Sound Judgment: Auditory – but not Visual – Information Reveals Musical Competition Winners

Author: Daniel A Scannell

Boston College Electronic Thesis or Dissertation, 2014

Sound Judgment:

Auditory – but not Visual – Information Reveals Musical Competition Winners

Dan Scannell

Senior Honors Thesis

Department of Psychology

Boston College

April 15, 2014

Advisor: Ellen Winner

Second Reader: Jeffrey Lamoureux

Abstract

Previous research reported that people can successfully determine the winner of a musical competition when viewing a six second film clip of the performer without sound (Tsay, 2013, 2014); in contrast, when given an audio-only film clip or a clip that combined auditory and visual information, people perform at chance. Given the well-known publication bias in psychology (Ioannidis, 2005), this surprising and counterintuitive finding begs replication. In Study 1, 112 participants were randomly assigned to a sound, video, or video-plus-sound condition and were asked to select the winning musician after viewing five pairs of clips, one showing the winner and the other showing a non-winning musician. Clips were presented for 60 instead of six seconds, with the goal of giving participants more information about the performance, a modification we predicted would enhance performance in the audio and audio-visual conditions. Contrary to Tsay (2013), participants performed at chance in all three conditions. To more directly replicate Tsay (2013), in Study 2, 69 additional participants were randomly assigned to either a sound, video, or sound plus video condition and were asked to select the winning musician after viewing five pairs of 6-second clips showing the winner and another, non-winning musician. Here again the results did not replicate Tsay (2013): Participants performed significantly above chance in only one condition – when only hearing the performance and not seeing it. These results suggest that previous findings showing increased performance in rating musical performances without sound may be spurious and due to sampling error, issues in experimental design, low power, publication bias, or some combination of these. This also shows the strong importance of replication studies.

Sound judgment:

Auditory not visual information reveals musical competition winners.

Settling into a comfortable seat, you await the local symphony's performance of your favorite piano concerto. How do you judge the performance of this rendition? Do you focus on the tone the pianist is able to produce, or do you attend more to the movements of the pianist's hands? Is your experience of the performance affected by the pianist's visual cues, like walking casually up to the piano in a sweatshirt and sneakers instead of formal attire? What if she or he slouched or scowled?

Visual impressions may play a critical part in how motivated we are to listen to performances, all before a single tone is played. Platz and Kopiez (2013) asked participants to evaluate the behaviors of a musician walking on stage, preparing to play. When the actions of the musician were considered appropriate, participants were more highly motivated to hear the performance than when the actions of the musician were deemed unacceptable or inappropriate (Platz & Kopiez, 2013). Thus, perhaps our evaluative judgments of a musical performance depend not only on what our ears take in but also on what our eyes see.

One reason to believe in the importance of visual information when experiencing music comes from people's motivation to flock to live performances. If music were a solely auditory experience, there would be no need for live concerts. People would be just as content listening to recordings on their mp3 players or on the radio as they would seeing their favorite artists at a local venue. But people crave the opportunity to see performers in person, perhaps due to the importance of visual stimuli in music performance (Davidson, 1993; Platz & Kopiez, 2012; Schutz, 2008). One simple explanation is that people enjoy the physical presence of celebrities, but it might also be because we feel that being able to see the artist perform enhances the experience (Auslander, 1999, p.73). "Seeing is believing" in popular music, such that we want to authenticate the musical experience in person, and we may be drawn to live performance just to see the artists for ourselves and authenticate the musical experience. Many musical performances have become more than just the music, with choreographed dance, light shows, and skits performed on stage. Witnessing a performance may instead be more immersive than listening alone, allowing for the performers to influence, communicate with, and persuade the audience into enjoying the music more than would be possible without this social engagement (Platz & Kopiez, 2012).

We are influenced by what we see. Indeed, seeing the emotion of the performer reveals much more than the music alone (Bergeron and Lopes, 2009; Davidson, 1993). In Davidson's 1993 experiment, participants were shown video-only, sound-only, or video-and-sound of solo performances. Only the movement of the performer could be detected in visual stimuli. When participants were given visual stimuli, they deemed the performance to have more expressivity and were able to distinguish between deadpan, standard, and exaggerated performances. Clearly, people can detect much about a musical performance by seeing it. Furthermore, previous studies looking at both musical students and renowned musicians have shown that visual information does influence judgment pertaining to the quality of a performance (Schultz, 2008). Wapnik et al. (1998) showed that participants viewing video-plus-sound clips rated performances as better than when viewing sound-only clips. The question of whether visual or auditory stimuli play a more important role in judging musical performances thus emerges.

While we may enjoy the visual experience of a live performance and are strongly influenced by our visual world, this does not necessarily mean that we use visual information, consciously or unconsciously, to evaluate the relative quality of the performance in competitions. Tsay (2013, 2014) conducted a series of studies in order to investigate the relative power of sound versus sight when people are asked to judge a musical performance.

In Experiment 1, Tsay (2013) asked 106 novice musician participants the importance of audio and visual cues when judging a competition. Participants were told that they could win a cash bonus if they could correctly identify the winner of a musical competition. They could choose to be given recordings with audio, sound, or both audio and sound. They were told that a "tax" would be placed on choosing the video and sound recording. Participants were significantly more likely to choose audio over the other two options, and even choose the taxed video and sound significantly more than video only recordings.

Experiment 2 presented 106 novice participants with video-only and sound-only six second clips of the top three finalists from 10 classical music competitions (Tsay, 2013). The task was to pick the winner out of the group of three finalists. Again, participants predicted that the audio only clips would be more helpful than the video only clips. However, results showed that participants were significantly more likely to choose the winner when presented with the video-only stimuli.

Experiment 3 expanded upon Experiment 2 by adding a video-plus-sound condition (Tsay, 2013), in which 185 novice participants were either presented 10 groups of three clips of video-only, sound-only, or video-plus-sound for six seconds in length. Consistent with the results of Experiment 2, participants were significantly above chance at choosing the winner when viewing video-only clips, significantly below chance when listening to sound-only clips, and at chance when viewing the video-plus-sound clips.

Experiments 4 and 5 tested expert musicians under the same conditions as Experiments 2 and 3 (Tsay, 2013). When polled before the start of Experiment 4, 35 expert musicians shared the belief with their novice peers that sound was more important than sight when judging a musical performance. However, these same expert musicians proved able to choose the winner significantly above chance only in the video-only condition and not the sound-only condition. In experiment 5, 106 expert musicians were above chance at choosing the winner with video-only, significantly below chance for sound-only, and at chance for video-plus-sound.

Experiments 6 and 7 examined movements and gestures in relation to the visual cues of musical performance (Tsay, 2013). In Experiment 6, 89 participants of varying degrees of musicality were shown six-second clips that had been distilled down to just the outlines of motion. When viewing these clips, experts were significantly above chance at choosing the winner. In Experiment 7, 262 participants were presented either the sound-only or video-only clips, again in 10 groups of three (winner and top two finalists) and asked to choose which clip out of the three per group was most confident, creative, involved, motivated, passionate, and unique performer. When presented video-only clips, participants chose the actual winner for each of these six adjectives significantly above chance. When presented audio-only clips, they performed at chance.

In supplementary information, Tsay (2013) mentions the following in regards to clip-length:

> "In supplemental tests of the primacy of visual cues, additional studies featuring the same between-subjects design as experiments 3 and 5 replicate the findings outlined in this paper with 3-s and 1-s recordings. The at-chance findings with sound-only and video-plus-sound recordings remain even with longer time intervals ranging up to 60-s recordings. These results suggest that the findings outlined in the current experiments remain meaningful for more extended periods of evaluation" (Tsay, 2013; pp. 3)

Nothing, however, is mentioned about significantly above chance findings for video-only clips shown for 60 seconds. No actual data for these supplemental tests were presented. Despite repeated attempts to contact Tsay, I was unsuccessful in obtaining these data.

A more recent study by Tsay (2014) employed the same testing procedure but in this case participants were asked to judge the winners of orchestral competitions and differentiate between ranked orchestras and their non-ranked counterparts. By viewing but not hearing short clips of the entire orchestra playing, or just the group leader, experts and non-experts alike were able to choose the winners of orchestral competitions and differentiate professionally ranked orchestras from their non-ranked peers when given six-second clips.

Need for study

Tsay's (2013, 2014) findings are very surprising. Why would even expert musicians fail to distinguish between winning performances of major piano competition when given the competition audio? These results could be explained by the lack of information conveyed in such a short clip or that participants are not privy to how musicians are judged in these competitions. Perhaps, the relatively smaller sample of expert musicians led to these results. It is also quite possible that this is an impossible task, regardless of musicality, and these findings are a statistical anomaly.

Whenever findings are this surprising, they need to be replicated. Many times, claims made by one researcher can be refuted by another (Ioannidis, 2005). Furthermore, much of the methodology and analysis currently in use can be manipulated to generate false-positive results (Simmons et al., 2011). Many studies also suffer from creating hypotheses post-hoc (Kerr, 1998). Such techniques and possible experimental flaws dictate that replication is a necessary component to ensure the validity of published results.

The goal of the studies reported here was to attempt to replicate the surprising findings reported by Tsay (2013). We made four alterations to the method. First, we showed participants clips for 60 rather than six seconds, to make the task easier, potentially helping subjects to make more informed decisions about performance quality. Second, while Tsay asked people to select the winner from a set of three performances, we used a set of two, to minimize any errors due simply to the constraints of human working memory (Cowan, 2008). Third, each clip in the pair was played twice, in alternating format (i.e., clip 1, clip 2, clip 1, clip 2), so that participants would have more exposure to the two performances. Fourth, instead of choosing among only semifinalists, participants chose between the winner and a player who did not make it to the final round, thus increasing the contrast between the two performances.

Additionally, Tsay (2013) tested whether musically trained versus untrained participants showed any differences in their performance assessments. To attempt to replicate this finding, we assessed participants' musical ability using the Musical Ear Test (MET; Wallentin et al., 2010), a measure of musical ability through perception of melodic and rhythmic phrases. Should a population difference exist in performance between musicians and nonmusicians, this continuous measure may have more explanatory power than a binary metric of musical experience. We predicted that musicality, as measured by the MET, would predict higher scores in differentiating winners from losers in piano competitions.

Study 1

Method

Participants. One hundred twelve Boston College undergraduates (44 males, 68 females) completed the study for course credit.

Materials.

Competition Video. Ten videos of competitors performing in international piano competitions were used. These videos were obtained from YouTube.com from the Van Cliburn International Piano Competition, the International Franz Liszt Piano Competition, and the San Marino Piano Competition; three competitions utilized by Tsay in her 2013 study. These videos were presented in 360p resolution on the participant's personal computer utilizing QuickTime Player within the participants' browsers. Each video clip was 60 seconds in length. Pairs of clips were matched for similar camera angles and musical qualities. Each pair consisted of two players in the same competition. One clip showed the winner playing at the most advanced, solo stage of the competition available online, and the other clip showed the non-winner during either the same or lower round, depending on video availability. The members of each pair were similar in terms of pace and technical difficulty conveyed. See Appendix 1 for details on clips.

Competition Audio. Twelve audio-only recordings of 128kbps quality were extracted from the competition video for each of the corresponding videos. Each audio clip was identical in length and content to audio component of the video clips.

The Musical Ear Test. The Musical Ear Test is used to assess musical ability and takes approximately 20 minutes to complete (Wallentin et al., 2010). The test consists of 52 pairs of audio containing melodic phrases and 52 pairs of audio containing rhythmic phrases. For each pair of audio clips, the participant is asked whether the pair of phrases are the same or different. Scores on this test can range from 0 – 104. This test correlates with musical expertise and can be utilized to clearly distinguish between different levels of musicianship.

Procedure. Participants provided written consent to participate in the study and were free to withdraw at any point during the study. The study was conducted online via Qualtrics. Participants were randomly assigned to one of three conditions: sound-only, video-only, or video-plus-sound, with the built-in randomization function of Qualtrics. Participants first took the MET, using the instructions and materials provided by the test authors (Wallentin et al., 2010) The audio instructions explained that the test consisted of 52 pairs of melodic phrases, followed by 52 pairs of rhythmic phrases. The instructions explained that the task of the participant was to listen to the pair of clips and then choose whether the pair of clips was the same or different. Two examples for pairs of melodic phrases were given before the test. Two examples for pairs of rhythmic phrases were given before the rhythmic section of the test. Participants listened to the pairs of sound clips on their personal computer. They then marked their answers for each pair via online prompt in the survey.

Next, participants were told that they would be listening/watching/listening and watching to pairs of clips of pianists competing in international piano competitions. The following instructions were presented to the participant: “You will be presented with pairs of audiovisual/visual/audio clips. Each pair of clips will be presented twice. Each clip presents a pianist performing in an international piano competition. One member of each pair went on to win the competition. After each pair has been presented twice, choose which pianist you think won the competition.”

In the sound-only condition, participants were presented only the audio from each clip of competition. In the video-only condition, participants were presented only the video without audio for each clip. In the video-plus-sound condition, participants were presented with the video and audio together for each clip. The presentation of winning and non-winning performers was randomly sorted, with the winner appearing first two times and the non-winner appearing first for the other three. After presentation of a pair of stimuli two times, participants were then asked via online prompt which of the two clips was performed by the winner of the competition.

Participants watched and listened to timed clips via Qualtrics survey on their personal computers.. Each pair of clips was presented in a random order that did not vary between participants.

Results

Participants received a score of 1 for each correct choice, and scores could therefore range from 0 to 5. A two-way analysis of variance (Condition by Gender) was performed on correctness scores. There was no main effect of Condition ($F(2, 111) = .137$, $p = .872$) or Gender ($F(1, 111) = .400$, $p = .671$); nor was there a Condition by Gender interaction ($F(2, 111) = .1.373$, $p = .319$). We then performed one-sample t-tests against chance within each condition to determine whether performance exceeded chance. Using a Bonferroni correction, alpha level was set at .016 because three separate *t*-tests were performed. Participants were collapsed across gender since gender did not prove significant in the analysis of variance. Performance was strongest in the sound-only condition, but in no condition was performance significantly higher than chance (sound-only: $t(43) = 1.73$, $p = .091$; video-plus-sound: $t(38) = .71$, $p = .483$; video-only: $t(28) = .924$, $p = .363$). Results are visualized in Figure 1.

To determine whether a particular pair influenced the overall result, we conducted an item analysis. Performance on each pair was tested against chance (50%), adjusting the alpha level to .016 after a Bonferroni correction for multiple comparisons. While there were minor differences in performance across the pairs, results were comparable to the main analyses (see Table 1).

The Musical Ear Test (MET) was calculated from 0-104, with scores ranging from 45-98. Participants received a score of 1 for each correct answer. An incorrect answer was given 0 points. MET scores were then regressed onto judgment scores for each condition. None of the three regression analyses yielded significant results; sound-only: $R^2 = .001$, $F(1, 42) = .050$, $p = .825$; video-plus-sound, $R^2 = .028$, $F(1, 37) = 1.083$, $p = .305$; video-only $R^2 = .115$, $F(1, 27) = 3.514$, $p = .072$. Thus, performance on the MET was unrelated to ability to determine the winner in each pair of clips.

Discussion

Results from Study 1 indicate that under all three conditions, participants were unable to significantly choose the winner in a pair of two competitors in international music competitions. Lengthening the clips from six to 60 seconds was presumed to make the task easier, but it is possible that this may have instead have made the task more difficult. Perhaps participants thought too hard about the choice and could not come to a conclusion and guessed instead. Perhaps relying on their immediate, intuitive reaction would have resulted in greater accuracy. Of course, judges in actual panels take a great deal of time. But while members of a panel can deliberate with their peers, the isolated individual does not have this luxury. When making a decision in isolation, perhaps the immediate gut response is more accurate than one made after thinking about the pairs for a minute.

Results indicating that the MET failed to predict higher test scores replicate findings from Tsay (2013). Expert musicians and novices in Tsay's original study did not differ in results, as shown in experiments 3 and 5. These results may not be surprising when the amount of time participants have to make a judgment is considered. The lack of thoughtful deliberation and dedicated time may make the task equally challenging for both expert musicians and novices, regardless of the length of clip being presented, whether six or 60 seconds.

We failed to replicate Tsay (2013, exps 3, 5) utilizing 60-second clips. This extra time, in comparison to Tsay (2013) may have confounded these results. Another study was conducted to better replicate the original study. Therefore, in Study 2, we used clips that were only six seconds long.

Study 2

Study 2 was designed to once again attempt to replicate the findings of Tsay (2013), Studies 3 and 5, but this time using only six-second clips, as in Tsay's original study. Since Tsay (2013) did not find any difference in performance between musically trained and untrained participants, and because in Study 1 we showed that scores on the MET did not predict performance, in Study 2 we did not administer any measure of musical ability.

Tsay's six-second clips were chosen based on the theory of "thin-slices" (Ambady & Rosenthal, 1992). The theory of thin-slices claims that much information can be gathered in a relatively short amount of time. A 1993 study by Ambady and Rosenthal limited stimuli to as short as six-second clips while still retaining the same data as much longer samples (Ambady & Rosenthal, 1993). When evaluating professors' performances, participants gave the same answers after just a six-second silent clip of the professors teaching on the first day as they would after taking the professors' courses for the entire semester. In the aforementioned case, the studies looked at judgments of behavior, not the quality of a musical performance. Given these findings, however, the possibility of thin-slice effects occurring with six-second clips seemed plausible.

Participants

Sixty-nine Boston College undergraduates (9 males, 60 females) completed the study for course credit. This study is ongoing; therefore, all data presented are preliminary.

Materials and Procedure

The first six seconds of the 60-second clips used in Study 1 were used as stimuli. The same procedure was carried out as in Study 1, with the exception of the removal of the MET. Participants were given the judgment task after giving consent to participate in the study and answering a few demographic questions. As in Study 1, participants were randomly assigned into the three conditions: sound-only, video-plus-sound, or video only.

Results

There was no effect of Gender on Test Score, as analyzed by a one-way ANOVA, $F(1, 67) = 1.878$, $p = .175$. A one-sample t-test against chance (set at 2.5) was performed for each condition. Again, alpha was sent at .016. Analysis of t-tests showed participants to be significantly above chance in the sound-only condition: $t\ (26) = 5.453$, $p < .001$, $d = 1.05$. Participants were at chance in the video-plus-sound condition: $t(18) = 1.966$, $p = .065$, $d = .45$. Performance in the video-only condition was at chance: $t(22) = -1.358$, $p = .188$, $d = .28$. These findings are shown in Figure 2.

To determine whether an outlier pair influenced these results, the same t-test was performed for each pair separately within condition, with chance set at .5. As before, alpha level was adjusted to .016 utilizing a Bonferroni correction In the sound-only condition, only pairs 2 and 3 showed above chance performance, Pair 2: $t(26) = 6.310$, $p < .001$, $d = 1.21$; Pair 3: $t(26) = 2.801$, $p = .009$, $d = .54$. In the video-plus-sound condition, only Pair 3 showed above chance performance: $t(18) = 3.980$, $p =.001$, $d = .91$. In the video-only condition, no pairs were significantly above chance, and one pair (#4) was significantly below chance: $t(22) = -3.214$, $p =.004$, $d = .67$ Table 2 lists the results of the t-tests for each pair.

Discussion

Results from Study 2 diverge sharply from the results of experiments 3 and 5 of Tsay's 2013 study. While Tsay reported that participants were significantly above chance at choosing the winner when presented video-only clips, the present study found participants to be at chance with video-only clips. Similarly, while Tsay reported participants to be significantly below chance at choosing the winner when presented sound-only clips, the present study found participants to be above chance. Both Study 1 and Study 2 presented here showed that when presented with video-plus-sound clips, participants were at chance at choosing the winner.

Thin-slice clips allowed for participants to choose the winning musician above chance in the sound-only condition. These short clips may be more suitable for the survey setting in which the study took place. By only allowing participants a short glimpse into the performances of these competitors, an informed decision could be made about the clip, given the relatively short amount of time participants utilized before making a decision. That these short clips did not allow for participants to make the proper choice above chance in the video-plus-sound or video-only conditions shows the importance of sound when judging a musical competition. When sight becomes a factor, participants may be given too much information in the video-plus-sound condition, leading to an inability to successfully choose the winners significantly above chance. In the video-only condition, the proper information was not conveyed to the participants in order to make a proper decision.

Pair 3 highlights the differences in conditions well. For both the sound-only and video-plus-sound conditions, participants were able to properly choose the winner significantly above chance. Once the audio component of the pair 3 clips was taken away, yielding the video-only condition, participants were no longer able to differentiate between the winning and losing musician. Where looks alone were unable to convey the ability of the musicians, sound was able to do so, reflected by the participants' ability to choose the winner in the two conditions that include sound.

General Discussion

The results of Studies 1 and 2 contrast with the data reported by Tsay (2013). Where Tsay found sight to trump sound when judging the winners of a professional music competition, the current study found sound to be more important. Study 1 failed to replicate Tsay's findings that when sound-only and video-plus-sound clips are lengthened to 60 seconds, participants were unable to choose the winning competitor significantly above chance. Added to this finding, however, is that the same result is noted when video-only clips are lengthened to 60 seconds as well. Utilization of the MET aligned with Tsay's results, in which experts musicians and novices were equally as likely to choose the winners in all three conditions.

After failing to replicate Tsay's results using 60-second clips, six-second clips were used to replicate the original experiments more exactly.. Results from this second study indicate the opposite of Tsay's findings. Of course, it is possible that the divergence in findings could be due to not using the same clips as Tsay used. Unfortunately we could not use the same clips as Tsay used since information on which clips she used is not available. The critical next step would be to redo the study using the same clips as Tsay, in the same triads (winner plus two runner ups).

As mentioned earlier, Tsay (2013) states that when clips were extended from six seconds to 60, and participants were assigned either to the video-plus-sound or sound-only condition, participants were unable to choose the winning competitor. But we are not told what occurred when video-only clips were extended from six seconds to 60. In the current study, it was found that participants were unable to guess the winner when given 60-second clips. It is important to determine whether or not this was a replication of Tsay's results, or a new finding all together.

Evidence for sound as the more important factor can be seen in screening techniques now used when holding try-outs for premier orchestras (Golden and Rouse, 2000). When impartial judgments are made using a screen, sound dominates, since no visual information is presented to the judging panel. Due to this screening technique, significantly more women have had a chance to try-out, make it to later rounds of the hiring process, and become hired in major symphony orchestras. What can be gleaned from this process is that visual information can interfere with our ability to make the proper judgments in musical performance. Our biases seep into our judgments, and the top performer does not always get the job. Not surprisingly then, male musicians dominate in musical competitions. Study 1 and Study 2 of the current investigation reinforce this point, as nine of the 10 clips presented to participants involved male competitors. When conducting a search for appropriate clips for the current studies, hardly any clips of female competitors could be located.

Further research questions are raised by the two current studies. Though thin-slicing was able to provide us with significant data using minimal time on the part of researcher and participant, the survey settings in which this study took place share almost no resemblance to actual musical competitions. It would be nearly impossible to subject participants to realistic competition settings, however some steps may be taken to better simulate the event. When presenting participants with longer musical excerpts, it may be worthwhile to have the participants reflect after each clip is shown, along with creating a reasonable amount of time for which the participant must deliberate between the two clips before answers which clip presented the winning musician. In this way, it may be possible to subject participants with more information from the musical competition while still receiving significant results. Finally, given Golden and Rouse's (2000) study, pitting male versus female contestants against each other in pairs of clips would reveal whether this gender bias is significant even when presented clips of only six seconds.

Acknowledgments

I thank Ellen Winner for her constant support as my thesis advisor, evaluating and critiquing my progress every step of the way, along with pushing me to strive for perfection. I also thank Samuel Mehr for his collaboration, including his insight into the analysis of data and invaluable comments while editing the final draft.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consiquences: A meta-analysis. *Psychological Bulletin, 111*(2), 256.

Ambady, N., & Rosenthal, R.(1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431-441. doi:10.1037/0022-3514.64.3.431

Auslander, P. (1999). Seeing is believing. In P. Auslander (Ed.), *Liveness: Performance in a mediatized culture* (1st ed., pp. 73). New York: Routlledge.

Bergeron, V., & Lopes D. M. (2009). Hearing and seeing musical expression. *Philosophy and Phenomenological Research, 78*(1), 1-16. doi:10.1111/j.1933-1592.2008.00230.x

Cowan, N.Chapter 20 what are the differences between long-term, short-term, and working memory? *Progress in brain research* (pp. 323-338) Elsevier. doi:http://dx.doi.org/10.1016/S0079-6123(07)00020-9

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*(2), 103-113. doi:10.1177/030573569302100201

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review, 90*(4), 715-741. doi:10.2307/117305

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), 696. doi:10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196.

Krumhansl, C L. (2010). Plink: "thin slices" of music. *Music Perception: An Interdisciplinary Journal, 27*(5), 337-354. doi:10.1525/mp.2010.27.5.337

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal, 30*(1), 71-83. doi:10.1525/mp.2012.30.1.71

Platz, F., & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musicae Scientiae, 17*(2), 167-197. doi:10.1177/1029864913486369

Simmons, J. P., Leif, D. N., & Simonsohn, U. (0410). *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant*

Schutz, M. (2008). Seeing music? What musicians need to know about vision. *Empirical Musicology Review, 3*(3), 83.

Tsay, C. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences,* doi:10.1073/pnas.1221454110

Tsay, C. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes, 124*(1), 24-33. doi:http://dx.doi.org/10.1016/j.obhdp.2013.10.003

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The musical ear test, a new reliable test for measuring musical competence. *Learning and Individual Differences, 20*(3), 188-196. doi:http://dx.doi.org/10.1016/j.lindif.2010.02.004

Wapnick, J., Mazza, J. K., & Darrow, A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education, 46*(4), 510-521. doi:10.2307/3345347

Appendix 1

Information on Clips.

Pair 1 presented clips from the 2009 Van Cliburn International Piano Competition, with each male contestant playing pieces of moderate pace, and with identical camera angles throughout the clips. Pair 2 presented clips from the 2007 Franz Liszt Piano Competition, with one female and one male contestant playing pieces of high pace, and with identical camera angles throughout the clips. Pair 3 presented clips from the 2010 San Marino Piano Competition, with each male contestant playing pieces of slow to moderate pace, and with similarly alternating camera angles throughout the clips. Pair 4 presented clips from the 2013 Van Cliburn International Piano Competition, with each male contestant playing pieces of slow pace, and with identical camera angles throughout the clips. Pair 5 presented clips from the 2011 Franz Liszt Piano Competition, with each male contestant playing identical pieces of slow to moderate pace, and with similarly alternating camera angles throughout the clips.

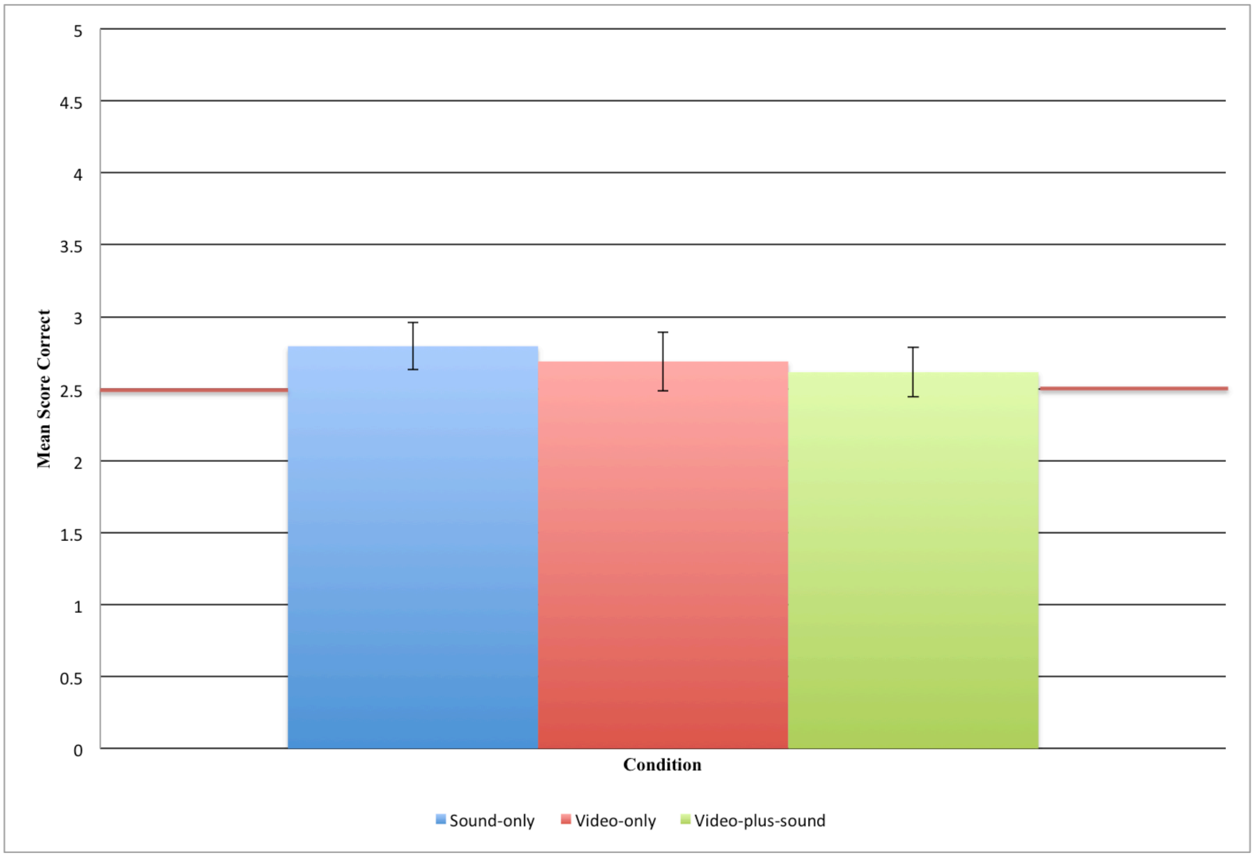
Figure Captions

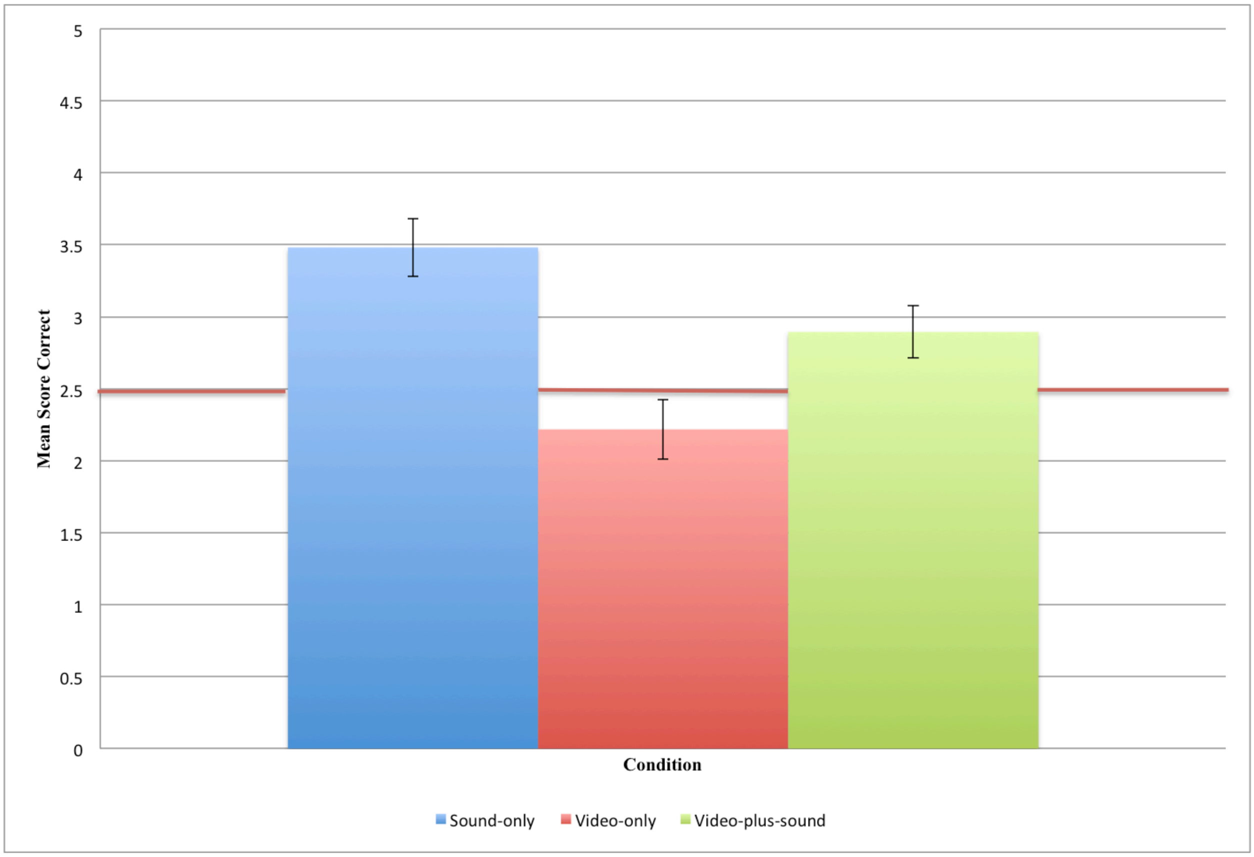
Figure 1. Mean Correct Scores Out of 5 (with chance at 2.5), Study 1

Figure 2. Mean Correct Scores Out of 5 (with chance at 2.5), Study 2

Table 1. List of Individual *t* - Tests Against Chance, Study 1

Table 2. List of *t* - Tests Against Chance, Study 2


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Mean Score Correct
Condition
Sound-only
Video-only
Video-plus-sound


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Mean Score Correct
Condition
Sound-only
Video-only
Video-plus-sound

Condition	Pair	dF	*t* - Score	Significance
Sound-only	1	43	-1.212	.232
	2	43	-.298	.767
	3	43	.599	.553
	4	43	1.859	.070
	5	43	.599	.553
Video-plus-sound	1	38	-.158	.875
	2	38	1.462	.152
	3	38	1.462	.152
	4	38	-1.125	.268
	5	38	-.158	.875
Video-only	1	28	-.183	.856
	2	28	1.727	.095
	3	28	-.183	.856
	4	28	-.926	.362
	5	28	1.727	.095

Condition	Pair	dF	*t* - Score	Significance
Sound-only	1	26	1.369	.183
	2	26	6.310	< .001
	3	26	2.810	.009
	4	26	.574	.574
	5	26	1.803	.083
Video-plus-sound	1	18	-.224	.826
	2	18	1.681	.110
	3	18	3.980	.001
	4	18	-.224	.826
	5	18	-.678	.506
Video-only	1	22	-1.994	.059
	2	22	1.994	.059
	3	22	-.204	.840
	4	22	-3.214	.004
	5	22	.204	.840