Effects of User Expertise on the Efficacy of Multimodal Presentations

Declan Hills
declanhills@usask.ca
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

Carl Gutwin
gutwin@cs.usask.ca
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

Stephen Brewster
Stephen.Brewster@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland

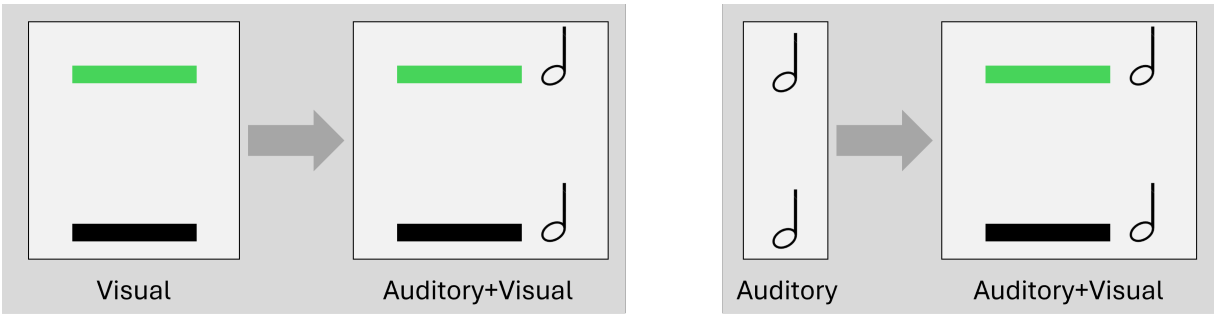


Figure 1: Does expertise in one modality affect the value of adding another modality? Our experiment investigates whether a user's expertise in the audio domain affects the benefit that they get when moving from a visual-only or audio-only presentation to an audio+visual presentation.

Abstract

Multimodal output presents information through multiple sensory channels, and has been shown to improve performance. However, previous work has not considered how expertise in one modality could affect the usefulness of multimodal output. Expertise could potentially reduce the effectiveness of adding another modality (e.g., by interfering with or distracting from the first). To investigate how expertise affects the value of multimodal output, we conducted a study where participants with three levels of audio experience interpreted differences between two values presented as visual, audio, or audio+visual information. We assessed the improvement in time and errors when moving from unimodal to multimodal presentation. Results showed substantial effects of audio expertise: experts improved much less than novices when visual information was added to audio, and improved much more than novices when audio information was added to visual. Our study shows that designers should consider user expertise when designing multimodal output.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

GI '26, Waterloo, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Multimodal Feedback, Expertise, Audio and Visual Feedback

ACM Reference Format:
Declan Hills, Carl Gutwin, and Stephen Brewster. 2026. Effects of User Expertise on the Efficacy of Multimodal Presentations. In *Proceedings of Proceedings of the 50th Graphics Interface Conference (GI '26).* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Multimodal output presents information to the user through multiple sensory channels (e.g., visual, auditory, or tactile). Multimodal output can be used for feedback (e.g., providing both visual and tactile information about a targeting movement [15]), notification (e.g., alerting the user to the occurrence of an event using both visual and audio cues [16]), or information presentation (e.g., showing a visualization to compare two values, with an accompanying auralization). In addition, the distinct output channels of a multimodal system can be used to present the same or different information: when the same information is encoded through different sensory channels, it can reinforce the message and make it more resilient to noise or attentional demands [2]; when different information is presented in each channel, the additional information is complementary to the primary message [9].

Previous research has shown that in several usage scenarios, multimodal presentation can improve performance compared to unimodal presentation. Adding feedback modalities such as haptic and audio to interactions such as icon location, scanning, and drag-and-drop mouse movements can lead to improved task completion time, accuracy, and accessibility [15] [10] [3]. These performance gains are particularly notable in the context of accessible computing; Jacko et al. found that users with visual impairments completing a

drag-and-drop task typically performed best when provided with tri-modal or bimodal feedback as opposed to unimodal feedback [8]. Additionally, multimodal output has improved performance in real-life applications such as PROMUSE, a protein structure alignment software that presents protein traits visually and as audio [6]. There are also examples of multimodal interfaces failing to improve performance - Lee et al. showed that combinations of haptic and visual feedback can be ineffective in drag-and-drop tasks [10], while Lin and Chen found the addition of audio feedback to menu item selection had little-to-no impact on user performance [12].

However, previous work has not considered how expertise in one modality can affect the usefulness of adding another modality. In this paper, we consider how a user's ability to recognize discrete pitches, a skill known as interval identification, could affect the usefulness of combined audio-visual output. Expertise in pitch identification could potentially reduce the effectiveness of multimodal presentation for several reasons: for example, if the user focuses entirely on the modality that they are familiar with, they may not get the reinforcement of the other modality; or if the second modality causes interference with the first, interpretation accuracy could be compromised.

Variability in expertise is perhaps most likely to occur in the audio domain. Different users may have widely different experience with musical training, leading to varying abilities in sub-tasks such as interpreting pitches, intervals, and rhythms – abilities that are critical for working with audio interfaces such as Earcons [1] that may be part of multimodal output.

To explore the effects of audio expertise on the value of multimodal output, we carried out a study that asked participants to identify a difference between two values that was encoded either unimodally (visual or audio only) or multimodally (simultaneous audio and visual presentation of the same information). Twelve differences were represented as stimuli using distance: in the visual domain, using the distance between two lines on the screen; in the the audio domain, using the distance between two musical notes.

Participants first carried out interpretation tasks with the two single modalities (with order counterbalanced), and then completed tasks with the multimodal presentation. We recorded completion time and errors for each trial, as well as participants' subjective confidence. Participants were divided into three expertise groups, based on the amount of musical training they had received: *novices* had no musical interval recognition training and between 0 and 2 years of music training overall, *intermediates* had between 3 and 9 years of combined musical interval identification and general music training, and *experts* had either general musical and interval identification training exceeding 10 years or self-identified as a musical interval expert during recruitment.

Our analysis focused on the difference in performance between unimodal and multimodal presentations, and we carried out the comparison separately for each single modality (i.e., Audio alone vs. Audio+Visual, and Visual alone vs. Audio+Visual). We then analyze the size of the performance difference for the three different levels of audio expertise. (Since multimodal was always seen after unimodal, we expect to see improvement due to increased experience with the task – but this effect is equal across all conditions, and so does not affect comparisons between expertise groups).

Our results show that there are substantial effects of audio expertise on the value of multimodal information presentation:

- Expertise affects how much a user benefits from adding a second modality: audio experts benefited from the addition of audio feedback significantly more than novices.
- Adding visual feedback significantly improved performance for audio novices, but only marginally benefited experts.
- The workload experienced by novices was consistent regardless of modality type, whereas audio experts reported lower Frustration and higher Perceived Performance when using audio information.
- Audio experts reported higher Mental Workload when moving from the audio-only to audio-visual interface while novices did not.

Our study shows that for difference information that can be presented either unimodally or multimodally (and that provides the same information in both modalities), there can be large differences in the value of a multimodal presentation, based on the user's experience with the audio domain. As a result, designers should consider the expertise of their users in the development of multimodal feedback.

2 Related Work

We examine prior research exploring the impact of integrating multimodal feedback into computing systems here, as well as several studies that have looked at multimodality and expertise.

2.1 Output Modalities

2.1.1 Earcons. An audio-counterpart to visual icons, Earcons are audio cues output by a computing system that provide information about the state of a system or provide user feedback [1]. Brewster et al. provided a set of guidelines for implementing pitch as a method of item selection from a set of audio icons, noting that pitches should be distinct for each item in a set [4]. Additionally, Brewster found that musicians were not significantly better at utilizing music-based Earcons than novices. A potential issue with this conclusion is that pitch-recognition expertise was assumed to be correlated to musicianship, but not all musicians are experts at pitch recognition [13]. Thus, analyzing the performance of pitch-recognition experts against novices may result in a more significant performance gap.

2.1.2 Multimodal Systems. Multimodal presentation of information allows a system to communicate with multiple human sensory receptors. This theoretically allows for users of a multimodal system to process information from discrete sensory channels, such as audio and visual, simultaneously [2]. Integrating multimodality into traditionally visual computing tasks has been shown to improve performance; for example, previous work tested the effects of different combinations of haptic, audio, and visual feedback in a mouse drag-and-drop task [15]. In this experiment, participants were asked to sort a series of files into specific folders as quickly and accurately as possible - feedback was provided when the file was dragged over the correct folder by earcons (audio), the mouse rumbling (haptic), or a blue highlight (visual). Vitense et al. found that certain bimodal combinations of the feedback mechanisms

decreased perceived workload in participants, and combining haptic and visual feedback decreased task time. Their implementation of audio feedback did not encode modal redundancy as compared to haptic and visual; many users waited for the audio chime to play before completing their task, which negatively impacted user performance [15]. The study additionally did not consider user's previous expertise in the audio domain or in drag-and-drop tasks.

Additionally, the temporal nature of certain information presented by interfaces is uniquely suited to combined Audio-Visual feedback, as Brewster et al. found when they explored effects of adding Earcons to a visual scanning task [3]. Scanning is a method of item selection, typically requiring users to wait as a cursor highlights individual items in grid, moving from one item to the next after a fixed amount of time. Brewster et al. integrated additional audio feedback into a scanning task by mapping a musical note related to each item's position on the item grid, taking advantage of music's natural mapping onto time [3]. Their study did not take user's musical experience into account, but preliminary results indicated that adding audio information to scanning systems may provide advantages over standard unimodal visual feedback.

2.2 Multimodal Feedback and Expertise

Beyond simple tasks such as dragging and selection, complicated visual information that requires specialized training to interpret can benefit from multimodal presentation. An example of this is PROMUSE - a program that provides both audio and visual representations of protein data [6]. Traits of protein structures typically displayed in the visual domain were mapped to instrumental cues including slap bass, electric piano, and trumpet. Accuracy scores were much higher when users were in the Audio and combined Audio-Visual feedback conditions as opposed to Visual-only. Protein identification experts, who had more experience with traditional visual representations of protein structures, outperformed novices significantly in the Visual-only condition, but were much closer to novice performance in Audio-only, suggesting that experts may not have a large advantage in new modalities. While they performed the same as novices in the Visual-only condition, expert musicians unexpectedly outperformed novices by a wider margin in the combined Audio-Visual task than the Audio-only condition, suggesting a more efficient parallel interpretation of data for experts than novices [6]. However, workload measurements and qualitative information about the expert and novice impressions of PROMUSE were not captured as part of this study, nor were musicians checked for pitch-recognition expertise.

Previous work by Gargiulo et al. suggests that some visual task experts respond differently to the addition of audio feedback than novices in brain-computer interfaces [5]. Their study asked participants to simulate moving a ball to a hole in a visual interface by imagining hand movements, with the ensuing brain activity triggering the movement of the ball. Participants received either Visual feedback through a monitor or combined Audio-Visual feedback, with distinct frequencies indicating a target hit or miss. Several novice subjects experienced large jumps in performance in the Audio-Visual condition, but participants well-trained in the visual interface received little-to-no benefit from the addition of audio feedback. It is unclear if the novices that over-performed in the combined Audio-Visual condition had previous audio expertise that contributed to their increased performance over their cohorts.

The link between audio expertise and multimodal feedback has not been thoroughly explored, but prior research has found a surprising relationship between Visual and combined Visual-Haptic feedback. He and Agah found notable differences in how gaming experts and novices responded to the addition and subtraction of haptic feedback in an Asteroids-like game [7]. The study found that novices were best when given delayed haptic feedback and worst when given visual-only, but experts were better with unimodal visual than multimodal delayed haptic feedback. This suggests that some haptic-visual multimodal presentations may help novice users while hindering unimodal experts, which differs from previous research suggesting that experts are better suited to process multimodal feedback in parallel.

3 Study Methods

We carried out a controlled experiment to better understand how increased expertise with an audio modality affects the value of adding a second modality to the presentation of information. The following sections provide details on the experimental task used in the study, the way in which this task was represented in the two modalities (visual and audio), the participants (and how expertise was determined), the study system and apparatus, and the experimental design.

3.1 Study Design

The study used a within-participants factorial design with two main factors: *Modality* (Audio, Visual, Audio+Visual) was a within-participants factor, and *Expertise* (Novice, Intermediate, Expert) was a between-participants factor. We also considered *Block* (1-5) and *Order* (Audio first or Visual first) as secondary factors.

Our dependent measures of performance were task completion time and errors (that is, total guesses per trial - 1), and these were used to derive our primary measures of interest – the difference between each unimodal presentation and the ensuing multimodal presentation (i.e., the degree to which performance improved with the addition of the other modality). Dependent measures of experience included subjective ratings of effort and preference. The order of unimodal presentations was counterbalanced using a Latin square. Each participant completed 180 trials (3 modalities $\times$ 5 blocks $\times$ 12 trials); with 60 participants, there were 10800 trials.

3.2 Experimental Task

The study used an interpretation task that could arise in several analysis domains: participants were asked to interpret a visual or audio presentation that encoded a difference between two values. Recognizing the distance between two discrete pieces of data is a common task: for example, in assessing the difference between bars in a bar chart. Difference was encoded using distance; either visual distance in pixels (as is common in visual presentations) or pitch difference in semitones (to mirror the encoding of pitch in several audio-based interfaces such as Earcons [1] and Promuse [6]). The information presentation in the two modalities was designed to be as similar as possible. Participants selected one of twelve choices for the distance between the values by clicking on one of

Figure 2: The twelve differences encoded by distance (each bar was shown in sequence for one second).

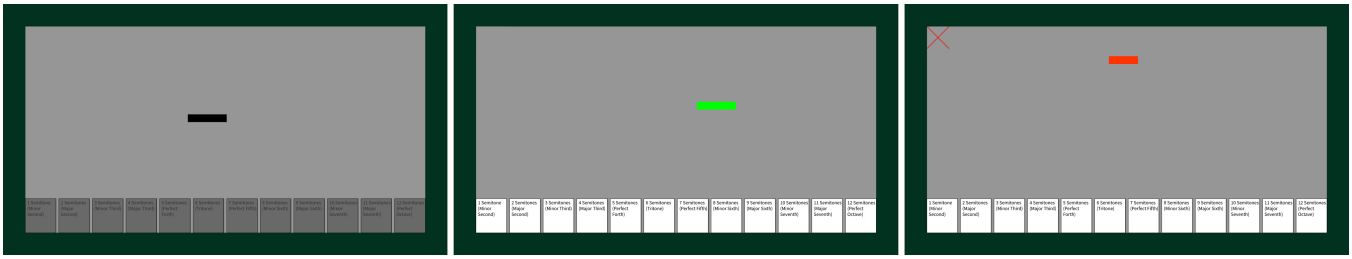


Figure 3: Visual modality presentation. Left: base bar (always shown in the same location). Middle: target bar (shown a varying distance from the base). Right: feedback provided after an incorrect answer (the guess was shown in red for one second, and then the target bar was shown again.

twelve buttons on the screen; the buttons were labelled with both a distance and the musical interval for that distance (e.g., "4 semitones - Major Third"). If the selection was incorrect, participants were shown feedback about their guess, and then allowed to select again until they chose the correct answer. The experimental task (using the visual modality) is shown in Figure 3 (the audio modality is demonstrated in the accompanying video).

3.3 Task Representation – Visual

Each trial displayed a sequence of two coloured bars on the screen, each shown for one second: the first bar was black and represented the base for interpreting the distance (the black bar was always shown in the same location); after one second, the black bar disappeared and a green target bar was shown at a location higher than the black bar. This unusual presentation was done to avoid the cross-modal performance issues found by Vitense et al. [15], where visual feedback lead to faster selection times than audio feedback due to participants waiting for audio cues to finish playing. The vertical distance between the black bar and the green bar encoded the difference (the distance was between 1 and 12 times the thickness of the bars). The vertical distances for the visual modality are shown in Figure 2. After an incorrect answer, the system showed the participant their guess (coloured red), replayed the presentation of the target bar (coloured green), and allowed the participant to answer again (as seen in Figure 3). The participant could then make further selections, until they chose the correct answer.

3.4 Task Representation – Audio

Each trial played a sequence of two notes, each lasting one second. The first note was always "middle C" (C4, approximately 261.63 Hz); the second note was between 1 and 12 semitones higher than the first, with the distance between the notes in semitones representing the notification type. This task allows us to assess the user's ability to perceive differences in the audio domain - similar to an Earcon task where users are selecting between 12 items, each one mapped to a musical note between C4 and C5. Participants used the same on-screen buttons to select a difference as their answer: if they answered incorrectly, the note of their guess was played for one second, followed by the correct second note (also for one second). Participants could then answer again, and continued in this fashion until they got the correct answer.

3.5 Task Representation – Visual+Audio

The multimodal presentation was simply a combination of the audio and visual presentations described above. The two modalities were presented simultaneously – the black bar and the middle-C note were presented for one second, followed by the target bar and note, also presented for one second. The same buttons were used to select a difference amount, and feedback given on an incorrect answer was a combination of the visual and audio feedback described above.

3.6 Ordering of Modalities

The study used two orderings that counterbalanced the presentation of the two unimodal conditions. The orders were: *Audio* → *Visual* → *Audio+Visual*, and *Visual* → *Audio* → *Audio+Visual*. The multimodal condition was always done last, to ensure that participants were familiar with both individual modalities before seeing a multimodal presentation. This means that the Audio+Visual condition had the benefit of additional training, so our analysis does not focus on the absolute difference between unimodal and multimodal presentations – instead, we are interested in the relative differences between these conditions, and the size of the differences for each level of expertise.

Figure 4: Sound-attenuating gaming pods where participants completed the experiment

3.7 Participants and Expertise

60 participants (19 men, 37 women, 4 non-binary) were recruited from a local university. We divided participants into three expertise groups (novice, intermediate, or expert) based on their previous training in the musical task of *interval recognition*, a common element taught in most music lessons, choral, or band training. Participants were determined to be novices if they had less than 2 years of prior music and interval recognition training; intermediates had up to nine years of music and interval training, and experts had more than ten years of training or self-identified as experts. There were 20 participants in each expertise category, although expertise was not evenly distributed across gender due to participant availability:

- Expert: 14 women, 4 men, 2 non-binary
- Intermediate: 10 women, 8 men, 2 non-binary
- Novice: 13 women, 7 men

Although there were more women experts in our participant group, we are unaware of any previous research that indicates interactions between gender, expertise, and multimodal feedback. Ethical approval was provided by the ethics board of the University of Saskatchewan.

3.8 Apparatus

The study was carried out in person with participants seated in sound-attenuating gaming pods (shown in Figure 4). Participants wore Sennheiser PC37X over-ear headphones for the audio information, and visual information was shown on a 27-inch LCD monitor. The study used a custom experimental system written in Processing/Java; the study system presented all trials and recorded all performance data (completion times and errors).

3.9 Procedure

The study system presented the difference-identification task; questionnaires were presented using an online survey tool. Participants completed informed consent and demographics forms, and were interviewed to determine their expertise level. Participants were then assigned to an order group for the presentation of the two individual modalities, using a Latin square. The system provided a brief introduction to the concept of encoding difference with distance (in both the audio and visual domain), and provided information about how to complete the tasks. Participants then carried out tasks in their uni- and multi-modal conditions as specified by their order group. For each condition, users first completed 12 practice tasks to become familiar with the modality and the task; participants then completed 5 blocks of 12 trials (each block presented each of the 12 differences once). Tasks were presented in pseudo-random order, but the sequence was for each condition to ensure a consistent experience (pilot testing determined that participants would not be able to remember the ordering of the 60 trials).

After each modality condition, participants completed a NASA-TLX survey as well as a short questionnaire asking about how they used the modality to complete the tasks. After the Audio+Visual condition, participants also completed an additional questionnaire asking the degree to which they used each modality in completing the multimodal tasks, and whether they found either of the modalities distracting. At the end of the session, participants were asked a set of semi-structured interview questions to find out more about their experiences and their use of the audio and visual information (Table 1).

Semi-structured Interview Question List
1. What steps did you follow to guess audio intervals?
2. What steps did you follow to guess visual intervals?
3. What steps did you follow to guess audio-visual intervals?
4. How did your approach differ between audio and visual intervals?
5. How did your approach differ between audio and audio-visual intervals?
6. How did your approach differ between visual and audio-visual intervals?
7. Did you have a preferred feedback mechanism?

Table 1: Questions asked to participants following the completion of the study and written survey.

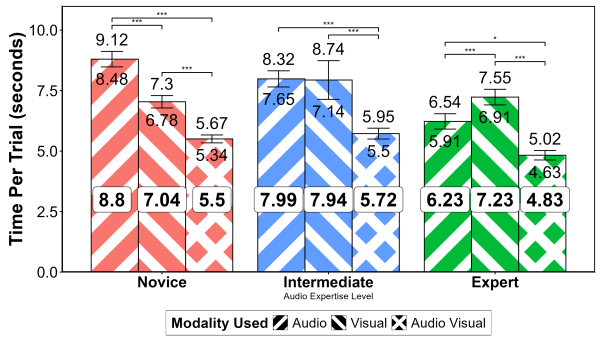
4 Results

Analyses for our two main dependent variables (completion time and errors) are reported below, in two parts: first, we present analyses of factors Expertise and Modality, and second, we present the relative analysis of performance improvements after adding a second modality. We then present subjective measures in a similar fashion. The effect size for significant ANOVA results is reported as generalized eta squared (η^2) [11, 14]. Follow-up tests were corrected using the Holm-Bonferroni method. Five participants were not able to complete the trials for all three feedback mechanism during the allotted study time and so were removed from the final result set (900 trials). Any trial where the total number of guesses exceeded 12 was removed as there were only 12 possible answer to a given question - these trials typically involved a participant experimenting with the feedback system (162 trials). Finally, any trial that exceeded 3 standard deviations in time or errors were removed to account for outlier performance (239). Ultimately, 1301 of the 10800 trials (12.0%) were removed from the dataset. In all charts below, error bars show 95% confidence intervals.

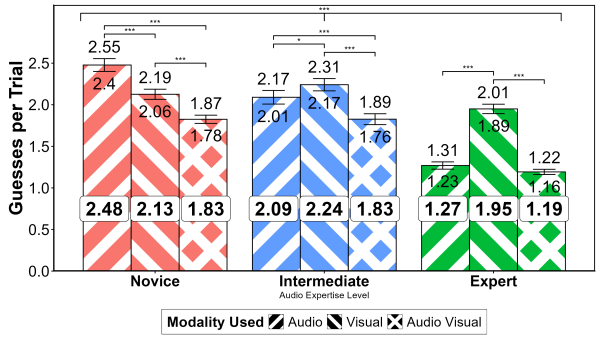
4.1 Performance by Expertise and Modality

We measured trial completion time as the time between the initial presentation of the base bar or note and the participant responding

with the correct answer. Figures 5 (a) and (b) show participant performance in these metrics by modality used and expertise level. (Recall that *expertise* relates to the participant's expertise in the audio domain, and in particular in the task of interval recognition, as introduced above).



(a)



(b)

Figure 5: (a) Mean time per trial and (b) mean errors per trial by Expertise and Modality (± 95% CI). Significant differences between categories are denoted "*" for $p \leq 0.05$ and "*" for $p \leq 0.001$.**

We carried out factorial analyses of the two main factors (although we note that our primary analyses involve the relative improvement between modalities – see next section). Table 2 shows the results of the 3 × 2 (Expertise × Modality) ANOVA. Overall, there was no difference in terms of completion time between different levels of expertise, but experts committed fewer errors. There was a significant difference based on Modality, with the Audio+Visual presentation faster and more accurate overall, but there was no difference on either measure between unimodal Audio and Visual presentations. For both completion time and errors, there was a strong interaction effect between Expertise and Modality, contributing to observed relative improvement (reported below).

Metric	Factor	DF (n,d)	F	p	η^2
Time	Expertise	2, 49	570.95	0.15	
	Modality	2, 98	56.85	**4.08e-17**	0.19
	Expertise × Modality	4, 98	5.59	**4.28e-04**	0.04
Errors	Expertise	2, 49	21.15	**2.39e-07**	0.40
	Modality	2, 98	71.22	**7.94e-20**	0.24
	Expertise × Modality	4, 98	25.89	**1.19e-14**	0.19

Table 2: Effects of Expertise and Modality on Completion Time and Errors.

4.2 Relative Improvement with Multimodal Presentation

To investigate the effects of expertise on the relative gain achieved by moving from a unimodal to a multimodal presentation (and to explore the interaction revealed by the basic analysis above), we measured the difference between participant time and errors in unimodal conditions (Audio-only and Visual-only) and the Audio+Visual condition. This was done by calculating each participant's change in time and error in the multimodal condition as a percentage of their original unimodal results (regardless of whether they received training in the audio-only or visual-only interface first), using the following formulas:

For completion time:

$$\textbf{A+V vs A}_{time} = \frac{A{+}V_{time} - A_{time}}{A_{time} * 100}$$

$$\textbf{A+V vs V}_{time} = \frac{A{+}V_{time} - V_{time}}{V_{time} * 100}$$

For errors:

$$\textbf{A+V vs A}_{errors} = \frac{A{+}V_{errors} - A_{errors}}{A_{errors} * 100}$$

$$\textbf{A+V vs V}_{errors} = \frac{A{+}V_{errors} - V_{errors}}{V_{errors} * 100}$$

These results are shown in Figures 6, 7, 8, and 9. As can be seen from the charts, there were marked differences in how the different expertise groups benefited from the addition of the other modality, for both completion time and errors. When moving from Visual-only to Audio+Visual (i.e., adding Audio), experts improved considerably more than novices; when moving from Audio-only to Audio+Visual (i.e., adding Visual), experts improved very little compared to novices. In all cases, the effects on intermediate expertise participants were between those of experts and novices.

We carried out one-way ANOVAs with factor Expertise on each unimodal-to-multimodal comparison (Table 3). For completion time, the effect of Expertise on moving to multimodal was significant for Audio+Visual vs Audio, but not for Audio+Visual vs Visual. For errors, both comparisons were significant.

4.3 Subjective Measures

We measured the NASA-TLX scales Mental Workload, Temporal Workload, Perceived Performance, Frustration, Effort, and Temporal Workload. Lower ratings imply lower workload for all scales except Perceived Performance, where higher is better. Figures 10 and 11 provide the overall NASA-TLX results for each significant

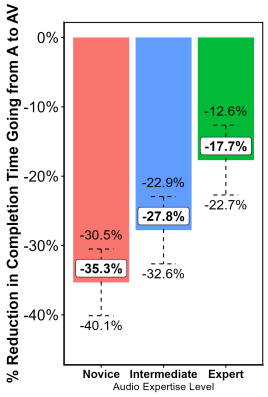


Figure 6: Percent reduction in time in A+V vs A (± 95% CI).

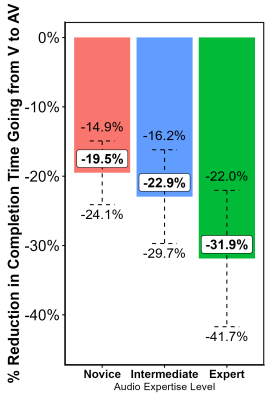


Figure 7: Percent reduction in time in A+V vs V (± 95% CI).

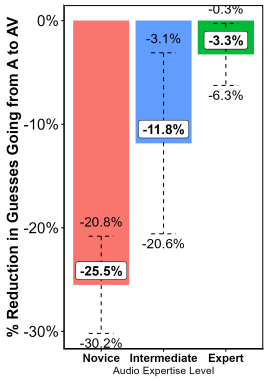


Figure 8: Percent reduction in errors in A+V vs A (± 95% CI).

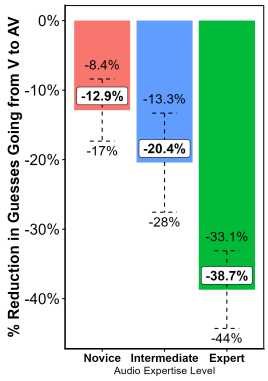


Figure 9: Percent reduction in errors in A+V vs V (± 95% CI).

Metric	Comparison	Factor	F	p	η^2
Completion Time	A→A+V	Expertise	14.15	**1.41e-05**	0.37
	V→A+V	Expertise	3.05	0.06	
Errors	A→A+V	Expertise	21.69	**1.79e-07**	0.47
	V→A+V	Expertise	15.20	**7.34e-06**	0.38

Table 3: Factorial analysis, within-participant improvement from unimodal to multimodal. DF(2, 49)

relationship. The fulls results of the two-factor ANOVA (Expertise x Modality) can be found in the appendix (Table 6).

Frustration, Perceived Performance, and Temporal Workload were significantly affected by expertise levels. Experts experienced less Frustration and Temporal Workload than novices, while reporting higher Perceived Performance. Intermediate participants reported a significantly lower mean Perceived Performance than experts; they also reported a worse Perceived Performance than novices, but this difference was not found to be significant.

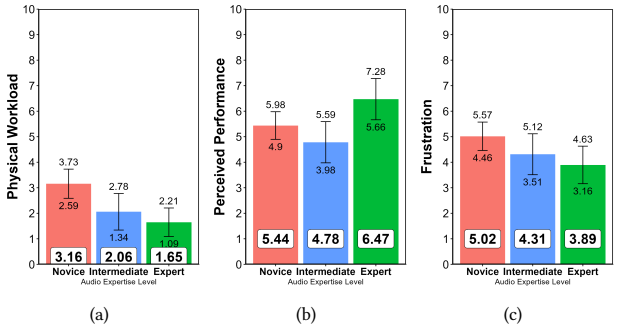


Figure 10: By Expertise Level: (a) Temporal Workload (b) Perceived Performance (c) Frustration (± 95% CI).

Modality used had a significant impact on reported scores for Mental Workload, Frustration, Perceived Performance, and Temporal Workload. The Audio+Visual condition led to a significant reduction in reported Frustration, Temporal Workload, and Mental Workload as compared to the Visual condition. The Audio+Visual condition improved Perceived Performance over both the Audio and Visual condition. Overall, no significant difference was identified in workload between the Audio and Visual conditions.

A strong interaction between Modality and Expertise was observed in Perceived Performance and Frustration workload (see appendix: Table 6). Novices reported similar Perceived Performance and Frustration regardless of modality used, while intermediate participants rated Perceived Performance as higher for Audio+Visual than for Audio or Visual, as well as reporting significantly lower Frustration levels for Audio+Visual. Experts reported similar Frustration and Perceived Performance scores whenever Audio feedback was present, but had significantly increased Frustration and lowered Perceived Performance when completing the Visual-only task.

4.4 Relative Workload Improvement with Multimodal Presentation

We calculated the relative improvement of participant workload scores moving from a unimodal to multimodal representation. In a manner similar to section 4.2, we calculated the difference between unimodal workload (Audio-only and Visual-only) and multimodal workload (Audio+Visual) for each TLX scale as follows:

$$\textbf{A+V vs A}_{workload} = \frac{A{+}V_{workload} - A_{workload}}{A_{workload} * 100}$$

$$\textbf{A+V vs V}_{workload} = \frac{A{+}V_{workload} - V_{workload}}{V_{workload} * 100}$$

The results of this calculation are in the appendices (Figures 15 and 16). Overall, experts saw a greater decrease in Frustration and higher Perceived Performance when moving from Visual-only to Audio+Visual as compared to novices. Conversely, audio experts faced an increase in Mental Workload when moving from Audio-only to combined Audio-Visual. This contrasts how novices experienced the same transition, where adding visual feedback led to decreased Mental Workload.

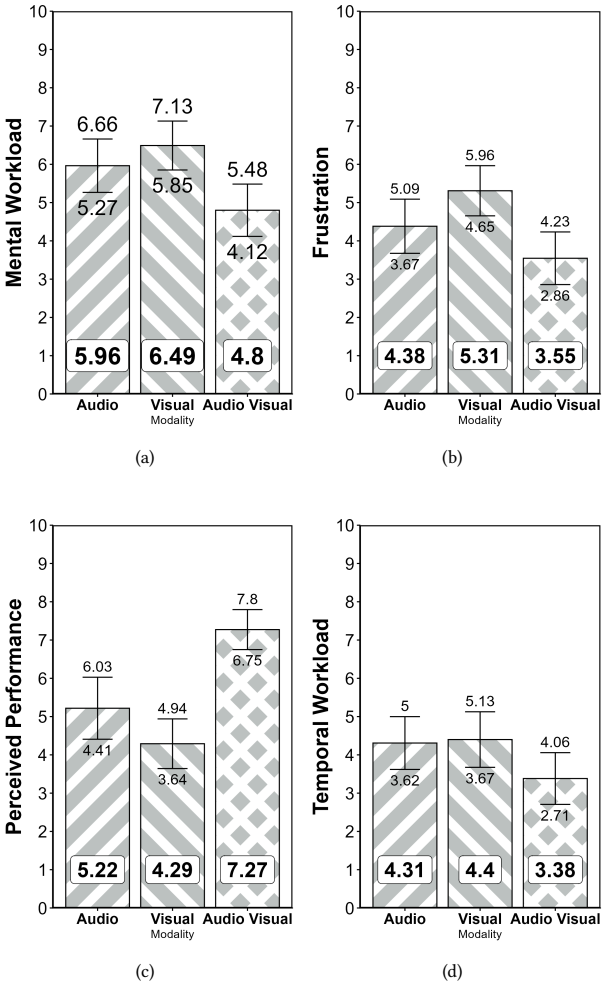


Figure 11: By Modality: (a) Mental Workload (b) Frustration (c) Perceived Performance (d) Temporal Workload (± 95% CI).

One-way ANOVAs with ART transformations were performed on the two unimodal-to-multimodal pairs (see Table 4). The differences in Mental Workload and Frustration were found to be significantly affected by expertise when moving either Visual-only or Audio-only to Audio+Visual, regardless of order. Perceived Performance was significantly impacted in both transitions.

4.5 Preferences

To assess participant preferences for the different modalities, we analyzed the written responses to the in-study survey. We found relevant differences in participant answers to the following questions:

(1) *"After receiving a new interval, which feedback mechanism did you rely on most to inform your guesses?"*
(2) *"After making an incorrect guess, which feedback mechanism did you rely on most to inform your next guess?"*

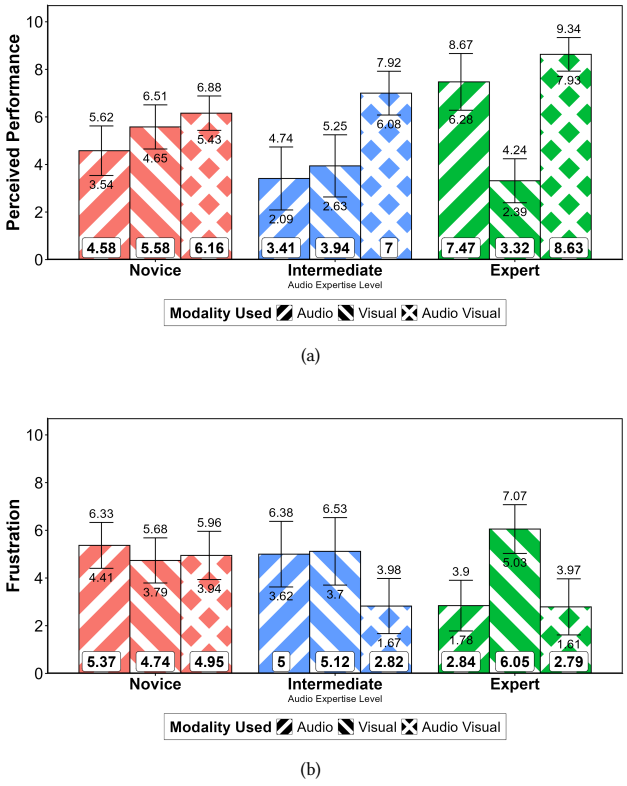


Figure 12: by Modality and Expert: (a) Perceived Performance (b) Frustration (± 95% CI).

TLX Scale	Comparison	Factor	F	p	η^2
Mental	A → A+V	Expertise	3.38	**0.04**	0.12
	V → A+V	Expertise	0.00	0.80	
Physical	A → A+V	Expertise	1.29	0.29	
	V → A+V	Expertise	1.63	0.21	
Temporal	A → A+V	Expertise	0.81	0.45	
	V → A+V	Expertise	0.90	0.41	
Perceived Performance	A → A+V	Expertise	5.73	**0.01**	0.19
	V → A+V	Expertise	10.77	**1.33e-4**	0.31
Frustration	A → A+V	Expertise	2.22	0.12	
	V → A+V	Expertise	4.82	**0.01**	0.16
Effort	A → A+V	Expertise	0.75	0.48	
	V → A+V	Expertise	0.25	0.78	

Table 4: Factorial analysis, within-participant change in NASA-TLX Scores from unimodal to multimodal. DF (n,d) = (2, 49) for all values.

(3) *"Rate what percentage of your answer was driven by audio feedback vs. visual feedback"*
(4) *"(If applicable) Rate how much the less-used modality hindered or helped your ability to identify intervals"*

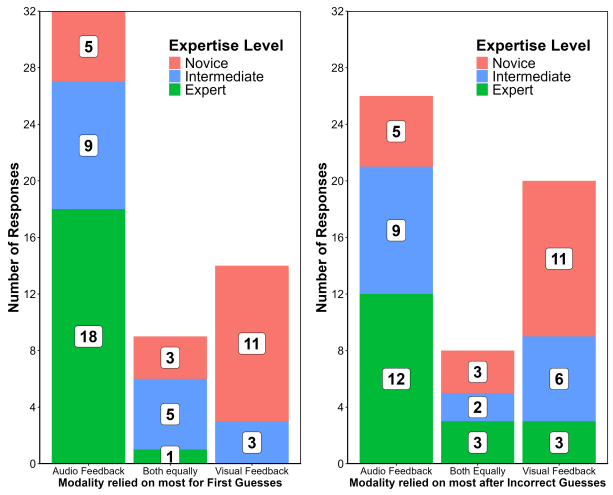
Participants selected either "Audio Feedback", "Both equally", or "Visual Feedback" for questions 1 and 2, a sliding scale ranging from "100% Audio" to "100% Visual" for question 3, and another

sliding scale ranging from "Significantly Hindered" to "Significantly Helped" for question 4. We performed a one-factor ANOVA (Expertise) for questions 3 and 4 (Table 5).

Question	Factor	DF	F	p
Q3	Expertise	2	13.60	**1.77e-05**
Q4	Expertise	2	4.76	**0.013**

Table 5: ANOVA analysis of participant responses to Questions 3 and 4

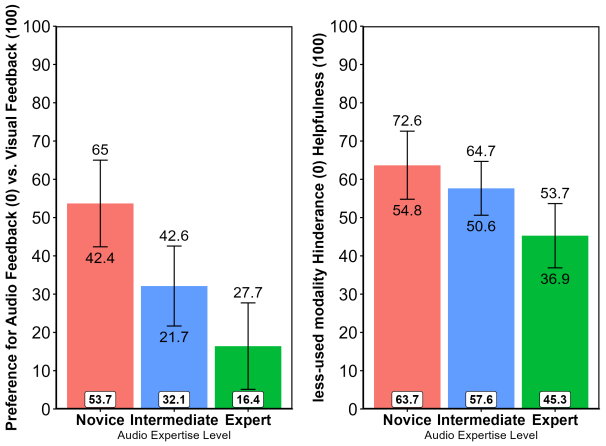
As seen in Figure 13 (a), participant's answers to question 1 were strongly connected to their expertise level. Those with higher levels of expertise were more likely to report relying on audio feedback to make the first guess of their trial in the multimodal condition (18/19 responding experts), while the majority of novices expressed a preference for visual feedback (11/19 responding novices). Of the 55 respondents, only 9 expressed an equal preference for audio and visual feedback (3 novices, 1 expert, and 5 intermediate participants). Similarly, participant answers to question 2 indicated that they tended to rely on either Audio or Visual feedback when correcting mistakes in the multimodal condition (Figure 13 (b)); only 8 of 54 participants stated that they used both audio and visual feedback equally. Somewhat predictably, audio experts were most likely to report audio as their preferred modality in the combined condition, while novices tended to lean on visual feedback.



(a) (b)

Figure 13: By Expertise: (a) First-guess modality preference (b) Preferred feedback modality for corrections

An analysis of Question 3 revealed that experts were significantly more likely to attribute their answers in the multimodal condition to audio-only than visual-only or audio and visual equally, while novices typically reported using an even mixture of audio and visual (Figure 14 (a)). Experts were also significantly more likely to report that the less-used modality made no difference or hindered their



(a) (b)

Figure 14: (a) Q3: Portion of multimodal answers driven by Audio-only (0) vs. Visual-only (100)
(b) Q4: Reported hindrance (0) vs. helpfulness (100) of second-most used modality

performance when answering Question 4, while novices tended to find that adding an extra modality was useful (Figure 14 (b)). Intermediate users did not have a preference different from experts or novices, but notably had a positive view of the less-used modality overall.

4.6 Semi-Structured Interviews

We analyzed participants' semi-structured interviews by transcribing their answers to each of the 7 questions (see Table 1), utilizing these to categorize participants based on whether they:

(1) used an expert-level strategy for Audio-only interval identification (regardless of expertise level)
(2) used visual feedback in the Audio+Visual condition
(3) used a different feedback mechanism to correct mistakes than they used to make an initial guess in the Audio+Visual condition

After categorization, participant groups were assessed for common themes, including strategies used for auditory and visual distance interpretation, preferred modality, and perceived differences between the Audio-only and Visual-only conditions.

28 of the 60 participants used expert-level strategies in the Audio-only and Audio+Visual conditions, including singing major scales aloud, associating the two notes in an interval with a popular song, or imagining playing the notes between the intervals on an instrument. Of these 28, we found that 12 people reported ignoring the visual feedback in the Audio+Visual condition, with several reporting they found it distracting or unhelpful. 15 of the remaining expert-level strategy users expressed a preference using audio-feedback generally but used the visual feedback for small corrections, and one participant reported a preference for using visual feedback to drive all choices.

Of the 32 non-strategy-using participants, only one reported that they ignored the visual feedback in the Audio+Visual condition. The remaining 31 all used visuals to correct mistakes, with seven using visuals to drive all interactions. Overall, participants who did not use an expert-level audio strategy (i.e., those who typically reported "guessing" the audio intervals) were much more likely to use both Audio and Visual feedback in the multimodal condition compared to experts.

5 Discussion

The study provided several main findings about the ways that prior expertise affects the usefulness of adding another modality to the presentation of difference information:

- The benefit of adding a second modality was strongly related to participant expertise: adding audio information led to a substantial benefit for participants who were audio experts, but very little benefit for audio novices.
- Similarly, adding visual information led to a substantial benefit for participants who were novices in the audio domain, but very little benefit for audio experts.
- Novices experienced similar overall workload regardless of the number or type of modalities used. This differed for experts, who reported lower Frustration and higher Perceived Performance when provided with audio in both the uni-and-multimodal conditions.
- Experts reported increased Mental Workload when moving from Audio-only to Audio+Visual, but novices did not.

In the following sections, we provide initial explanations for these results, discuss how the findings can be generalized and applied to real-world settings, and outline limitations of the research and directions for future work.

5.1 Explanations for Main Results

5.1.1 Why Adding the Expert Modality Primarily Helped Experts. That experts performed best in the Audio-only and Audio+Visual condition is unsurprising - when given their preferred feedback mechanism, experts were able to lean on best-practice and years of training interpreting the information provided by system. When experts transitioned from unfamiliar unimodal presentation to multimodal, they were able to perform as if they were just using unimodal in their preferred modality. Their multimodal performance did not significantly eclipse that of their preferred unimodal, suggesting that they do not adapt their interpretation of the dataset - one audio expert described extra visual feedback as "nice to have (...) but I really don't use it at all".

5.1.2 Why Adding the Non-expert Modality Did not Help Experts. We found little evidence that the addition of visual feedback significantly improved expert performance. Experts reported ignoring Visual feedback altogether (e.g., one participant noted that they "didn't pay attention to the visual" during the study, and another who described the visuals as a "hindrance" closed their eyes during the Audio+Visual trials). He and Agah found that extra modalities can be perceived as distracting [7]; we found that while experts generally met or raised their performance transitioning from Audio-only to Audio+Visual, they did experience higher Mental Workload in the multimodal condition than novices (e.g., one participant stated: "every time I looked at it, it skewed my answer and made me irritated"). Several experts commented that the extra modality helped them correct mistakes (e.g., "when I got something incorrect, which happened about, I think, three or four times, I did rely on both the audio and the visual"), but this is not reflected in the results: experts made only 3.3% fewer errors on multimodal trials than unimodal, a negligible improvement when order effects are considered.

Hansen and colleagues were surprised to observe that musical experts in a complex protein interpretation task outperformed novices by a larger margin in combined Audio-Visual than in Audio [6], and theorized that experts were able to process information more readily in parallel. We did not see this behaviour; novices were able to narrow the performance gap between themselves and experts in the combined Audio+Visual condition, and were more likely to utilize strategies that integrated both audio and visual stimuli. It is possible that this is due to the relative simplicity of our spatial distance matching task; using audio to represent several dimensions may results in a wider performance gap between novices and experts.

5.2 Generalizing the Findings

5.2.1 Application of Results. Our results indicate that pitch recognition experts perform faster and more accurately than novices in a redundantly-encoded combined audio-visual recognition task. This has potential implications for systems that utilize pitch characteristics alongside visuals: for example, any system using Earcon-style audio notifications in addition to visual feedback should consider the audio expertise of their users, as those users are more likely to ignore the visual characteristics of the interface (or be distracted by them). In addition, systems such as Promuse [6] may be affected by unimodal expertise: experts in their study may have leaned on audio more heavily in the multimodal condition while novices focused on visual, leading to the gap in performance being accentuated in the combined condition.

5.2.2 Interval Training and Auditory Interfaces. Brewster et al. provided guidelines for utilizing pitch as a method of item selection using Earcons, while controlling for general musicianship[4], but we found that musicians who specifically had expertise in interval identification were faster and more accurate in the Audio+Visual than novices and those with an intermediate-level of training. The novice and intermediate groups were not significantly different, which does suggest that a moderate level of musical expertise may not impact performance in multimodal interfaces. Our results suggest that future research that explores the intersection between auditory interfaces and musicianship should more carefully consider which specific musical skills may transfer to new technologies. In addition, our study considered only one dimension of auditory presentation, and further studies should consider the effects of expertise in other dimensions such as amplitude, timbre, or rhythm.

5.3 Limitations and Future Work

Our study has several limitations, each of which provide directions for further work. First, the identification task that we had participants complete was not a real-world application such as

icon selection. Although our task included the atomic actions that build complex real-world applications (participants estimated visual distances as they would when reading a chart and estimated pitch differences as they would in an Earcons interface), further studies are needed with audio and visual information embedded in real-world tasks.

Second, the visual presentation of our system different from many traditional visual tasks - most systems portray visual information persistently while our visual cues were dynamic and only visible for a few seconds. This was done to maintain consistent feedback across both our visual and audio modalities, allowing for better direct comparison between the visual and audio system. However, future work should explore differing visual encoding mechanisms in a multimodal context.

Third, our study focused on interpretation of distance as encoded in audio and visual information, but several other task types exist, and further studies are needed to consider the effects of expertise on recognition, notification, and guidance tasks.

Fourth, we focused on redundantly-coded information in our study. While many systems do use this approach (e.g., mobile notifications), complementary encoding is also common. Our findings may extend to these scenarios (for example, experts who focused on a single modality in our study may similarly focus on that modality for complementary encoding), but further research is needed to explore these scenarios.

Fifth, although we do not believe that our results were affected by the imbalance of expertise across gender in our participants, future work can investigate demographic interactions with expertise and multimodal interpretation.

6 Conclusion

Audio expertise can play an important role in the recognition of multimodal information. It is not surprising that users rely on their preferred modality, but our study provided new results showing that that additional visual information did not improve expert performance. We found little evidence that experts utilized visual information better than novices in spite of their audio experience. Additionally, novice multimodal performance was strongly related to their performance with visual information (their preferred modality). Although multimodal systems have been shown to be valuable both from accessibility and performance standpoints, our results show that expertise can have a strong effect on the value of adding a second modality – application designers should carefully consider the expertise of their users when designing multimodal feedback.

References

[1] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and Icons: Their Structure and Common Design Principles. *Human–Computer Interaction* 4, 1 (1989), 11–44. doi:10.1207/s15327051hci0401_1 arXiv:https://www.tandfonline.com/doi/pdf/10.1207/s15327051hci0401$_1$

[2] Paul Brett. 1997. A comparative study of the effects of the use of multimedia on listening comprehension. *System* 25, 1 (1997), 39–53. doi:10.1016/S0346-251X(96)00059-0

[3] Stephen Brewster, Veli-Pekka Räty, and Atte Kortekangas. 1996. Enhancing Scanning Input with Non-Speech Sounds. 10–14. doi:10.1145/228347.228350

[4] Stephen Brewster, Peter Wright, and Alistair Edwards. 1993. An evaluation of earcons for use in auditory human-computer interfaces. 222–227. doi:10.1145/169059.169179

[5] Gaetano Gargiulo, Armin Mohamed, Alistair Mcewan, Paolo Bifulco, Mario Cesarelli, Craig Jin, Mariano Ruffo, Jonathan Tapson, and André van Schaik. 2012. Investigating the role of combined acoustic-visual feedback in one-dimensional synchronous brain computer interfaces, a preliminary study. *Medical devices (Auckland, N.Z.)* 5 (09 2012), 81–88. doi:10.2147/MDER.S36691

[6] Marc D Hansen, Erik Charp, Suresh Lodha, Doanna Meads, and Alex Pang. 1999. PROMUSE: a system for multi-media data presentation of protein structural alignments. In *Biocomputing'99*. World Scientific, 368–379.

[7] Fei He and Arvin Agah. 2001. Multi-Modal Human Interactions with an Intelligent Interface Utilizing Images, Sounds, and Force Feedback. *Journal of Intelligent and Robotic Systems* 32 (10 2001), 171–190. doi:10.1023/A:1013969524676

[8] Julie Jacko, Ingrid Scott, François Sainfort, Leon Barnard, Paula Edwards, V. Emery, Thitima Kongnakorn, Kevin Moloney, and Brynley Zorich. 2003. Older adults and visual impairment: What do exposure times and accuracy tell us about performance gains associated with multimodal feedback? *Conference on Human Factors in Computing Systems - Proceedings*, 33–40. doi:10.1145/642611.642619

[9] Simeon Keates and Peter Robinson. 1998. The use of gestures in multimodal input. In *Proceedings of the Third International ACM Conference on Assistive Technologies* (Marina del Rey, California, USA) *(Assets '98)*. Association for Computing Machinery, New York, NY, USA, 35–42. doi:10.1145/274497.274505

[10] Ju-Hwan Lee, Ellen Poliakoff, and Charles Spence. 2009. The Effect of Multimodal Feedback Presented via a Touch Screen on the Performance of Older Adults. In *Haptic and Audio Interaction Design*, M. Ercan Altinsoy, Ute Jekosch, and Stephen Brewster (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 128–135.

[11] Timothy R. Levine and Craig R. Hullett. 2002. Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research* 28, 4 (2002), 612–625. doi:10.1111/j.1468-2958.2002.tb00828.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2958.2002.tb00828.x

[12] Zhongzhen Lin and Chien-Hsiung Chen. 2023. The influence of dynamic page interaction and multimodal operation feedback on the user experience of central bank digital currency. *Displays* 80 (2023), 102516–.

[13] David Little, Henry Cheng, and Beverly Wright. 2018. Inducing musical-interval learning by combining task practice with periods of stimulus exposure alone. *Attention, Perception, Psychophysics* 81 (08 2018). doi:10.3758/s13414-018-1584-x

[14] S.F. Olejnik and James Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological methods* 8 (12 2003), 434–47. doi:10.1037/1082-989X.8.4.434

[15] H. S. VITENSE, J. A. JACKO, and V. K. EMERY. 2003. Multimodal feedback: an assessment of performance and mental workload. *Ergonomics* 46, 1-3 (2003), 68–87. doi:10.1080/00140130303534 arXiv:https://doi.org/10.1080/00140130303534 PMID: 12554399.

[16] David Warnock, Marilyn R. McGee-Lennon, and Stephen Brewster. 2011. The impact of unwanted multimodal notifications. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (Alicante, Spain) *(ICMI '11)*. Association for Computing Machinery, New York, NY, USA, 177–184. doi:10.1145/2070481.2070510

A Appendix

Metric	Effect	DFn	Sum Sq	Sum Sq. res	F	p
Mental Workload	Expert	2	13857.19	358522.38	2.84	0.06
	Modality	2	33531.91	340282.17	7.24	**0.00**
	Expert:Modality	4	14143.80	358218.17	1.45	0.22
Physical Workload	Expert	2	52646.10	309048.42	12.52	**9.52e-06**
	Modality	2	372.85	350613.43	0.08	0.92
	Expert:Modality	4	2965.11	349306.07	0.31	0.87
Temporal Workload	Expert	2	983.82	372309.85	0.19	0.82
	Modality	2	15520.11	357746.04	3.19	**0.04**
	Expert:Modality	4	11413.82	362178.08	1.16	0.33
Perceived Performance	Expert	2	43235.35	328193.00	9.68	**0.00**
	Modality	2	106790.97	262433.96	29.91	**1.26e-11**
	Expert:Modality	4	87620.80	285490.94	11.28	**5.22e-08**
Frustration	Expert	2	19502.20	353232.45	4.06	**0.02**
	Modality	2	35622.16	337501.74	7.76	**0.00**
	Expert:Modality	4	42932.84	330228.65	4.78	**0.00**
Effort	Expert	2	6664.52	366648.43	1.34	0.27
	Modality	2	2419.69	371294.45	0.48	0.62
	Expert:Modality	4	4661.96	368914.64	0.46	0.76

Table 6: Table containing two-way ANOVA analysis ofall Workload metrics

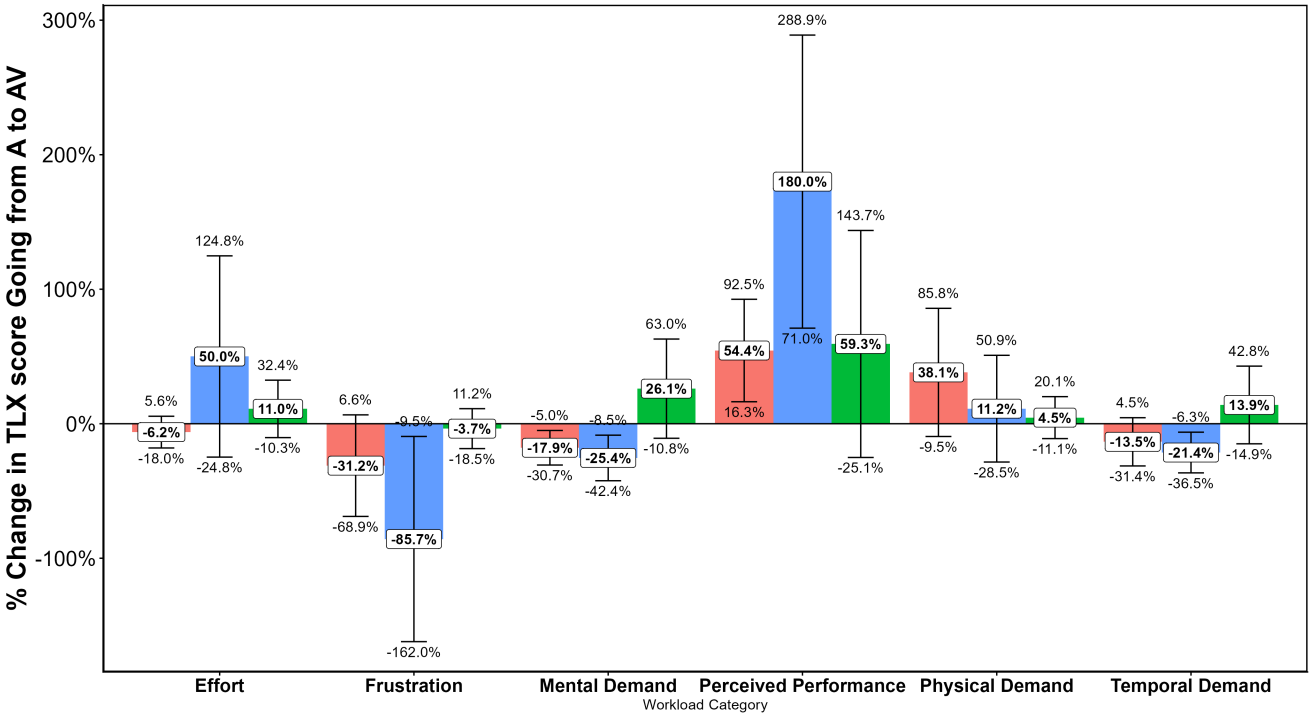


Figure 15: Percent change in workload for A→A+V (± 95% CI).

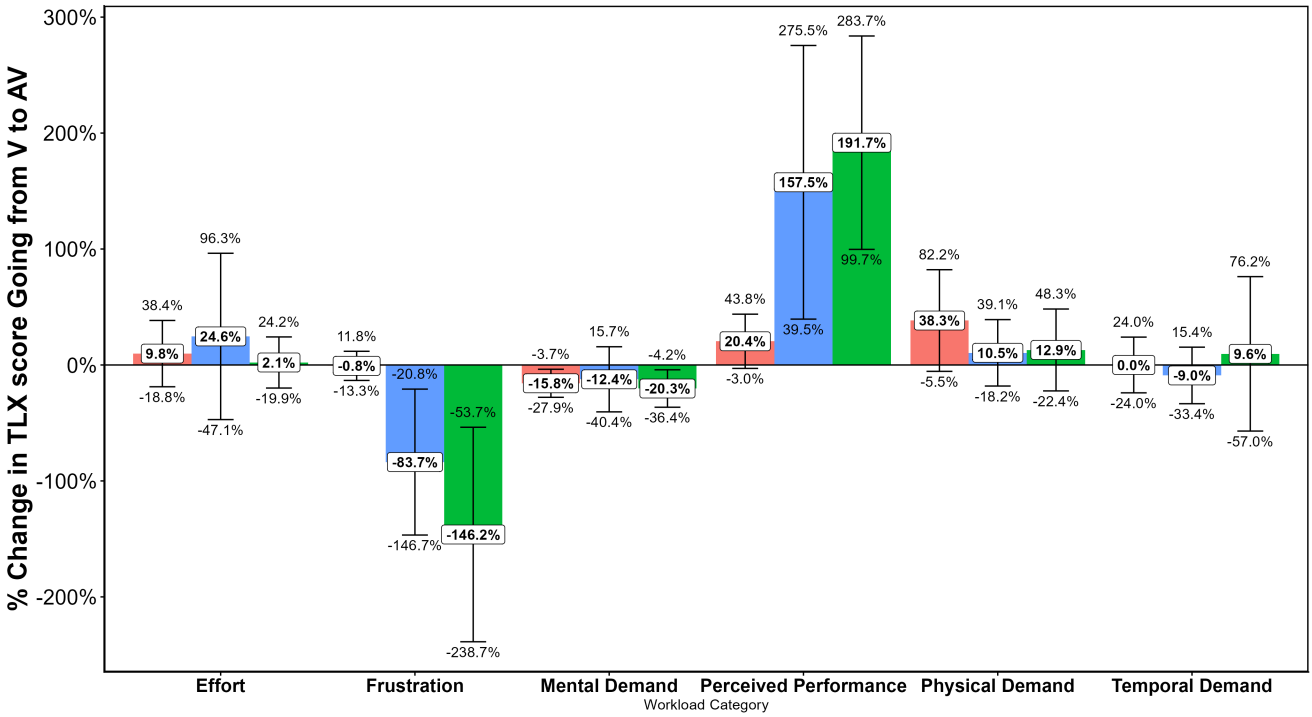


Figure 16: Percent change in workload for V→A+V (± 95% CI).