye-movement measures, thus validating our sample-split. Our main findings were that there is an eye-audiation span, and the evidence for it is stronger in non-expert music readers.

The finding of gaze-lagged temporal representations was based on two complementary lines of evidence. First, we found that temporal representations of a rhythmic pattern n are *not concurrent with gaze on n*: when testing for temporal vs. visual priming effects on first-pass viewing time, temporal priming of pattern n did not occur, while visual priming did. The presence of visual, but not temporal priming was also testified by the higher amount of regressions from non-primed visual patterns compared to primed ones (Different and Homophones vs. Same), while this was not true for non-primed vs. primed temporal patterns (Different vs. Homophones). Second, we found evidence that the temporal representation of target pattern n is *concurrent with the pattern to the right, n+1*. When bar 1 (*n*) was temporally dense, bar 2 showed increased first-fixation durations and first-pass viewing times. This is consistent with the possibility that participants were loaded with the temporal representations of bar 1 while gazing at bar 2, meaning that the representation of bar 1 was gaze-lagged.

Our study was novel in accessing the online cognitive processes in silent music reading. With our paradigm, we showed that it is possible to tap what happens in the mind during silent music reading using eye movements, and this is the major reason why our findings are important.

From a more general viewpoint, we broke the challenge of accessing internal temporal representations induced by visual input. In our case, these representations were induced by the score, but this scenario is not exclusive to the music domain: implicit prosody – the internal representation of prosodic patterns (e.g., the emotional tone of utterances) as silent text-reading unfolds (e.g., Yao & Scheepers, 2015) – is another example of covert behavior that may benefit from a methodological approach similar to ours.

We took a first step in showing that there is an eye-audiation span, but some issues may deserve further investigation. First, we cannot be certain that our eye-tracking results reflect audiation processes and nothing else, since we engaged readers in a task that adds at least two processing demands. One was to memorize the rhythmic sequence for further comparison with the auditory probe. As a memorization task, it also probably required readers to integrate the sequence as a unified representation (see, e.g., Cara & Gómez, 2016). Dissociating the effects of audiation, integration, and memorization on eye movements should thus be a concern in future uses of our paradigm. Memorization demands during reading may be eliminated by presenting the auditory probe prior to the visual music notation, but then one should be aware of the possibility that readers engage in matching processes (memorized auditory sequence against visual probe) rather than pure audiation. Eliminating integration while keeping audiation seems like a bigger challenge. One possible approach might be asking participants to read silently larger sequences than the ones we presented, so as to make integration less easy. They could be asked to stop at random moments and reproduce the last pattern they heard.

Second, we interpreted the effects of bar 1 Density on the first-fixation duration and the first-pass viewing time of bar 2 as evidence of lagged temporal representations. However, it is also possible to interpret them in a different way, namely if we focus on the fact that they showed up at a processing stage as early as the first-fixation on bar 2: although it is possible that gaze-lagged *temporal representations* emerge right at the onset of *n*+1 (bar 2) processing (our interpretation), it is also possible that increased first-fixation durations on bar 2 for dense bar 1 patterns reflect the maintenance of (dense) *visual representations* in memory. A way of disentangling these two hypotheses in future research could be using an interference paradigm combined with gaze-contingency techniques (e.g., Duchowski, 2002): as readers do their first fixation on *n*+1, they would listen to pattern n (the target of the lagged temporal representation). Evidence of interference would support the emergence of lagged temporal representations during the first fixation on *n*+1.

The second main finding of our study was that the eye-audiation span was more apparent in non-experts than in experts: the pattern of priming effects on first-pass (visual but not temporal) was common to experts and non-experts, but evidence of lagged representations was only seen in non-experts. This confirms our prediction that the amodal temporal representations of expert musicians would attenuate the presence of an internally unfolding rhythm in experts' minds (Drai-Zerbib & Baccino, 2014; Drai-Zerbib et al., 2012) and thus the evidence of lagged temporal representations. This is also consistent with a comment from one of our expert participants, who described his first-pass reading as the extraction of a "rhythmic sketch". However, a different interpretation is also possible. It may be the case that experts do have an internal unfolding-rhythm that generates lagged representations, but they are faster or more efficient in dealing with them: the temporal representations of bar 1 emerged on bar 2, but experts were more efficient than non-experts in processing them in parallel with the initial (gaze-concurrent) processing of bar 2 (see Silva et al., 2016). This would be consistent with the fact that both experts and non-experts showed effects in the same direction (see Figure 4), but only non-experts reached significance. Further studies may determine whether experts develop temporal representations during silent reading that are more amodal than those of non-experts by engaging experts and non-experts in both silent reading and sight-reading (playing the rhythm with an instrument or voice) and then correlating the indices of the eye-audiation span (indices of lagged temporal representations, load on following bar) with those of the eye-hand span (distance between gazed target and executed target in a given moment) of each participant: if it is true that expert readers generate amodal temporal representations, we should expect that the correlation between the eye-hand span and the eye-audiation span is weaker in experts compared to non-experts. Unlike non-experts, experts should not recruit sight-reading-like processes of temporal control during silent reading (no true eye-audiation span) due to their amodal representations. This is why the indices of the two measures should correlate less than in non-experts, for whom a true eye-audiation span would be more strongly expected.

In addition to our main findings – that there is an eye-audiation span, more apparent in non-experts – our study provided some clues on the association between processes in silent reading and the timescales (first-fixation, first-pass viewing time, total viewing time) for eye-movement analysis. Concerning visual priming processes, we saw them in first-pass viewing, but not in first-fixation duration. This suggests that the chunked processing of a bar (necessary to trigger the priming effect) does not occur during first-fixation. Concerning the process that better discriminated between experts and non-experts (the allocation of gaze across critical vs. non-critical bars), we saw it emerging in total viewing-time, but not during first-pass. In light of the answers to the post-experimental questionnaire and the nature of the task (memorization), it seems likely that such process corresponds to memorization strategies, which were more often reported by experts. One hypothesis that emerges is, thus, that first-pass viewing time affords the auditory imagery process, while second-pass viewing time is post-imagetic and devoted to consolidate the retention of the sequence. Dissociating audiation from memorization in future studies may shed more light on this issue.

Conclusion

In our study, we addressed the methodological challenge of inspecting the online temporal representations of silent music readers, using eye-movements to enter the black-box of the internal music extracted from the score. Specifically, we found that the extracted temporal representations (rhythm) lag behind gaze, a mechanism we referred to as the eye-audiation span. Our findings contribute to the possibility of mind-reading music readers, and they may also be relevant to other domains where temporal representations are extracted from visual input, such as implicit prosody in text reading. In addition, we found expertise effects on the eye-audiation span that are in line with the hypotheses that experts' music representations are amodal.

Acknowledgements

We are grateful to Professor Pedro Nobre for logistics support, and to Carolina Dias for her help with data collection.

Ethical statement

Participants signed informed consent according to the Declaration of Helsinki. Participants were healthy adults, and the task they performed in this behavioral experiment did not generate any stress or discomfort. At our university, formal requests for ethical approval are not required under these circumstances.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Portuguese Foundation for Science and Technology under grant UID/PSI/00050/2013.

ORCID iD

Susana Silva https://orcid.org/0000-0003-2240-1828

Notes

1. A search for post-1997 publications on "Silent Music Reading" at Google Scholar provided no more than four useful titles.
2. An example may be found in one of the most memorable scenes in the movie *Amadeus* (Saentz & Forman, 1984), where spectators are confronted with Salieri listening to an accurate version of Mozart's music in his mind while looking at the scores.

References

Arthur, P., Blom, D., & Khuu, S. (2016). Music sight-reading expertise, visually disrupted score and eye movements. *Journal of Eye Movement Research*, *9*(7). https://doi.org/10.16910/jemr.9.7.1

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., ... Singmann, H., ... Green, P. (2015). *Package'lme4'*. Available at: https://cran.rproject.org/web/packages/lme4/index.html

Boyle, J. D. (1970). The effect of prescribed rhythmical movements on the ability to read music at sight. *Journal of Research in Music Education*, *18*(4), 307–318.

Brodsky, W., Henik, A., Rubinstein, B.-S., & Zorman, M. (2003). Auditory imagery from musical notation in expert musicians. *Perception & Psychophysics*, *65*(4), 602–612. https://doi.org/10.3758/BF03194586

Brodsky, W., Kessler, Y., Rubinstein, B.-S., Ginsborg, J., & Henik, A. (2008). The mental representation of music notation: Notational audiation. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 427–445. https://doi.org/10.1037/0096–1523.34.2.427

Buswell, G. T. (1921). The relationship between eye-perception and voice-response in reading. *Journal of Educational Psychology*, *12*(4), 217–227. https://doi.org/10.1037/h0070548

De Luca, M., Pontillo, M., Primativo, S., Spinelli, D., & Zoccolotti, P. (2013). The eye-voice lead during oral reading in developmental dyslexia. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00696

Drai-Zerbib, V., & Baccino, T. (2013). Multisensory integration for expert musicians investigated using eye tracking. *Multisensory Research*, *26*(0), 77–77. https://doi.org/10.1163/22134808–000S0052

Drai-Zerbib, V., & Baccino, T. (2014). The effect of expertise in music reading: cross-modal competence. *Journal of Eye Movement Research*, *6*(5). https://doi.org/10.16910/jemr.6.5.5

Drai-Zerbib, V., Baccino, T., & Bigand, E. (2012). Sight-reading expertise: Cross-modality integration investigated using eye tracking. *Psychology of Music*, *40*(2), 216–235. https://doi.org/10.1177/0305735610394710

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Elliott, C. A. (1982). The relationships among instrumental sight-reading ability and seven selected predictor variables. *Journal of Research in Music Education*, *30*(1), 5–14. https://doi.org/10.2307/3344862

Fourie, E. (2004). The processing of music notation: Some implications for piano sight-reading. *Journal of the Musical Arts in Africa*, *1*(1), 1–23. https://doi.org/10.2989/18121000409486685

Frisson, S., Koole, H., Hughes, L., Olson, A., & Wheeldon, L. (2014). Competition between orthographically and phonologically similar words during sentence reading: Evidence from eye movements. *Journal of Memory and Language*, *73*, 148–173. https://doi.org/10.1016/j.jml.2014.03.004

Goolsby, T. W. (1994). Eye movement in music reading: Effects of reading ability, notational complexity, and encounters. *Music Perception: An Interdisciplinary Journal*, *12*(1), 77–96. https://doi.org/10.2307/40285756

Gordon, E. E. (1993). *Learning sequences in music: Skill, content, and pattern. A music learning theory* (4th ed.). Chicago, IL: GIA.

Gromko, J. E. (2004). Predictors of music sight-reading ability in high school wind players. *Journal of Research in Music Education*, *52*(1), 6–15. https://doi.org/10.2307/3345521

Hoppe, C., Splittstößer, C., Fliessbach, K., Trautner, P., Elger, C. E., & Weber, B. (2014). Silent music reading: Auditory imagery and visuotonal modality transfer in singers and non-singers. *Brain and Cognition*, *91*, 35–44. https://doi.org/10.1016/j.bandc.2014.08.002

Hubbard, T. L. (2010). Auditory imagery: Empirical findings. *Psychological Bulletin*, *136*(2), 302–329. https://doi.org/10.1037/a0018436

Hubbard, T. L. (2013). Auditory imagery contains more than audition. In S. Lacey & R. Lawson (Eds.), *Multisensory imagery* (pp. 221–247). New York: Springer.

Inhoff, A. W., Solomon, M., Radach, R., & Seymour, B. A. (2011). Temporal dynamics of the eye–voice span and eye movement control during oral reading. *Journal of Cognitive Psychology*, *23*(5), 543–558. https://doi.org/10.1080/20445911.2011.546782

Jakubowski, K., Farrugia, N., Halpern, A. R., Sankarpandi, S. K., & Stewart, L. (2015). The speed of our mental soundtracks: Tracking the tempo of involuntary musical imagery in everyday life. *Memory & Cognition*, *43*(8), 1229–1242. https://doi.org/10.3758/s13421–015–0531–5

Laubrock, J., & Kliegl, R. (2015). The eye-voice span during reading aloud. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01432

McPherson, G. E. (1994). Factors and abilities influencing sight-reading skill in music. *Journal of Research in Music Education*, *42*(3), 217–231. https://doi.org/10.2307/3345701

Pan, J., Yan, M., Laubrock, J., Shu, H., & Kliegl, R. (2013). Eye–voice span during rapid automatized naming of digits and dice in Chinese normal and dyslexic children. *Developmental Science*, *16*(6), 967–979. https://doi.org/10.1111/desc.12075

Penttinen, M., & Huovinen, E. (2011). The early development of sight-reading skills in adulthood a study of eye movements. *Journal of Research in Music Education*, *59*(2), 196–220. https://doi.org/10.1177/0022429411405339

Penttinen, M., Huovinen, E., & Ylitalo, A.-K. (2013). Silent music reading: Amateur musicians' visual processing and descriptive skill. *Musicae Scientiae*, *17*(2), 198–216. https://doi.org/10.1177/1029864912474288

Penttinen, M., Huovinen, E., & Ylitalo, A.-K. (2015). Reading ahead: Adult music students' eye movements in temporally controlled performances of a children's song. *International Journal of Music Education*, *33*(1), 36–50. https://doi.org/10.1177/0255761413515813

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., & Pollatsek, A. (1997). Eye movements, the eye-hand span, and the perceptual span during sight-reading of music. *Current Directions in Psychological Science*, *6*(2), 49–53. https://doi.org/10.1111/1467–8721.ep11512647

Rosemann, S., Altenmüller, E., & Fahle, M. (2015). The art of sight-reading: Influence of practice, playing tempo, complexity and cognitive skills on the eye–hand span in pianists. *Psychology of Music*, 305735615585398. https://doi.org/10.1177/0305735615585398

Saentz, S. (Producer), & Forman, M. (Director). (1984). Amadeus [Motion Picture]. United States: Orion Pictures.

Schotter, E. R., Angele, B., & Rayner, K. (2011). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5–35. https://doi.org/10.3758/s13414–011–0219–2

Silva, S., Reis, A., Casaca, L., Petersson, K. M., & Faísca, L. (2016). When the eyes no longer lead: Familiarity and length effects on eye-voice span. *Frontiers in Psychology*, 1720. https://doi.org/10.3389/fpsyg.2016.01720

Simoens, V. L., & Tervaniemi, M. (2013). Auditory short-term memory activation during score reading. *PLoS ONE*, *8*(1), e53691. https://doi.org/10.1371/journal.pone.0053691

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002

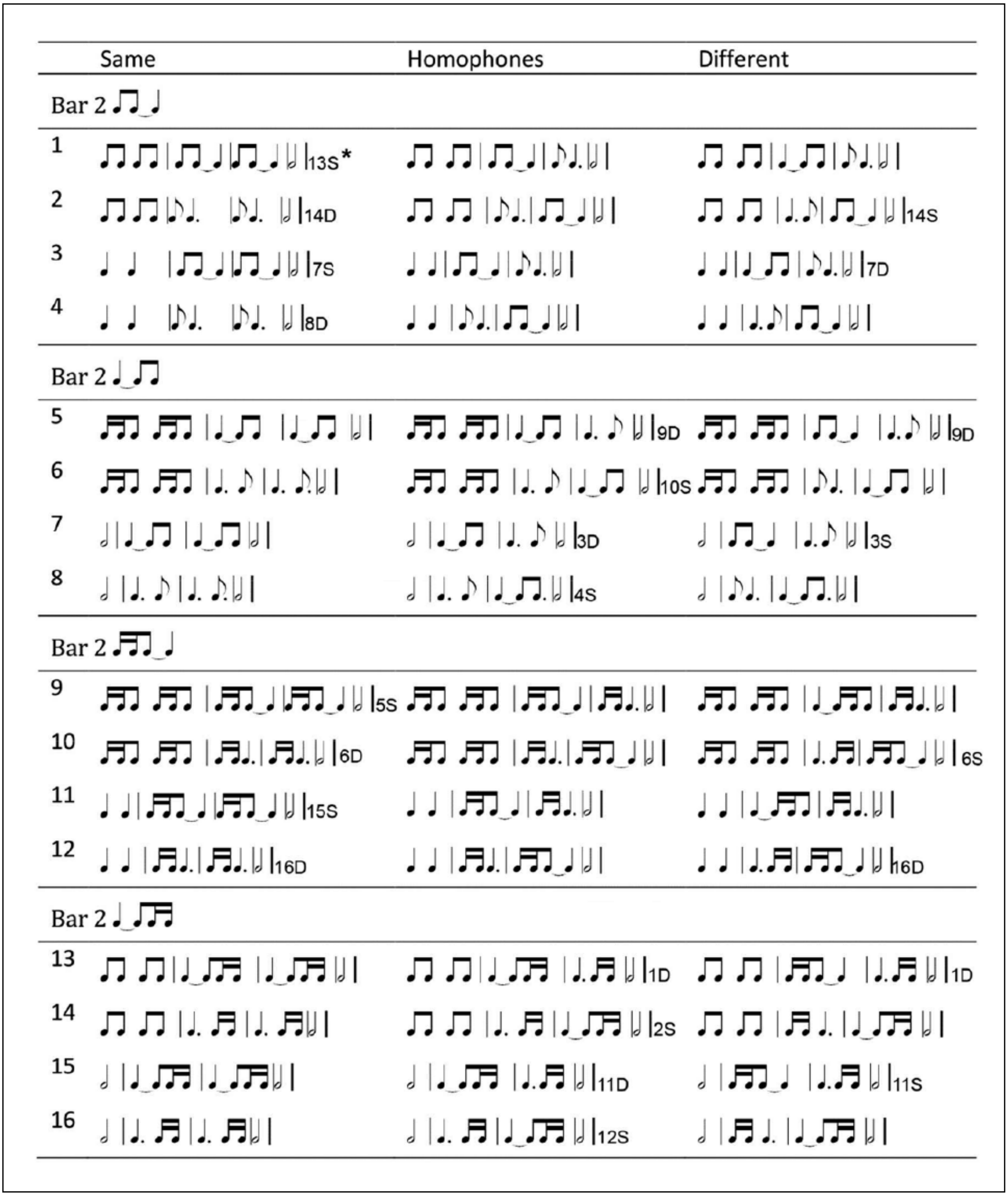
Truitt, F. E., Clifton, C., Pollatsek, A., & Rayner, K. (1997). The perceptual span and the eye-hand span in sight reading music. *Visual Cognition*, *4*(2), 143–161. https://doi.org/10.1080/713756756

Yao, B., & Scheepers, C. (2015). Inner voice experiences during processing of direct and indirect speech. In L. Frazier & E. Gibson (Eds.), *Explicit and implicit prosody in sentence processing* (pp. 287–307). Cham, Switzerland: Springer International.

Appendix I

The 16 stimuli (presented under three Similarity versions) are organized here according to the pattern used at bar 2. There were four basic patterns at bar 2. Each pattern appeared in both tied and dotted versions (2 × 4). Each of these versions was combined either with a dense or a sparse motif at bar 1 (2 × 2 × 4 = 16).

Half the stimuli were YES trials (audio probe matches visual sequence), and the other half were NO trials (audio probe does not match visual sequence). In half the NO trials, the first bar of the audio probe was common to the visualized sequence (difficult judgement) while in the other half it was not (easier judgement); also in half these trials, the repetition structure of bars 2 and 3 (repeat vs. non-repeat) was common to the audio probe and the audio sequence (also a difficult judgement) while in the other half it was not.



Stimulus material.

Audio probe used at each NO trial (Was this what you heard inside your mind? – No); S = same, H = homophone, D = Different. For instance, for visual stimulus 1/Same, we used the audio of stimulus 13/Same (13S) as probe; for 2/Same we used 14/Different (14D). Stimuli from YES trials have no indications.