Summer 2024

Is Chunking Multimodal?: Music Reading Expertise Effects on Eye Movements During a Cross-modal Visual Search Paradigm

Nicole Arco
University at Albany, State University of New York, narco@albany.edu

Is chunking multimodal?: Music reading expertise effects on eye movements during a cross-modal visual search paradigm

by

Nicole M. Arco

A Thesis

Submitted to the University at Albany, State University of New York

In Partial Fulfillment of

the Requirements for the Degree of

Master of Arts

College of Arts & Sciences

2024

ABSTRACT

Experts show performance advantages during visual search due to their extensive experience with domain-specific stimuli. Experts form memory representations for meaningful visual patterns, called chunks, that group together multiple chess features. Prior work suggests that the ability for experts to precisely encode a search template facilitates visual search performance (e.g. Hout & Goldinger, 2015). In music, expert musicians might also form chunks (see e.g., Maturi & Sheridan, 2020), although it is unclear what constitutes a chunk in music. The current study addressed the possibility that chunks are multimodal by introducing a new auditory-visual cross-modal version of the visual search paradigm introduced by Maturi and Sheridan (2020), while monitoring eye movements, comparing experts and non-musicians. Results support the idea that chunks are possibly multimodal in music: compared to non-musicians, experts had higher accuracy which was magnified in the cross-modal condition, indicating experts' performance advantages. In addition, compared to non-musicians, experts had longer dwell times in the target region for the cross-modal condition only, suggesting that experts fixate in more relevant regions and are successfully integrating auditory information into visual information, and that process takes time.

ACKNOWLEDGEMENTS – I would like to thank my advisor Dr. Heather Sheridan for all of her help on this project, I could not have completed it without her.

TABLE OF CONTENTS

I. Abstract……ii

II. Acknowledgments……....iii

III. List of Figures……...vi

IV. List of Tables……vii

V. Introduction…….....1

a. Expertise and chunking theory……1

b. Visual search……2

c. Current study……3

VI. Method……...4

a. Participants……...4

b. Materials……..5

c. Apparatus…….....5

d. Procedure…….....6

VII. Results……7

a. OSPAN…….....8

b. Accuracy……..9

c. Reaction Time Measures……..9

d. Dwell-based Analyses……10

VIII. Discussion……17

a. Interpretation of results……..18

b. Music reading literature……20

c. Strengths and limitations……20

d. Future directions…….21

IX. References……23

LIST OF FIGURES

I. Figure 1……..7

II. Figure 2……12

III. Figure 3……12

IV. Figure 4……13

V. Figure 5……14

VI. Figure 6……14

VII. Figure 7……15

VIII. Figure 8……16

IX. Figure 9……17

X. Figure 10……..17

LIST OF TABLES

I. Table 1……..11

II. Table 2……..12

III. Table 3……..13

IV. Table 4……..14

V. Table 5……..15

VI. Table 6……..16

Introduction

Expertise facilitates visual search performance in many domains, such as chess, medicine and sports (for review, see Brams et al., 2019). One mechanism that could be contributed to expertise advantages during visual search is that experts are adept at perceptual encoding of domain-specific information, which allows them to encode precise representations of the target they are searching for (i.e., the *search template* or *attentional template*), which facilitates attentional guidance during visual search tasks (for a related discussion, see Maturi and Sheridan, 2020). To further explore the mechanisms that support the encoding of search templates, we introduced a multimodal visual search paradigm in the domain of music to examine the impact of expertise on eye movements during visual search. Given that music expertise is multimodal, this paradigm allowed us to explore the contributions of the multisensory knowledge representations of experts to the encoding of precise visual search templates. We will next discuss prior work investigating perceptual processing in experts and then describe the current cross-modal visual search paradigm.

Chunking and template theories (Chase & Simon, 1973; Gobet, 1996) of expertise assume that experts learn to group visual features into meaning patterns (i.e., chunks). Chunking theory provides a possible explanation for why experts show domain-specific performance advantages in many different domains, such as chess and radiology (e.g., Chase & Simon, 1973; Gobet & Simon, 1996; 1998; Reingold et al., 2001; for review, see Reingold & Sheridan, 2011). In the domain of chess, evidence of chunking can be seen by expert players learning how to encode patterns that group together multiple different chess pieces, as opposed to individual chess pieces (e.g., Reingold & Sheridan, 2011). In chess, chunking and template theory

emphasizes visual processing, and current conceptualizations of chunking in the literature don't allow for the possibility that chunks may integrate information from multiple modalities.

Although there is extensive evidence that chunking may facilitate visual expertise in a variety of tasks (for a review, see Bilalić, 2017; Gobet, 2016; Sala & Gobet, 2017), more work is needed to understand the specific mechanisms by which chunking may support visual search performance. One possible mechanism by which expertise could facilitate visual search is that chunking could enable experts to form more precise visual search templates in their working memory. More precise visual search templates could lead to a more efficient search, given that prior work has shown that manipulations of the precision of the search template can impact visual search efficiency. Specifically, visual search and attentional guidance become less efficient when the search template is less similar to the target or additional features are included (e.g. Hout & Goldinger, 2015), for example, if the search template is presented in a different orientation or a different color than the target. A different line of work compared a word cue presented as the search template instead of a specific picture, with the word cue resulting in a less efficient search compared to the picture (Malcolm & Henderson, 2009; Schmidt & Zelinsky, 2009). Building on this work showing that precise search templates facilitate attentional guidance during visual search, we aimed to further explore the impact of expertise on visual search. Experts have detailed memory representations for domain specific stimuli, which could allow them to encode more precise visual search templates than novices.

Previously, Maturi and Sheridan (2020) tested the hypothesis that chunking by experts can facilitate the encoding of a visual search template. Specifically, experts and non-musicians searched for a specific visual bar of music (i.e., the search template) within a larger music array that was presented at the same time as the search template. Eye movements revealed that the

experts had a qualitatively different search pattern from the novices that was characterized by fewer, but longer dwells on the search template compared to non-musicians. Maturi and Sheridan (2020) interpreted their results as evidence that experts were using chunks to precisely encode the search template in visual working memory, which meant that experts needed to return to the search template less frequently, compared to non-musicians.

Extending Maturi & Sheridan (2020), we introduced a cross-modal manipulation, to test if chunks could be multimodal and if music experts use multimodal chunks to support the encoding of complex search templates from their domain of expertise. There are many advantages for using a music-related search paradigm to study the multisensory representations of experts. Music is more visually complex and music may also contain chunks, such as features or individual notes that are grouped together into arpeggios and chords (Maturi & Sheridan, 2020). Music also has a multisensory component to a greater extent than other domains, such as chess, and prior works shows that music experts acquire multisensory knowledge (Drai-Zerbib & Baccino, 2011; 2018; Hoppe et al., 2014; Lee & Lei, 2016; Schön & Besson, 2005; Schürmann et al., 2001; Simoens et al., 2012; Wong & Gauthier, 2009).

Therefore, to further explore how experts use multisensory knowledge to encode search templates, the current study's visual search task included a modality manipulation that contrasted two conditions: 1) auditory encoding of the search template, followed by visually searching for the target (i.e., cross-modal presentation; See Figure 1, Panel A), and 2) visual encoding of the search template, followed by visually searching for the target (unimodal presentation; See Figure 1, Panel B). This novel auditory-visual cross-modal visual search paradigm will test if the assumptions that chunking is multimodal and that multimodal integration is a possible mechanism by which chunks support precise search templates and visual search performance by

experts. We hypothesized that music experts acquire multimodal chunks that allow them to precisely encode the search template, even in the cross-modal condition. To the extent that these assumptions are correct, we predicted that expertise differences will be greater for cross-modal processing compared to unimodal processing, because the experts would be able to use multisensory chunks to derive visual representations of the search template from auditory information. Also, we predicted that the experts would look at the search target within the search array faster than the non-musicians, even if the search target was encoded in a different modality (i.e. cross-modal presentation). Finally, replicating Maturi & Sheridan (2020), we predicted that experts would spend more time looking at relevant regions (i.e. the target), as seen by longer dwell-times on the target compared to the distractor region.

Methods

Participants

Sixty-six [1]participants were recruited (33 expert musicians, mean age = 22.55; 33 non-musicians, mean age = 18.64; 47 female). The expert musicians had at least 10 years of musical training and could read music (Maturi & Sheridan, 2020; Sheridan & Kleinsmith, 2021). The non-musicians were participants with less than two years of musical training and the non-musicians self-reported that they could not read music (Maturi & Sheridan, 2020; Sheridan & Kleinsmith, 2021). The participants were either compensated in course credit or $15 cash.

[1] We followed the recommendation of Brysbaert and Stevens (2018) that at least 1600 observations in each condition are required to provide adequate statistical power. We met this recommendation because our sample size of 33 participants (and 51 items per condition) provided a potential maximum of over 1683 observations per condition. We included a larger sample size than is typical in the expertise literature, because it was difficult to predict the effect size that we would observe given the novelty of our cross-modal visual search conditions.

Materials

The visual search materials consisted of 104 two-line excerpts of lesser-known musical scores from the Baroque and Classical music repertoire. Participants were instructed that they would either see or hear a target bar of music (i.e., the search template) and immediately following, be presented with a musical excerpt (i.e., the search array) in which they must visually search for what they previously saw or heard. This experiment used a 2 x 2 design (group: expert or non-musician x modality presentation: cross-modal presentation (i.e. auditory-visual) or unimodal presentation (i.e. visual-visual). We randomly selected two bars of music from the search array, one from the first line of music and one from the second line of music, to serve as the target bar and the distractor bar. We counterbalanced across participants which bar of music would be designated the target bar, and the second bar would serve as the distractor. Across participants, the same bars served as both targets and distractors, so that each bar served as a control for itself.

Apparatus

Eye movements were monitored using the SR Research EyeLink Portable Duo system with high spatial resolution and a sampling rate of 1000 Hz. The head was stabilized using a head and chin rest. The gaze-position calibration error was less than 0.5^{o} for all participants. Auditory stimuli were created using Finale Music Notation Software, Version 25 (2016) and were played through SONY MDR-7506 Sound Monitor Headphones, Casque Hi-Fi Digital. Visual stimuli were displayed on a 24-inch Asus CG248QE computer screen with resolution of 1920 x 1080 pixels. The screen was 56.5 cm away from the participant. The degree of visual angle, in pixels, for the search arrays were on average 73.77° in width and 26.97° in height, ranging from 73.45°-76.97° in width and 27.07°-27.93° in height. The degree of visual angle for

the target and distractor bars were on average 18.60° in width and 14.13° ranging from 18.60° - 20.17° in width and 12.72°-13.73° in height. The musical excerpts were presented in black on an off-white background.

Procedure

Testing was conducted in one session, and testing time was approximately one hour. Participants first completed an Operation Span Task (OSPAN; Turner & Engle, 1989) to test working memory followed immediately by a music-related visual search task. The OSPAN task presented participants with 66 math-word combinations (e.g. (3x4) + 2 = 14 ? String), with set sizes varying from two to six. After viewing two to six items, participants were prompted to recall the words they saw, in the order they saw them, by typing those words in a box.

Immediately following the OSPAN task, participants completed the eye tracking component. During each trial, participants were first presented with a bar of music (i.e., the search template) that was presented either auditorily or visually, followed immediately by a visual presentation of a two-line excerpt of music (i.e., the search array; see Figure 1). Participants were instructed to locate and fixate on the target bar of music (i.e., the search template that they saw or heard earlier in the trial), as quickly and accurately as possible, and to press a button on a button box while fixating on the target measure. The participant's fixation location at the time of the button press was recorded to allow us to assess if they accurately located the target. The trial ended after the button press.

There were two practice trials, one for the cross-modal condition and one for the unimodal condition, which were followed by 102 experimental trials (51 trials were in the cross-modal condition and 51 in the unimodal condition). There was a short break in the middle of the

experiment to allow participants to remove their head from the head and chin rest and to rest their eyes. Following the break, the eye tracker was re-calibrated. No single participant saw or heard the same search template or saw the same search array more than once.

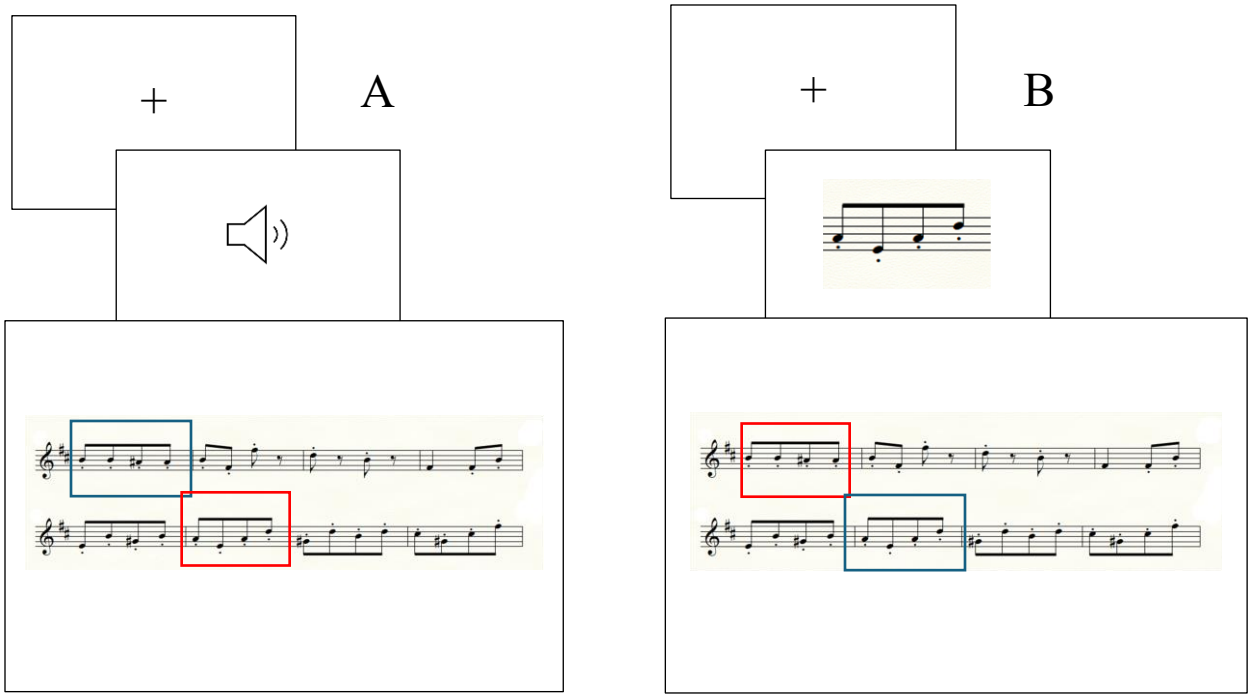

Figure 1. An illustration of the procedure for the cross-modal presentation and unimodal presentation trials. As shown in Panel A), in the cross-modal condition, the participants were presented with a fixation cross while listening to the target bar of music, and immediately afterwards they were be presented with the search array (see target bar outlined in blue, distractor bar outlined in red, which was not be shown in the study). As shown in Panel B), in the unimodal condition the participants were presented with an image of one bar of music, and immediately afterwards they were presented with the search array (again, see the target bar outlined in blue and distractor bar outlined in red, which was not be shown in the study).

Results

To examine group differences in working memory, we conducted a two samples *t*-test for the OSPAN scores, comparing experts and non-musicians. To examine the effects of expertise and cross-modal integration we looked at the following measures: 1) accuracy – proportion of correct trials, 2) reaction time (RT) – the time from the beginning of the trial until a response was

made, 3) time to first fixation – the time interval between the start of the trial and the start of the first fixation in the target region, 4) total dwell time – the sum of all fixations across the entire trial in a given IA (interest area), and 5) first-run dwell time – the sum of all fixations in a given IA during the first dwell in this IA. Analyses were conducted on all trials[2]. Data was analyzed using the *lme4* package (Bates et al., 2015) in RStudio. Continuous variables were analyzed using linear mixed models (LMMs) using the lmer function, and accuracy and time spent in the target region, binary variables, were analyzed using the glmer function. We began by analyzing the full model[3], if the models failed to converge, they were systematically trimmed until they converged (see Barr et al., 2013). For accuracy, reaction time, and time to first fixation we reported the main effects for group (expert, non-musician) and modality presentation (unimodal, cross-modal), as well as any interactions between musical experience and modality presentation. For the dwell-based analyses we also reported main effect for IA region (i.e. relevancy), as well as any interactions between musical experience, modality presentation, and relevancy. A post-hoc analysis, using the *emmeans* package (Lenth, 2023) was also conducted to specifically examine the differences between groups and conditions.

OSPAN

Experts scored significantly higher on the OSPAN task ($M = 38.36$) compared to non-musicians ($M = 33.06$), $t(64) = 1.99$, $p < .05$. Because of this significant group difference, the raw OSPAN scores were converted to standardized *z*-scores and included as a co-variate in all

[2] Data was analyzed using all trials and accurate trials only. The pattern of findings did not change between all trials and accurate trials, only levels of significance changed for some measures. Therefore, we only are reporting results of all trials.

the models. The results of the analyses did not change when the *z*-scores were added to the model.

Accuracy

Supporting our predictions that experts are encoding precise representations of the search template, even if it was encoded auditorily, experts were more accurate compared non-musicians and there was higher accuracy for unimodal presentation compared to cross-modal presentation (see Table 1 and Figure 2). Experts had higher accuracy for both unimodal and cross-modal presentations, compared to non-musicians, but this difference was magnified in cross-modal presentation.

Reaction Time Measures

There was no main effect of group for RT. There was a main effect of modality presentation such that there were faster RTs for unimodal presentation compared to cross-modal presentation (see Table 2 and Figure 3). There was a two-way interaction between expertise and modality presentation such that there was greater effect of group (experts had faster reaction times compared to non-musicians) for unimodal presentation compared to cross-modal presentation.

There was no main effect of group for time to first fixation. There was a main effect of modality presentation such that there were faster times to first fixations for the unimodal presentation compared to the cross-modal presentation (see Table 3 and Figure 4). There was a two-way interaction between expertise and modality presentation such that experts had faster times to first fixation for both the cross-modal and unimodal presentation, but this was only significant in the cross-modal condition.

Dwell-based Analyses

Both total dwell time and first-run dwell time had very similar pattern of results, so they will be discussed in parallel. There was no main effect of group. However, there was a main effect of modality presentation such that there were shorter dwells times for unimodal presentation compared to cross-modal presentation (see Table 4; Figures 5 and 6). There was also a main effect of relevancy, such that there were longer dwell times in the target region compared to the distractor region. There was a significant three-way interaction between expertise, modality presentation, and relevancy such that in the cross-modal condition, experts had longer total dwell times and first-run dwell times in the target region compared to non-musicians.

To further explore the three-way interactions for total dwell time and first run dwell time, we conducted follow-up LMMs in which we analyzed group by relevancy for the cross-modal presentation and unimodal presentation separately (see Tables 5 and 6; Figures 7, 8, 9 and 10). For both of these follow-up LMMs, there was no main effect of group. However, there was a main effect of relevancy such that there were longer dwells in the target region compared to the distractor region. For the cross-modal condition only, there was a significant two-way interaction between expertise and relevancy such that experts had longer total dwells and first run dwells in the target region, compared to non-musicians. In the unimodal presentation, we saw a different pattern of results. Specifically, for total dwell time, the experts spent numerically less time looking in both the target and distractor regions, compared to non-musicians. For first-run dwell time again we see a different pattern of results, such that experts spent numerically less time looking in the distractor region compared to non-musicians, whereas both experts and non-musicians spent the same amount of time looking in the target region.

In addition, there was a two-way interaction between expertise and modality presentation such that experts had longer dwells than non-musicians for the cross-modal presentation, whereas non-musicians had longer dwells than experts for the unimodal presentation. There was also a two-way interaction between expertise and relevancy such that experts, compared to non-musicians, had longer dwells in the target region compared to the distractor region. The follow-up LMMs were consistent with this two-way interaction, such that, in both the cross-modal and unimodal presentations, the experts spent less time looking at the distractor region, compared to non-musicians, suggesting that experts are more sensitive to relevant information. Lastly, there was a two-way interaction between modality presentation and relevancy such that there longer dwells in the target regions compared to the distractor regions for both cross-modal and unimodal presentations, but this difference was greater in the unimodal presentation.

These results suggest that not only are experts able to fixate on more relevant regions, but since there are only longer dwell times for cross-modal presentation only, it is possible that it takes time to integrate the auditory information into visual, which results in longer dwells times in the target region for cross-modal presentation only.

	Accuracy (proportion correct)			
	b	*SE*	*z*	*p*
Group	**-1.82**	**0.24**	**-7.23**	**< .001**
Modality presentation	**1.38**	**0.07**	**20.94**	**< .001**
Modality presentation x Group	**0.62**	**0.13**	**4.71**	**< .001**

Table 1. Results for the LMM for accuracy. Significant trials are shown in bold.

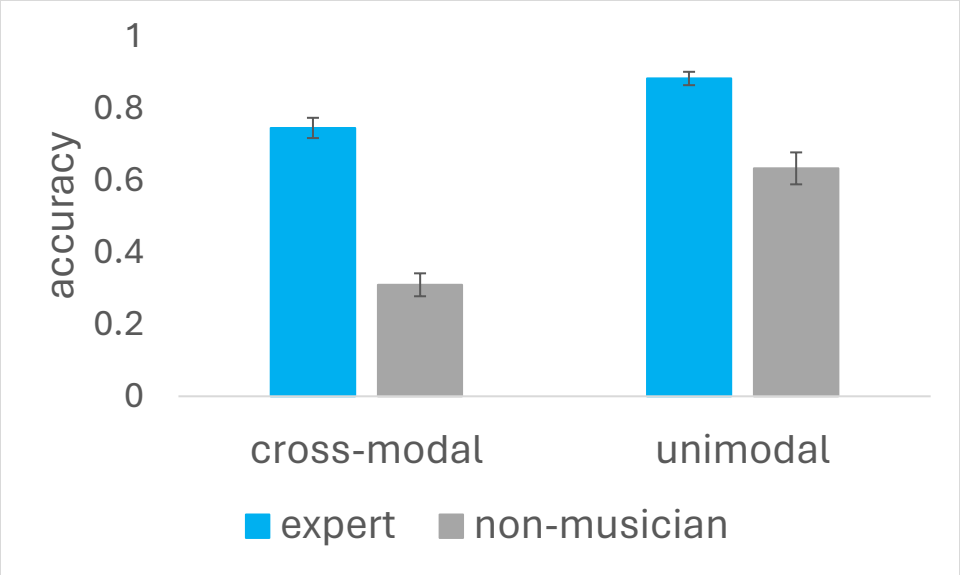

Figure 2. Accuracy (proportion correct)

	Reaction Time (ms)			
	b	*SE*	*t*	*p*
Group	627.86	678.82	0.93	0.36
Modality presentation	**-3996.67**	**99.14**	**-40.32**	**< .001**
Modality presentation x Group	**431.52**	**198.32**	**2.18**	**0.03**

Table 2. Results for the LMM for reaction time (ms). Significant trials are shown in bold.

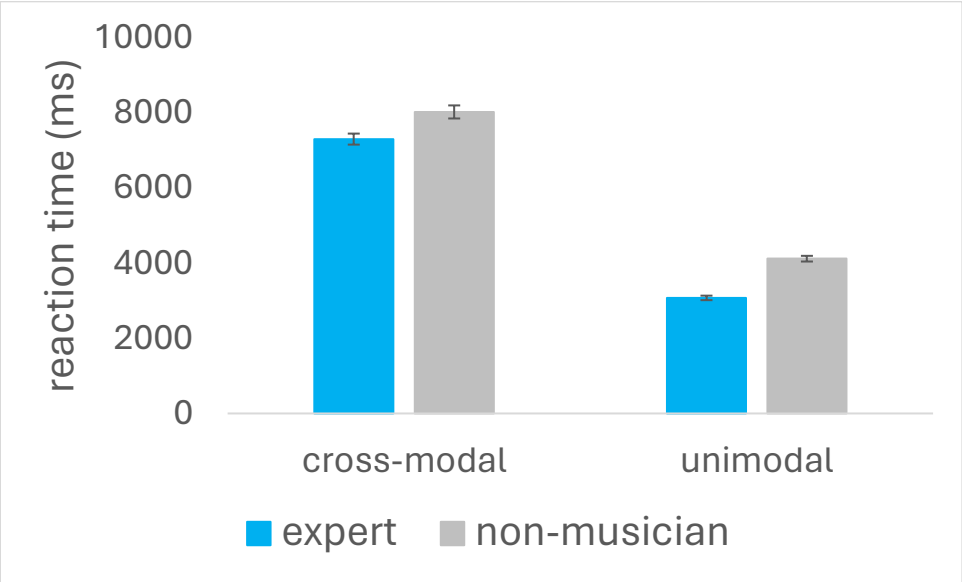

Figure 3. Reaction Time (ms)

	Time to First Fixation (ms) - all trials			
	b	*SE*	*t*	*p*
Group	205.04	107.08	1.92	0.06
Modality presentation	**-1024.42**	**41.63**	**-24.61**	**< .001**
Group x Modality presentation	**-219.36**	**83.27**	**-2.63**	**< .01**

Table 3. Results for the LMMs for first fixation to trial end (ms). Significant results are shown in bold.

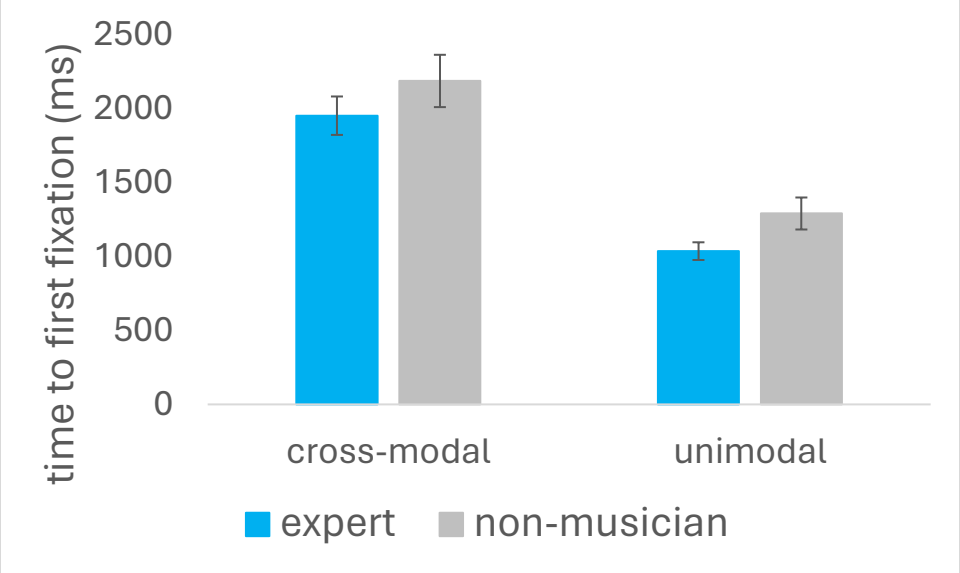

Figure 4. Time to First Fixation (ms)

	Total Dwell Time (ms) - all trials			
	b	*SE*	*t*	*p*
Group	13.60	153.17	0.09	0.93
Modality presentation	**-565.17**	**25.10**	**-22.51**	**< .001**
Relevancy	**1153.10**	**25.08**	**45.97**	**< .001**
Group x Modality presentation	**327.87**	**50.22**	**6.53**	**< .001**
Group x Relevancy	**-346.46**	**50.10**	**-6.92**	**< .001**
Modality presentation x Relevancy	**-195.76**	**20.39**	**-3.89**	**< .001**
Group x Modality presentation x Relevancy	**845.28**	**100.79**	**8.39**	**< .001**
	First Run Dwell Time (ms) - all trials			
	b	*SE*	*t*	*p*
Group	-3.99	48.78	-0.08	0.94
Modality presentation	**-66.14**	**13.27**	**-4.99**	**< .001**
Relevancy	**478.27**	**13.26**	**36.07**	**< .001**
Group x Modality presentation	**128.83**	**26.54**	**4.85**	**< .001**

Group x Relevancy	**184.20**	**24.48**	**-6.96**	**< .001**
Modality presentation x Relevancy	**135.19**	**26.63**	**5.08**	**< .001**
Group x Modality presentation x Relevancy	**251.39**	**53.26**	**4.72**	**< .001**

Table 4. Results for the LMMs for total dwell time and first-run dwell time (ms). Significant results are shown in bold.

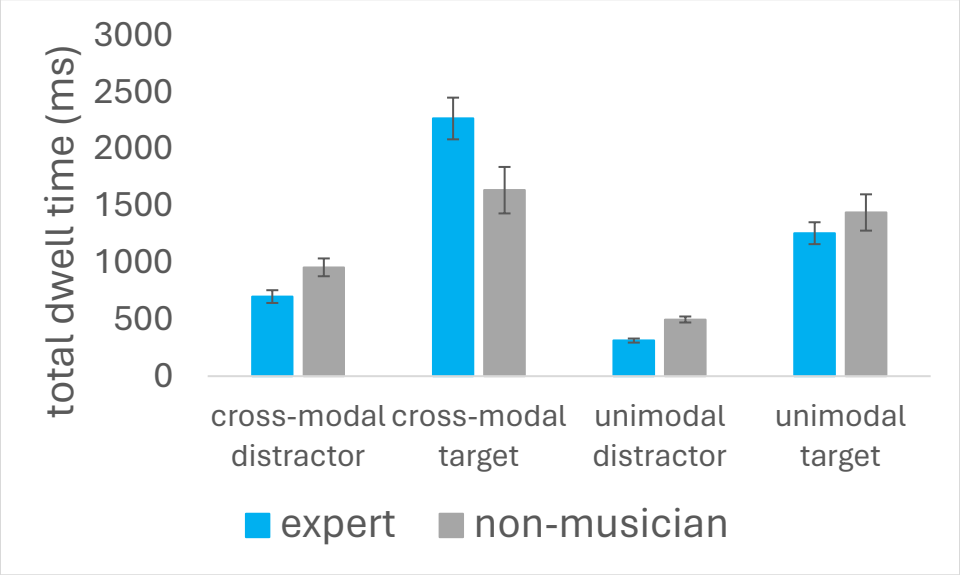

Figure 5. Total Dwell Time (ms)

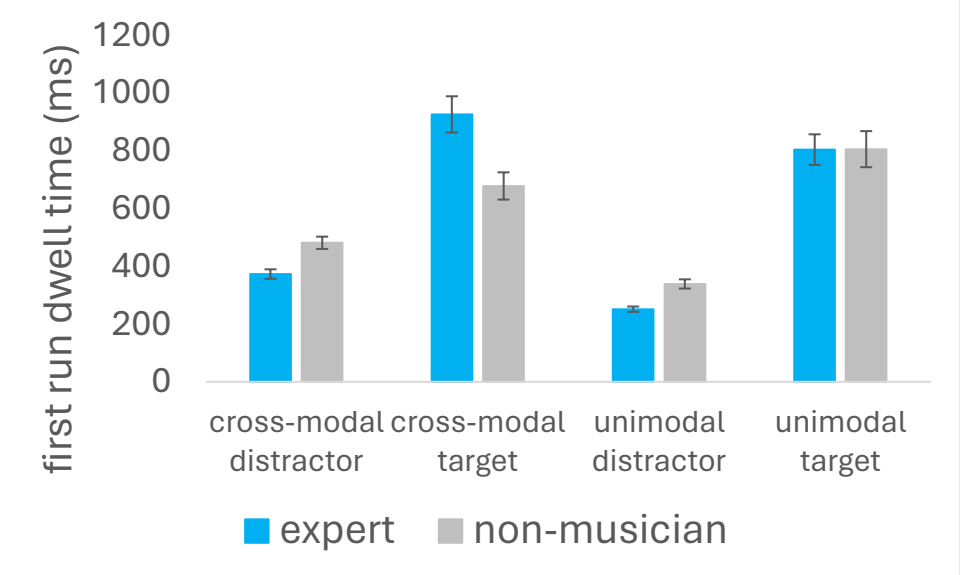

Figure 6. First Run Dwell Time (ms)

	Total Dwell Time (ms) – cross-modal condition			
	b	*SE*	*t*	*p*
Group	-189.59	191.60	-.99	.33
Relevancy	**1272.55**	**40.34**	**31.54**	**< .001**
Group x Relevancy	**-780.55**	**80.66**	**-9.68**	**< .001**
	Total Dwell time (ms) – unimodal condition			
	b	*SE*	*t*	*p*
Group	211.76	132.34	1.60	.12
Relevancy	**1058.23**	**26.34**	**40.17**	**< .001**
Group x Relevancy	100.16	52.62	1.90	.06

Table 5. Results for the LMMs for total dwell time (ms) for the cross-modal and unimodal conditions. Significant results are shown in bold.

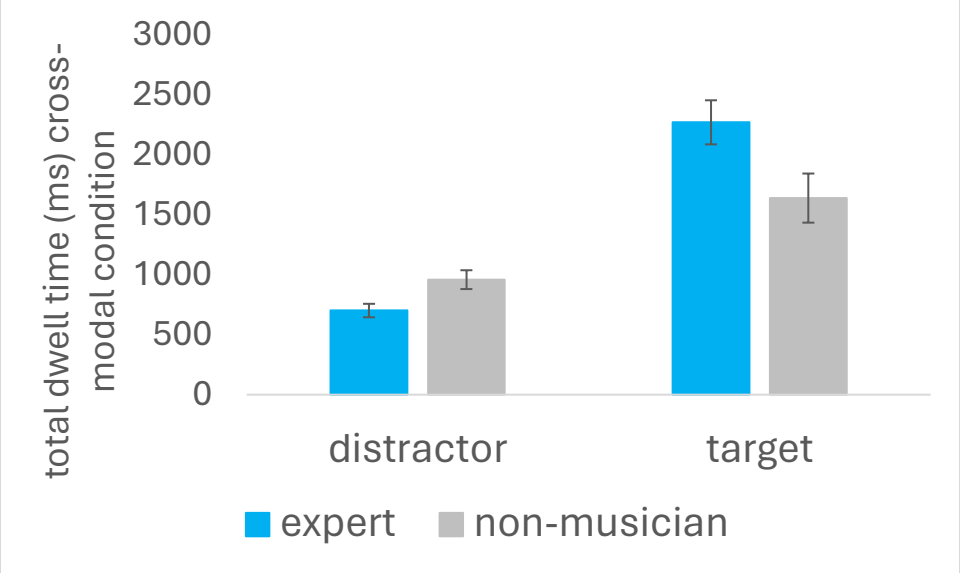

Figure 7. Total Dwell Time (ms) for the cross-modal condition only

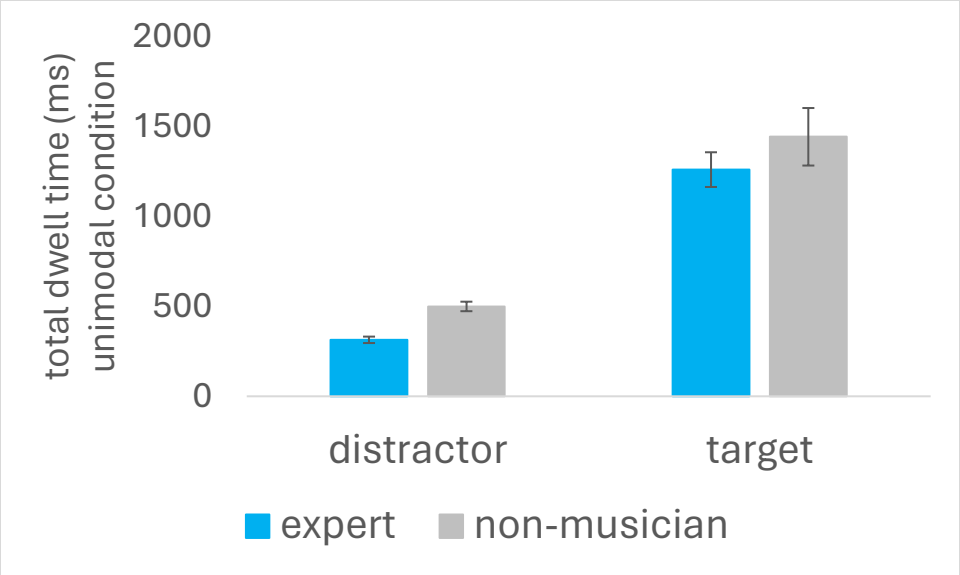

Figure 8. Total Dwell Time (ms) for the unimodal condition only.

	First Run Dwell Time (ms) – cross-modal condition			
	b	*SE*	*t*	*p*
Group	-71.46	50.81	-1.41	.17
Relevancy	**411.07**	**19.42**	**21.17**	**< .001**
Group x Relevancy	**-306.89**	**38.83**	**-7.90**	**< .001**
	First Run Dwell time (ms) – unimodal condition			
	b	*SE*	*t*	*p*
Group	61.09	56.61	1.08	.29
Relevancy	**546.54**	**17.71**	**30.86**	**< .001**
Group x Relevancy	-50.58	35.38	-1.43	.15

Table 6. Results for the LMMs for first-run dwell time (ms) for the cross-modal and unimodal conditions. Significant results are shown in bold.

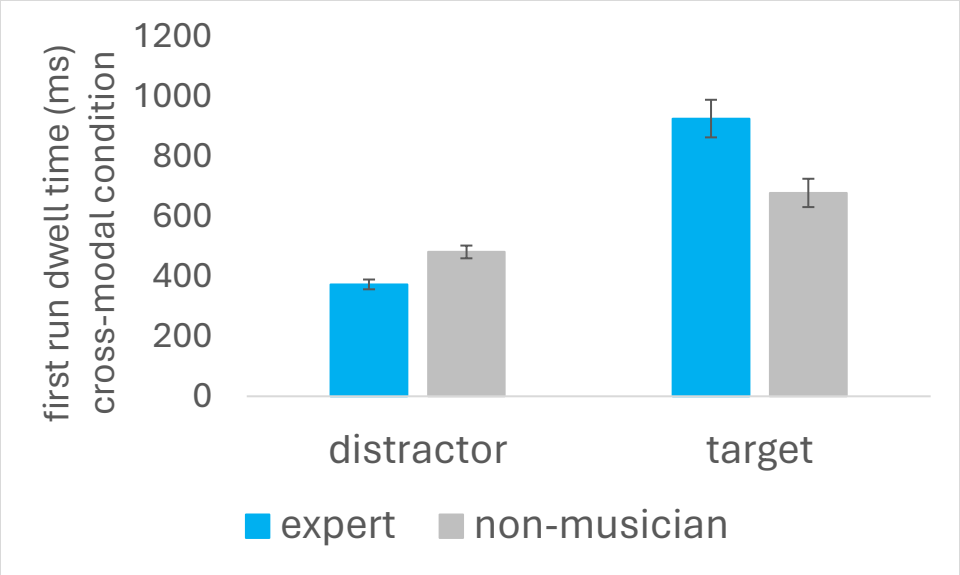

Figure 9. First Run Dwell Time (ms) for the cross-modal condition only.

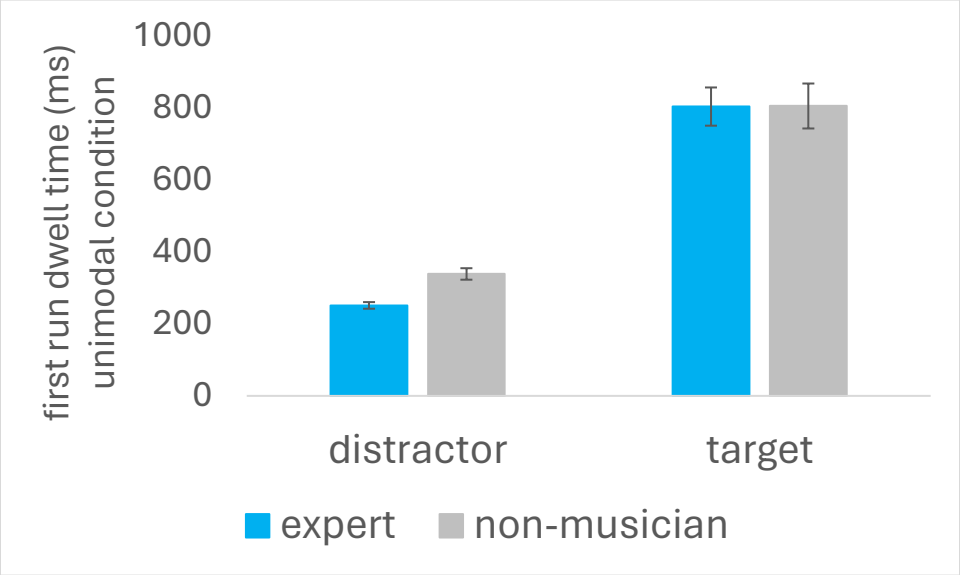

Figure 10. First Run Dwell Time (ms) for the unimodal condition only.

Discussion

According to chunking and template theories (see e.g., Chase & Simon, 1973; Gobet & Simon, 1996; 1998), experts development memory representations of meaningful visual patterns,

chunks as well as larger memory structures called templates. referring to meaningful visual patterns made up of domain-specific features. Chunking may be one mechanism that supports domain-specific performance advantages in experts. Extending on previous research exploring chunking in music (Maturi & Sheridan, 2020), we included an auditory manipulation to test if chunks for music experts were multimodal, which would enable them to use auditory information to form visual search templates. To test experts' ability to precisely encode a search template, eye movements were monitored while participants visually search for a target bar (i.e. search template), that was encoded either auditorily or visually, within a larger musical excerpt (i.e. search array). Since musicians are known to engage in multimodal processing (e.g. Drai-Zerbib & Baccino, 2011), we predicted that expert musicians would be able to precisely encode visual search templates, even if the template was encoded in the auditory modality. We expected there would be greater expertise effects for the cross-modal condition, compared to the unimodal condition.

In support of our hypothesis, experts were more accurate at locating the target bar within the search array, compared to non-musicians, and this effect was magnified in the cross-modal condition. This suggests that experts were able to precisely encode the search template, even if it was encoded auditorily, which enabled them to efficiently search for the target. This replicates findings from Maturi and Sheridan (2020), in which experts were more accurate while visually searching for a target.

Although experts did not have significantly faster reaction times, experts were faster at moving their eyes to the target, as indicated by faster times to first fixations on the target (see Figure 3). This effect was also magnified in the cross-modal condition, suggesting that experts were able to successfully form a visual representation corresponding to the auditory information,

and this visual representation allowed them to rapidly locate the target within the array. This is consistent with the previous medical expertise literature, where expert radiologists were faster at moving their eyes to the target compared to non-experts, with some studies showing experts fixate on an anomaly within one second of viewing the image (for review, see Reingold & Sheridan, 2011).

With dwell-based analyses, we replicated previous work that shows experts are able to fixate in more relevant regions (e.g. Bilalić, 2018; Maturi & Sheridan, 2020; Reingold & Sheridan, 2011). There was a significant three-way interaction between expertise, relevancy, and modality presentation (see Figures 5 and 6). This interaction reflected the fact that relevancy and expertise interact for the cross-modal condition, but not the unimodal condition. In the cross-modal condition, experts had longer total dwells and first-run dwells in the target region than the distractor region, compared to non-musicians, suggesting that experts fixate more on relevant regions (see Figure 7). For total dwell time, in the unimodal condition, although not significant, numerically experts spent less time in both the target and distractor regions compared to non-musicians, suggesting that experts are more familiar with the stimuli, resulting in less time fixating in the target region (see Figure 8).

The dwell-based results suggest that experts spend more time fixating on relevant regions the cross-modal condition only because it takes time to integrate the auditory information into a visual representation. These results are consistent with the medical expertise literature, in which expert radiologists not only fixate faster on the abnormalities, but also spend more time fixating on these regions compared to non-experts (for review, see Reingold & Sheridan, 2011). Because expert musicians are known to engage in multimodal processing (e.g. Drai-Zerbib & Baccino,

2018), it also supports the hypothesis that chunking in music may be multimodal, and that auditory-visual integration may be an underlying mechanism for experts.

Our study also builds on the music reading literature. Previous music reading and eye tracking studies explore the possibility of chunks in music by examining parafoveal processing. If experts process music scores in larger chunks, they may be engaging in greater parafoveal processing, which gives them an advantage in their domain (for review see, Sheridan et al., 2020). Parafoveal advantages have also been observed in other areas of expertise, such as chess and radiology, in which experts make fewer, but longer fixations to take in more information with the parafovea (for review see, Reingold & Sheridan, 2011). Convergent evidence of chunking can be seen in prior music reading studies. Chunking has been observed to facilitate change-detection in experts (Sheridan & Kleinsmith, 2021), observations of different chunking strategies (pitch, rhythm, measure boundaries) used by experts (for review see, Madell & Hérbert, 2008), and in computational modeling which utilized the CHREST (Chunking Hierarchy and REtrieval STructures; Gobet & Lane, 2012) model to correctly categorize music scores using chunking (Bennett et al., 2020). Our study provides additional evidence of chunking in music and that it may be multimodal in experts.

Strengths of this study include examining the novel cross-modal visual search task to examine the possibility of chunking being multimodal and the underlying mechanism of that being auditory-visual integration. Results do support the possibility of chunking being multimodal in musicians. Another strength of this study is that non-musicians were also able to complete this task, which allows for a strong contrast between expert musicians and non-musicians. Lastly, were able to collect data from a large sample of experts and gather additional information on eye movements of experts during visual search of a complex object that they

have a lot of practice with. This allows for greater generalizability to other domains that have a visual search component, such as radiology, TSA agents, military operations, athletics and video games. Limitations include that we still do not know what a chunk is in music, although there are some ideas in the literature, such as beaming or chord patterns (for further discussion, see Sheridan, et al., 2020). Lastly, more work is still needed to understand if auditory-visual integration is a mechanism of multimodal chunking in experts.

If chunking is multimodal in music, future directions include interfering with the integration of auditory-visual information. For example, in the current study, we found that expert musicians were able to form precise visual representations of the search template, even if it was encoded auditorily. This would suggest that experts are able to transfer auditory information into visual. We could disrupt this process by introducing interfering background music during the encoding process. If experts are encoding domain-specific stimuli using auditory and visual strategies, by disrupting this process, experts' ability to form precise representations would be hindered, suggesting that experts are using multimodal chunks to precisely encode the search template.

Future directions also include examining the extent to which the multimodal chunk persists in visual-working memory over time by disrupting the retrieval process (i.e. the search array). Results of the current study suggest that auditory-visual integration takes time, which is why experts have longer dwells in the target region for the cross-modal condition. If we play interfering background music during the visual search or retrieval process, it is possible that experts' ability to integrate auditory-visual information will be disrupted resulting in less efficient search strategies.

In addition, future directions include exploring different levels of expertise, such as novice, intermediate, and expert musicians. Future research could use the current music-related visual search paradigm and include a group of novice and intermediate musicians. This would allow us to explore expertise on a continuum, providing additional evidence for how and when chunks in music are formed and if those chunks are possibly multimodal. Finally, future research could also explore how expertise develops by using a longitudinal design to explore if learning is gradual or if there are sudden changes in learning.

References

Barr D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. https://doi.org/10.3389/fpsyg.2013.00328

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bennett, D., Gobet, F., & Lane, P. C. (2020). Forming concepts of Mozart and Homer using short-term and long-term memory: A computational model based on chunking. In *42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Human, Animals and Machines*

Bilalić, M. (2017). Perceptual Expertise. In *The Neuroscience of Expertise* (pp. 34-9). Chapter, Cambridge: Cambridge University Press.

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., & Williams, A. M. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin, 145*(10), 980-1027. doi: 10.1037/bull00002077

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Chase, W. G. & Simon, H. A. (1973). The mind's eye in chess.

Cox, G. E. & Criss, A. H. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review, 127*(5), 792-828. http://dx.doi.org/10.1037/rev0000195

Drai-Zerbib, V., Baccino, T., & Brigand, E. (2011). Sight-reading expertise: Cross-modality integration investigated using eye tracking. *Psycholoyg of Music, 40*(2), 216-235. doi: 10.1177/0305735610394710

Drai-Zerbib, V. & Baccino, T. (2014). The effect of expertise in music reading: Cross-modal competence. *Journal of Eye Movement Research, 6*(5), 1-10.

Drai-Zerbib, V. & Baccino, T. (2018). Cross-modal music integration in expert memory: Evidence from eye movements. *Journal of Eye Movement Research, 11*(2), 4. doi: 10.16910/jemr.11.2.4

Gobet, F. (2016). Entrenchment, Gestalt formation and chunking. In H.-J. Schmid (Ed.), *Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning*. Berlin: Walter de Gruyter.

Gobet, F. & Lane, P. C. R. (2012). Learning in the CHREST cognitive architecture. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1920-1923). Springer. doi: https://doi.org/10.1007/978-1-4419-1428-6_1732

Gobet, F. & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several chess boards. *Cognitive Psychology, 31*, 1 - 40.

Gobet, F. & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*(3), 225-255. doi: 10.1080/741942359

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. doi: http://dx.doi.org/10.1016/j.bandc.2014.08.002

Hout, M. C. & Goldinger, S. D. (2015). Target templates: the precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception, & Psychophysics, 77*, 128-149. doi: 10.3758/s13414-014-0764-6

Lenth, R. (2023). _emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.8.7, https://CRAN.R-project.org/package=emmeans

Madell, J. & Hérbert, S. (2008). Eye movements and music reading: Where do we look next? *Music Perception, 26*(2), 157-170. doi: 10.1525/MP.2008.26.2.157

Malcolm, G. L. & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Visual, 9*(11):8, 1-13. doi: http://journalofvision.org/9/11/8/, doi:10.1167/9.11.8

Maturi, K. S. & Sheridan, H. (2020). Expertise effects on attention and eye-movement control during visual search: Evidence from the domain of music reading. *Attention, Perception, & Psychophysics, 82(*5), 2201-2208. doi: 10.3758/s13414-020-01979-3

Reingold, E. M., Charness, N., Schultetus, R. S., & Stampe, D. M. (2001). Perceptual automaticity in expert chess players: Parallel encoding of chess relations. *Psychonomic Bulletin & Review, 8*(3), 504-510.

Reingold, E. M. & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In S.P. Liversedge, I.D. Gilchrist, & S. Everling. (Eds.) *Oxford Handbook on Eye Movements* (pp. 528-550). Oxford: Oxford University Press.

Sala, G. & Gobet, F. (2017). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. *Memory & Cognition, 45*, 183-193. doi: 10.3758/s13421-016-0663-2

Schmidt, J. & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology, 62*(10), 1904-1914. doi: https://doi.org/10.1080/17470210902853530

Sheridan, H. & Kleinsmith, A. (2021). Music reading expertise affects visual change detection: Evidence from a music-related flicker paradigm. *Quarterly Journal of Experimental Psychology*, 1-10. doi: 10.1177/17470218211056924

Simoens, V. L. & Tervaniemi, M. (2012). Auditory short-term memory activation during score reading. *PLoS ONE, 8*(1): e53691. doi: 10.1371/journal.pone.0053691

Turner, M. & Engle, R. (1989). Is working memory capacity task-dependent? J. *Memory Language, 28*, 127-154.

Wong, Y. K. & Gauthier, I. (2009). A multimodal neural network recruited by expertise with musical notation. *Journal of Cognitive Neuroscience, 22*(4), 695-713.