Learning to Play a Song with Online Videos: Strategies and Challenges of Musical Instrument Learners

ANONYMOUS AUTHOR(S)

Learning to play a musical instrument from online public videos is becoming increasingly popular as they offer easy access to expert tutorials and diverse instruction styles. Despite its popularity, video-based music learning is limited due to uneven quality of instruction in videos and absence of adaptive guidance. However, its exact nature is still understudied. We investigated the learning process, needs, and challenges of video-based music learners by conducting a diary study with 20 guitar and piano learners. Based on the analysis of diaries, interviews, and observation notes, we present (1) a model of video-based music learning process, highlighting the importance of metacognitive strategies for practice planning and resource utilization; (2) the criteria for assessing the instructional quality of videos, grounded on what learners found to be conducive to their learning; (3) the learners' challenges in finding and utilizing learning resources. We conclude by discussing implications and generalizability of our findings.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *E-learning*; Sound and music computing.

Additional Key Words and Phrases: video-based learning, music learning, instructional quality, self-regulated learning

ACM Reference Format:
Anonymous Author(s). 2021. Learning to Play a Song with Online Videos: Strategies and Challenges of Musical Instrument Learners. In *CHI '21: CHI Conference on Human Factors in Computing Systems Proceedings, May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Playing musical instruments is a popular activity [1, 2]. As a leisure activity, playing an instrument provides the benefits of the learner's mental wellbeing and offers hedonic values [61, 63]. Especially in situations such as the COVID-19 pandemic, where the range of social and leisure activities is limited under social isolation orders, playing a musical instrument is a meaningful way to spend time, as it offers self-development [26] and helps reduce anxiety and depression [33]. However, learning to play an instrument has unique challenges, as it involves training fine motor skills and practicing the sophisticated coordination of cognitive and physical activities [60]. Effective musical instruction is a sophisticated process that involves careful selection of resources, planning learning approaches, and the ability to integrate new information with prior knowledge [58].

Meanwhile, online videos are considered a core educational resource for independent musicians and hobbyists [38], as they offer music learners informal, diverse, and accessible learning opportunities [11]. Music content on these platforms is extensively produced and consumed: user-generated music videos account for 15% of videos on YouTube [29], and, as of August 2020, a YouTube search of the keywords "music tutorial" produces 453 million results. Several prior studies have examined how people use videos to learn software [34] and language [47]. However, we know little

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
Manuscript submitted to ACM

about the exact components, strategies, and challenges of video-based music learning processes. This knowledge gap is one of the core barriers in offering proper supports for video-based music learners and providing video creators with the guidance for producing effective video content. In particular, video-based music learning could be problematic in two aspects: difficulty in self-regulating practices and the uneven quality of user-generated, instructional videos. In this paper, we focus on musical repertoire practices, i.e., *playing a song*, which means practicing with the goal of completing a performance [49], because we found that the majority of music instructional videos on YouTube are song instruction [79]. Regarding other types of practices, technique practices are closely connected to and realized through repertoire practice; creative practices have already been largely explored in the context of online courses [68] and virtual collaboration [9, 12].

We define video-based music learning as autonomous music learning with videos as the primary learning resource. Autonomous learning — learning on one's own — means that the learners receive no adaptive guidance or feedback. Thus, music learners are required to possess strong self-awareness of their skills to self-evaluate [7, 48]. This behavior is called metacognitive regulation, defined as the "active control over the cognitive processes engaged in learning" [42]. Research shows that college music students with more experience and higher levels of expertise usually demonstrate better metacognitive regulations in their individual practices [27]. However, for learners who start learning with public videos, this process becomes difficult because they do not have instructors to teach metacognitive regulations or provide suitable resources for them as a starting point [28].

Because the quality of video-based learning depends greatly on the quality of video resources [52], characterizing instructional quality of music-related videos is important for helping video creators produce more effective instructional videos. Guidelines not specific to music instruction exist for creating effective instructional videos, such as providing previews of tasks [70] and minding the "6-minute rule" [39]. For effective music instruction, we need to consider not only the characteristics of the resource [78] but also the learner's background [44, 65], preferred learning style, and expertise level [5]. Furthermore, because learners might utilize more than just music tutorial videos — for example, covers, live performances, and non-video supplementary resources — the types of videos [38] and the connections between different types of learning resources [11] are also factors that determine how beneficial a video is. This study adds to the literature what components and characteristics of music-related videos learners found to be conducive to their learning process.

In addition to improving the effectiveness of video-based music learning from the video creator's perspective, we also studied the needs and challenges that learners face in their learning process. Potential challenges include the limitations of general-purpose video navigation interfaces for music learning tasks and encumbrance when controlling the interface for video search and navigation [21, 54, 81]. Although there are a few video interfaces designed to support video-based instruction for "how-to" videos on general topics [18, 37], we have yet to find any support for locating suitable resources for music learning processes and semantic navigation in music-related videos.

In this paper, we conducted interviews and diary studies with 20 guitar and piano learners of various skill levels to answer three research questions: (RQ1) what are the learning process and resource utilization strategies for learning to play a song from videos? (RQ2) what characterizes the instructional quality of a video in learning to play a song? (RQ3) what are the needs and challenges that learners face in their learning process? We identified the importance of metacognitive strategies for practice planning and resource utilization and constructed a model of video-based music learning processes for learners with different levels of metacognitive skills. Based on what learners found to be conducive to their learning, we proposed five criteria for assessing the instructional quality of music-related videos.

Finally, we identified the learners' needs and challenges in setting up the practice environment, locating suitable resources, and effectively utilizing videos.

This study makes three core contributions:

- **C1.** A model of the video-based music learning process. focusing on metacognitive regulation strategies,
- **C2.** Guidelines for assessing the instructional quality of music-related videos, and
- **C3.** List of learners' needs and challenges in video-based music learning.

2 RELATED WORK

2.1 Video-based learning strategies

Public videos are widely used for autonomous learning. Online information seeking behaviors and strategies have been studied in the context of physical tasks or craft making tasks, specifically how learners use keywords in search queries, combine resources, and regulate their attitudes while searching and browsing [72]. In classrooms, learners engage with video resources by pausing/resuming, navigating to specific points in the videos, and taking notes [14]. Their engagements can be measured on MOOC (Massive Open Online Courses) as quantitative variables (e.g., dropout rate and interaction peaks); these variables are shown, statistically, to be affected by or correlated with video navigation behaviors and video contents [36].

Videos are a medium for online music instruction. Among the activities of pedagogical online music instruction, public videos are most beneficial for "seeking information and explanation" activities [11] but lacking the "immediate feedback via synchronous communication" between instructors and learners [76]. Among different platforms of public videos, YouTube is popular for its accessibility and abundance of contents [3]. It is a preferable resource for high school students to informally learn to play instruments [29].

Although we know the overall contexts in which videos are advantageous, we did not find literature that specifically studied the video-based instrumental music learning process and how different types of music-related videos could be utilized in that process. In this paper, we studied the strategies that learners used to learn a song, starting from picking the song to finishing it, using any types of public videos as the primary resource.

2.2 Self-regulation and metacognitive skills in music learning

Self-regulated learning is an integral part of autonomous learning. It requires the learner to control their "behaviors, motivation, affect, and cognition" and set their own goals by and for themselves [62]. It could be applied to learning in informal settings, such as from family members, museums, and mass media [10]. Specifically in video-based learning, self-regulation contributes to higher engagement and retention, and can be even more effective with interactivity [20]. Self-regulated learning is also highly correlated with the learner's enjoyment [23], which helps create intrinsic motivation in the learner as well as a sense of community in adult learners [46].

Three types of self-regulation are crucial for individual music practices: behavioral (regulation of practice methods), environmental (regulation of practice space), and covert (regulation of mental state) [7]. For music learners to self-regulate their practices effectively, their instructors need to help them develop metacognitive skills, such as how to structure practice time "based on practice task's complexity and difficulty" [28]. A learner's knowledge and experience are highly correlated with their metacognitive skills, in that advanced learners show stronger self-awareness in their own skills and more accurate analysis of the musical piece than novices [27].

Although self-regulation has been studied extensively as a supplemental component to formal musical instruction, we do not know how videos can be used in their practice sessions. In our study, we focused on how learners with different levels of metacognitive skills utilize videos in each step of their learning process differently. We also identified learners' challenges in self-regulating their practice sessions.

2.3 Guidelines for effective instructional videos

Several assessment frameworks have been proposed for evaluating the instructional quality of tutorial videos. Morain et al. proposed the well-known set of criteria for evaluating instructional videos, derived from YouTube user's ratings; the criteria consist of three categories: physical, cognitive, and affective [52]. Other criteria are based on other characteristics of videos, such as video lengths and user's comments [84]. For lecture videos on MOOC, learners' interaction patterns can indicate how engaging a video is [25]. The learners' rating is also a popular indicator of instructional quality; the sentimental and objective reasoning behind the rating [66], as well as the correlation between a learner's performance and their rating [65], have been extensively studied.

For music learning, a survey of "music education" videos on YouTube (both instructional and non-instructional) provided statistics on diversity, perceived participants and audience, genres, and other demographic information [79]. Further content analyses on the video comments and ratings revealed how users left comments under music educational videos on YouTube [78] and how the users' feedback to a video — the likes/dislikes counts, user's comments, etc. — affected a new audience's perception of the video's quality [80].

This paper extends the scope of videos from instructional videos to all types of music-related videos. We developed specialized video evaluation guidelines for music instruction based on what learners found to be conducive to their learning process. The guidelines consist of a list of essential components of instructional videos, the types of musical information that is best conveyed by each component, and the desirable characteristics that makes each component effective for instructional purposes.

2.4 HCI studies on video interfaces for learning

Design of video interfaces for learning is a well-trodden path in HCI research. We highlight three aspects of prior studies that are related to our findings. First is to remedy the isolated learning environment and the unidirectional information flow from video to learner. Video interfaces can use learnersourcing to structure the content and promote better learning [77], make tutorial interactive [55], or induce learners to create a social learning environment beyond videos [40]. Secondly, video interfaces can leverage semantic cues to facilitate the user's browsing and navigation. Such cues include textual contents [51], interaction data from learners [35], people [73], objects and actions [56], and transcripts [51, 85]. Thirdly, interfaces that tackle situational impairment, specifically encumbrance [54], can use alternative input modes such as voice [15] or gestures [67].

The existing music learning interfaces are mostly designed for keyboard learners and do not involve video resources. Their features include guiding the learner's hands with visual cues projected onto the instrument [71, 83], automatically adjusting difficulty level [87], and automatically evaluating the performance's correctness [64].

This paper adds fresh design implications to these bodies of work, with a keen focus on remedying difficulty in self-regulations, a deficit of resources, the musical instrument's constraints on device setup, and other problems in choosing and navigating videos on existing online platforms.

3 METHOD

We conducted a diary study to capture video-based learning experiences. Diary study has been employed for studying music learning process, strategies, and styles [13, 82], as it works well for capturing the learning process over time and for exploring various variables of the learning process [24]. Every participant had to complete seven practice sessions, each at least 30 minutes long, and filled out the diary template for the session. The participants were free to practice more than one session per day and could choose to skip practices for a day or more if necessary (e.g., they might not have access to the instruments or might be busy with work on some days.)

We supplemented the diary study by running an observational study to capture specific critical incidents in the practice sessions and what our participants actually did, rather than what they said they did. We observed one of the seven practice sessions, rather than observing all seven, because their practice settings and learning strategies tended to be similar across the sessions. Each observation was followed by a ∼ 30-minute retrospective interview, where we asked the participants to review the recorded video footage of their practice session and explain their behaviors.

We chose piano and guitar as the target musical instruments for two reasons. Firstly, they are the two most popular instruments in the world for both formal and informal instructions [16]. Secondly, they allowed us to cover various formats of tutorial videos, such as synthesia tutorials, lessons with verbal explanations, lessons with human demonstration, and animated sheet music.

3.1 Participants

We recruited 20 music learners (10 for each instrument) using snowball sampling through outreach at the regional music learning societies and personal contacts with acquaintances. The inclusion criteria include: age 19 year-old or older, prior experience with piano or guitar, but not a professional musician. We used a screening survey and purposive sampling to balance the number of learners per instrument and skill level. As suggested in Hallam's work [27], the skill level was categorized into beginner, intermediate, or advanced based on a combination of the participant's self-reported experience with the instrument and further refined based on their metacognitive regulation strategies demonstrated in the diary study (e.g., resource management, session planning, and self-evaluation) [19]. Details of participant demographic information and skill levels can be found in Appendix A.

In this paper, each participant is referred as $\mathrm{P}N^{superscript}_{subscript}$, where N is the assigned participant number, the superscript is their skill level (*bgn* for beginner, *int* for intermediate, *adv* for advanced), and the subscript is their instrument (*pn* for piano or *gt* for guitar).

3.2 Task and diary materials

During the diary study, the participants were instructed to learn target song(s) of their own choices. To promote a natural and stress-free music learning context, we did not set the number of songs they ought to finish. For example, if a chosen song did not have adequate online video resources or was too difficult for their current skill level, they were allowed to start over with another song. Yet, we asked them to finish at least one song, so that we could observe the complete learning process. For each session, the participants had to use at least one public video as the primary resource. The participants were allowed to also use other music learning resources as they saw fit.

After each practice session, the participants answered structured questions on a given template of an online diary. We asked the participants to record the general information about the practice session and evaluate the instructional quality of each video they watched during the session. The general information consisted of the length of their practice sessions,

goals, satisfaction, and self-perceived efficiency of the session. We also asked them to take a photo of their practice space to show how they set up devices and instruments. For each video, we asked the participants to provide the link to the video, a description of how they found the video, why they decided to use it, the helpful and unhelpful/improvable characteristics of the video, and how they utilized the video.

3.3 Procedure

Table 1. Our study procedure, consisting of four stages, with approximately 8 hours of total participation.

	Method	Task	Data collected	Time spent
Screening survey	Questionnaire	—	Participant's demographics, self-reported musical skills	5-10 minutes
Pre-study interview	Interview and observation	Video searching	Music learning experience and video searching strategies	30 minutes
Diary study	Electronic diary and observation with retrospective interview	Learn to play song(s)	Video-based music learning strategies and video's instructional quality	7 hours (7 sessions; each of approximately 30 minutes practicing and 30 minutes writing the diary)
Post-study interview	Interview	—	Reflection on the diary study	30 minutes

The overall procedure is shown in Table 1. Prior to the diary study, we asked each participant to fill out a screening survey and conducted a semi-structured interview with them to collect information about their musical experience. During the diary study, we observed one of the practice sessions (30-60 minutes) through a video call to gain detailed insights into the video searching and browsing tasks, video navigation, and other practice strategies. Finally, after the diary study, we conducted a post-study interview, where we asked the participant to reflect on their experience from 7 practice sessions of the diary study and elaborate on their challenges and needs. The participants were compensated at the rate $15 per hour (total $120 per participant.)

3.4 Data collection and analysis

We collected all data via online platforms, in accordance with the social distancing regulations in the COVID-19 situation. This included screening survey and online diaries via Qualtrics, and interviews and observations via Zoom. During the observation, the participant used the same device to watch the video and share their screen while video-calling on Zoom. Interviews and observation sessions were audio-recorded and transcribed for later analysis.

For data analysis, we performed a reflexive thematic analysis starting by in-vivo coding of the diary entries and interview transcripts. The lead researcher coded the transcripts and diaries, with another researcher reading through part of them and discussing potential themes identified based on the codes in multiple iterations. Then, we reviewed the codes, inductively, with the notes taken during the observation sessions to come up with 42 initial themes. We refined and organized our themes based on our research questions. Finally, we reviewed our final themes against the data to ascertain their accuracy and against the existing literature to identify novel findings.

4 FINDINGS

All participants finished the study procedure, except $P11^{bgn}_{pn}$, who withdrew after completing 6 practice sessions because they felt stressed about reporting their learning progress. Their data from the pre-study interview and practice sessions

was included in our analysis. We collected 139 diary entries from 107 hours and 20 minutes of music learning and practicing. The participants watched a total of 117 music related videos, covering 8 genres: song tutorial, music video & original recording, cover performance, lyric video, technique tutorial, musician's interview, music theory lesson, and finger exercise. We also collected 69 different search queries and 39 photos of practice environments. Every participant finished at least one song, although what they mean by finishing a song varied by individuals: some finished when they could remember the score; others finished when their performances were fluent.

By analyzing and triangulating multiple modes of data, we identified three major themes: learning strategies for video-based learning, criteria for assessing instructional quality of a video, and the learner's challenges. Although we conducted the study with piano and guitar, we found that many findings were common between the two instruments, which we will further discuss in section 5.1. The main differences between guitar and piano video-based learning in our data were only their tutorial formats and the specific playing techniques. For the rest of this section, all the findings apply to both instruments, unless otherwise noted.

4.1 Metacognitive strategies of video-based music learners

We discovered patterns of self-regulated learning strategies. The learners showed different levels of metacognitive skills in managing their resources, motivations, and cognitive strategies. Metacognition played a big role in how the learners structured their learning process (see section 4.1.1 on the process model, sections 4.1.2 and 4.1.5 on session-level strategies and section 4.1.4 on resource-utilization strategies.)

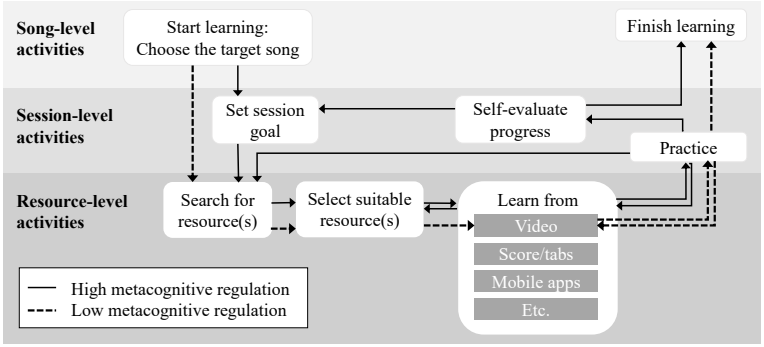


Fig. 1. A model of the video-based music learning process for learners with different levels of metacognitive skills. Each box represents an action taken in the process, and each arrow represents the chronological order in which the actions happen. Details of these differences are specified in section 4.1.

4.1.1 Model of video-based music learning. We identified a structured process of learning activities that our participants tended to follow. We organized the common patterns into a process model as shown in Fig. 1. These activities took place at three different levels depending on the relationship between the goal of the activity and the nature of the practice: song, session, and resource. We observed significant differences between the processes of learners with high vs. low metacognitive regulation (i.e., advanced vs. novice learners). Therefore, we structured these activities into two pathways. There are three important differences between the pathways.

- There are iterative steps in the learning process. For learners with low metacognition, this only happens between learning from a video and practicing, since they would adhere to one video to learn the whole song. For learners

with higher metacognition, the process is more complex as there are several possible alternatives at each step. This topic is discussed further in section 4.1.4.

- Learners with low metacognition watch only videos, while learners with high metacognition may utilize heterogeneous resources. Learners need experience and musical knowledge to specify compatible resources of different kinds and combine their contents together. Moreover, since learners with low metacognition usually watch only one video, the videos tend to be strictly tutorials, whereas learners with high metacognition may utilize various types of videos for diverse purposes. This topic is also discussed further in section 4.1.4.
- Goal-setting and self-evaluation activities are indicative of high metacognitive skills (see the two session-level activities in Fig. 1). This topic is discussed further in sections 4.1.2 and 4.1.5.

It should be noted that these activities occur on a spectrum of metacognitive levels. For example, intermediate learners exhibited varying mixtures of the activities in the two pathways.

4.1.2 Learners set goals for each practice session differently depending on their levels of metacognition. Beginners usually did not set goals for practice sessions (afterwards, they filled in their goals in the diary with what they have achieved in the session.) Most of the times, their goal was to "follow the video" ($P14^{bgn}_{pn}$). But such goals might not be suitable for their skills. As a result, they would unexpectedly get stuck and/or diverge from their goals. Moreover, there was no connection between the progress made in one session to the goal of the next session.

On the other hand, advanced players had more specific goals, which they set based on past progress, self-awareness of their skills, and accurate evaluation of the song's difficulty relative to their skills. Moreover, when advanced learners could not achieve their goals, they made plans to tackle that in the next session ($P3^{adv}_{pn}$, $P15^{adv}_{gt}$). This echoes findings from studies of formal music instruction: advanced music students' learning strategies are goal-oriented and their learning involves more self-awareness of their skills and utilizes more musical knowledge [27].

4.1.3 How music learners search videos: query formation and thumbnail-based visual filtering. We compiled the search queries that we collected from the search-and-browse tasks and the practice diaries. Then, we analyzed them for common patterns, which can be formulated as the syntax representation of query combinations as follow:

[(Optional) Song contexts] + [Song name] + [Instrument name]
+ [(Optional) "Tutorial"] + [(Optional) Descriptor(s) for target level of difficulty]

In this syntax, the song difficulty is evaluated based on the arrangement of the song that the learner intends to learn, which means it can be based on the original song or a cover version. For piano learners, the level of difficulty could be determined by "how many fingers to press" simultaneously ($P8^{bgn}_{pn}$) and "speed of the song" ($P20^{adv}_{pn}$). Most learners only used videos from the search results and did not look at the recommended videos at all. This is because they "know what [they are] looking for" ($P3^{adv}_{pn}$) with the search query, and because the recommended videos were usually "from same channel but different songs" ($P13^{bgn}_{pn}$). In general, the recommendations were more helpful for "finding new songs" to learn ($P5^{bgn}_{gt}$).

Most learners looked at thumbnails first when they chose whether to watch a video. A video's thumbnail provides a rough summary of the visual components in the video, such as the performer's hands and the annotation overlays. Therefore, it serves as the first filter of the search results. Some learners also looked for familiar faces on the thumbnails, e.g., a tutorial "from a channel [they have] watched before" ($P2^{bgn}_{gt}$); others looked at the font and background to determine "how professional the production is" ($P9^{bgn}_{gt}$). With a synthesia-style tutorial, the learners could also estimate

the level of difficulty by looking at how many bars are "falling at the same time" and the speed in which they are falling ($P19^{adv}_{pn}$).

On top of thumbnails, our learners also looked at other information to decide whether the video would be beneficial for their practice session. On the search results, the title indicates a tutorial's level of difficulty. View counts and upload years are often evaluated together as secondary factors to determine the popularity of the videos. Video lengths may also be a factor, as beginners tend to pick longer videos because that implies more detailed teaching ($P2^{bgn}_{gt}$). Based on these factors, the learners would pick a video to skim through. They may look for a "good performance" ($P3^{adv}_{pn}$) (the definition of a good performance can be found in Section 4.2.1) or more specific contents, e.g., "teach fingerpicking style" ($P7^{bgn}_{gt}$) or teach the chord they "can't figure out by ear" ($P17^{adv}_{gt}$). Learners only looked at video descriptions for links to outside resources, e.g., sheet music or tabs. Otherwise, they did not look because descriptions are "full of advertisements" ($P4^{int}_{pn}$).

4.1.4 Music learners employ diverse strategies to utilize videos and supplementary musical resources in their learning process. Beginners liked to watch a video in a linear order, following the steps outlined by the instructor; more advanced learners would skip to the part they want to learn first. For example, $P20^{adv}_{pn}$ wrote in her diary about the second video she used for learning her target song, "I actually skipped to the latter half of the piece, because I thought the arrangement of the first half in the 1st video was good enough." The more advanced a learner was, the less patient they were to listen to the things they "already know about" ($P1^{int}_{gt}$). This is partly because they had more specific goals of what they wanted to get out of a video and partly because they had a more complex way of structuring their learning, e.g., $P10^{int}_{gt}$ filled in the intro of his target song with chords, and then learned the accurate riff after he had finished learning everything else for the song.

Depending on the types of learning goals, the learners' skill levels, and resource availability, learners often used some combinations of the following three strategies for utilizing videos. A popular strategy is to imitate the hands or play according to the verbal explanation. This strategy allows the learners to learn the accurate finger placements and movements. Another strategy is to write down annotations on paper based on the instructor's hands or verbal explanation ($P10^{int}_{gt}$, $P11^{bgn}_{pn}$, $P14^{bgn}_{pn}$). This only applies to videos that (1) don't provide annotations, (2) provide only "one unit" of annotations at a time, or (3) have a unique arrangement that is incompatible with any annotations from outside sources. A similar strategy to this is to "pause the video" on a frame and "look at score" while playing repeatedly ($P7^{bgn}_{gt}$, $P9^{bgn}_{gt}$). The last strategy is performing along with the video. If there were many repetitions of the performance at different speeds, the learners could "let [the video play] without pausing" ($P2^{bgn}_{gt}$). If there were not enough rounds of demonstration, the learners would replay segments of the video and manually control the playback speed.

On top of these popular strategies, some advanced learners had other unique, less obvious ways of utilizing videos and other tools. For example, $P20^{adv}_{pn}$ watched a live performance to compare the song's key with the tutorial video she planned to use; $P17^{adv}_{gt}$ watched a video, which explained a specific techniques that he lacked to play the target song, and $P3^{adv}_{pn}$ watched a video to understand the harmonization style of a specific composer. Other tools that participants used include the OS audio driver for transposing audio output ($P20^{adv}_{pn}$), a mobile application for verifying alternate ways of playing a chord on guitar ($P18^{adv}_{gt}$), and a software for rendering MIDI performance from tabs ($P15^{adv}_{gt}$).

4.1.5 Advanced music learners evaluate their session performance based on prior knowledge and experiences. Based on prior knowledge and experiences, advanced learners were aware of their abilities and the factors preventing them from achieving their best performances. They considered these factors when self-evaluating ($P3^{adv}_{pn}$, $P18^{adv}_{gt}$). For satisfaction

Table 2. A summary of the guidelines for creating desirable music instruction videos. The left column is the guidelines for making effective music instruction videos (sections 4.2.1–4.2.5). The middle column is the type of instructional content that each guideline is associated with. The rightmost column describes the specific, desirable characteristics of instructional content as they relate to the guideline.

Guidelines	Target information	Desirable characteristics
Demonstrate an error-free performance	Timing, tempo, groove, overall execution	(If not the original artist) Accurate transcription, arrangements, and performance
Put the currently-played score on screen in segments	Score	The annotation should be on the screen all the time in segments (not just what is currently played). Its timing should match what the instructor is playing/explaining.
Show performer's hands clearly in familiar orientation	Score, fingering, hand and finger movements	A straight angle, clear lighting and focus, show both hands in an intuitive orientation
Consider the target learners' skill levels when explaining techniques	Playing techniques (including hand and finger movements), song analysis, practice tips	The level of details should correspond to the difficulty level indicated in the title.
Help learners manage their expectation in the video title and lesson introduction	Lesson overview, practice strategy	List topics to be covered and identify difficulties to signal the arrangement's complexity and to manage learners' expectation

rating, they considered whether they had achieved the goal, and for efficiency rating, they considered whether they achieve the goal under the time they expected to. $P17^{adv}_{gt}$ explained, "Sometimes, it'll take 15 minutes to learn something. That's very efficient. And other times it'll take an hour and a half. And I'm still learning what I want to learn. It's satisfying, but it's not very efficient."

Less advanced learners, on the other hand, evaluated their session performance based on their feelings, i.e., how enjoyable or frustrating the session was. For instance, $P6^{int}_{pn}$ explained that, when he started practicing, he "sometimes go on without actually knowing how much I could do." And when he evaluated himself to fill in the practice diary, it was based on feelings ("You can feel if you're frustrated about how you're playing.") Contrary to the advanced learners, they did not evaluate themselves with respect to the session goal.

4.2 Guidelines for assessing the instructional quality of music-related videos

Our learners found certain characteristics of videos to be conducive to their learning process. We distilled the data, grouped them by instructional contents, identified the desirable characteristics of the contents, and then developed them into prescriptive language of guidelines summarized in Table 2.

4.2.1 Demonstrate an error-free performance. In a tutorial, showing the instructor's demonstrative performance helps affirm the instructor's ability and sets the learners expectation about the difficulty of the performance they are going to learn. For example, $P17^{adv}_{gt}$ wrote in the diary that he chose a video because "[the instructor] plays the song well and plays with correct technique." In contrast, $P16^{int}_{pn}$ watched a video but ended up not using it because "[the instructor] ability to play through was bad, so I didn't trust her skill." A good performance should exactly correspond to the lesson and/or provided annotations and should comply to the original recording of the song.

4.2.2 Put the currently-played score on screen in segments. Annotation is a necessary reference for learners' practices before they can learn the score by heart. The on-screen annotation is more helpful for learning than a separate sheet music, since it is synchronized with the performance or explanations in the video. It also saves the learners from having to switch between different application windows, which is especially inconvenient on phones and tablets ($P13^{bgn}_{pn}$). However, for the on-screen annotation to be helpful, it should have context. That is, it should contain segment(s) of the music, e.g., a verse of the song, not just the notes that are currently played. If it contains only the currently played chord or note, it will not be usable when the learners pause the video or need to "see the next notes ahead of time" ($P6^{int}_{pn}$).

Providing annotation, either on-screen or separately published by the instructor, is also helpful when the learners has doubts about the correctness of the lesson; they can easily verify the provided annotation with other versions, e.g., from tab websites. If no annotation is provided, annotations from different sources may be of "different arrangements" ($P18^{adv}_{gt}$), making it difficult for learners to compare which part of the arrangement is the same or different, especially if the learners are novices or are not familiar with the song.

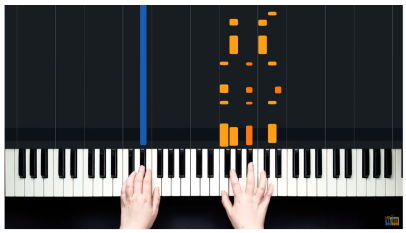
The suitable type of annotation for a video depends mainly on the learner's ability to interpret annotations and how they want to play.

- **Learner's ability:** beginners need extra information to translate from annotations to action. For beginner guitar learners, the chord charts should also be available for beginners, in addition to chord names, because beginners may "not know how to play [a chord] just by looking at the letters" ($P9^{bgn}_{gt}$). For beginner piano learners, pitch-leveled details of chords or the score should be written in letters, not in the 5-line staff format, since most beginners cannot read it.
- **How learners want to play:** $P20^{adv}_{pn}$, who finished learning one song for singing along and one as an instrumental performance, explained that "A music sheet is almost equivalent to trying to extract the exact notes from an accompaniment, but the chords are a lot more vague... If I'm seriously devoted to learning the song, so that I can perform it for people, I prefer having an exact arrangement that I can follow. I use the chords mostly when I just want to sing to myself." We observed this pattern in other learners who aimed to sing along ($P4^{int}_{pn}$, $P7^{bgn}_{gt}$) and who aimed for instrumental performances ($P10^{int}_{gt}$, $P19^{adv}_{pn}$) as well.

4.2.3 Show performer's hands clearly in familiar orientations. Although learning the score by watching hands is usually harder than by reading annotations, learners can approximate where the notes will be along the keyboard or the guitar neck in advance by watching the instructor's hand movements. With guitar, beginners watched the hand movement between chords a lot, trying to learn the technique of how to transition smoothly or make a clean sound (e.g., with the F chord.) With piano, many learners copied the fingering from the video. If the hands were not visible in the videos, it took longer for them to "figure out the fingering" ($P19^{adv}_{pn}$) by themselves; other learners simply did not think about fingering and changed it with every round they practice ($P16^{int}_{pn}$). The later strategy was less efficient, likely because using different fingering every time made it harder to develop muscle memory [50].

Both hands of the performer should be shown in a familiar orientation, which is different for guitar and piano, as shown in Fig. 2. Moreover, the lighting and camera focus should be clear in order to accurately show where each finger goes on the instrument. Bad examples, as shown in Fig. 3, are not suitable for learning finger placements. However, they may still be used for learning some aspects of finger movements, such as the rough chord shapes or the hand placements.

Guitar — front view, both hands are fully visible



Piano — top view, hands pointing upward

Fig. 2. Examples of desirable hand displays in guitar and piano tutorials

Dark lighting and out-of-view right hand is not visible

Out of focus and low resolution

Fig. 3. Examples of problematic hand displays

4.2.4 Consider the target learners' skill levels when explaining techniques. Since beginners liked to stick to one video and watch it from the beginning to the end linearly, they preferred videos with comprehensive explanations. If the title contains keywords indicating a low skill level, such as "easy" or "beginner", the video should feature detailed explanations about basic techniques, not assuming that the learners already know them. $P7^{bgn}_{gt}$ wrote in the diary: "I found the video by searching on Youtube. Keywords: 'can't help falling in love guitar beginner'… I chose this video because I thought it was a beginner level video... [the instructor] was talking way too fast for beginners to catch up." On the other hand, if the title does not contain "easy" or "beginner", the explanation should be concise and not cover the basic details, since listening to explanation takes away the time in which the learners could spend playing the instrument for leisure purposes. For example, $P1^{int}_{gt}$ commented that, for a video he used, "[the instructor] went into a lot of depth talking about how which fingers are supposed to go at which frets on the guitar. That's something that I know because I've already learned that ... If it was my ideal lesson, he would not talk about that."

4.2.5 Help learners manage their expectation in the video title and lesson introduction. The performance's complexity should be stated explicitly in the video's title because beginners like to include it in their search queries. In the lesson introduction, the instructor should list the techniques that will be needed for the song, identify difficulties, and list the exact contents they are going to teach. For example, $P2^{bgn}_{gt}$ wrote in the diary that it was helpful when "the instructor warned that the first riff is awkward. So, I managed my expectations." The learners may choose to learn a song as a mean to learn a technique used in the song, so if the instructor "assumes [the learners] already know" ($P7^{bgn}_{gt}$) it and does not cover it in their tutorial, they should make that assumption explicit. The instructor should also "play what they're going to teach you" ($P4^{int}_{gt}$), so that the learners would not waste their time learning something and discover later that they "don't like how it sounds as a whole" ($P16^{int}_{pn}$).

4.3 Challenges in video-based music learning

We identified five main challenges in video-based music learning related to the learner's metacognitive regulation, resource availability, practice environments, and two difficulties in using video interface. These challenges diminish the effectiveness of the learning process and reduce the pleasure learners get from the practice sessions. It should be noted that every participant chose YouTube to do this study, although we did not specify the platform in the instruction. We compared the features of YouTube interface presented in this section with three other popular platforms: Dailymotion, Vimeo, and Facebook. We found that their interfaces were composed of a subset of those features, implying that that our findings are comprehensive and generalizable.

4.3.1 Beginners have trouble combining different instructional resources in support for their video-based learning. Beginners preferred "all-in-one" ($P7_{gt}^{bgn}$) video that taught everything they needed to complete their target song, but such video did not always exist. When they started out learning from a video and found some contents to be missing or poorly explained, they usually kept pushing through with that video, trying to mimic the instructors as much as possible ($P9_{gt}^{bgn}$, $P13_{pn}^{bgn}$). They explained two main reasons for not looking for alternative or supplementary materials. The first reason is the idea of finding resources for specific problems "never occurs" to them, even when they could identify the problem ($P7_{gt}^{bgn}$, $P14_{pn}^{bgn}$), indicating the lack of metacognitive skills [27]. The second reason is different resources may not be compatible with each other. $P1_{gt}^{int}$ recounted a struggle from a practice session: "I went to look for tabs... And it was the same note as the teacher played. But it was very different from the recording that I had. And it took me a while to understand that the recording I had was not in the same key." $P9_{gt}^{bgn}$ also commented that using videos from different instructors was difficult because "the chords [the current instructor] taught was completely different than what I used to learn [from the previous video]."

4.3.2 Advanced learners experience a deficit of videos that suit their sophisticated goals. Instructional videos that featured an ideal arrangement for advanced learners did not always exist. Although it was possible with their skills to learn from non-instructional videos, it required greater effort and was very inefficient. Learners might try to learn by watching the original artist's hands in a live performance. However, for old or obscure songs, it was difficult to find a performance video that clearly showed hands, so the learners might try to figure the score out by ear ($P3_{pn}^{adv}$, $P17_{gt}^{adv}$). If the song was not meant for the instrument (guitar or piano), they might follow a cover artist, again by watching hands or by ear ($P19_{pn}^{adv}$, $P20_{pn}^{adv}$). For creative learners, a satisfying cover might not exist either; in this case, they might choose to develop their own arrangement of the song ($P3_{pn}^{adv}$).

Videos that teach advanced techniques are also lacking. We observed this in intermediate and advanced guitar learners ($P1_{gt}^{int}$, $P10_{gt}^{int}$, $P17_{gt}^{adv}$, $P18_{gt}^{adv}$), who had knowledge about musical styles and techniques. (We did not observe technique learning in any advanced piano learners during the diary study, but $P3_{pn}^{adv}$ mentioned having "watched masterclass videos to learn fingering techniques" before.) Because there were not many instructors who could teach the difficult techniques, the learners' only choice was to try imitating the performer and figure out the techniques by themselves ($P18_{gt}^{adv}$). Based on Whitaker et al.'s study, such "lesson videos" make up only 6% of music educational videos, and the most popular ones are for beginners [79], showing there is a significant lack of video resources of learning advanced techniques.

4.3.3 The device setup is constrained by the instrument's physical architecture and acoustic nature. Learners' practice environment setup and the video device's interaction mode were constrained by the need to mitigate the encumbrance problem [54] for each instrument, as shown in Fig. 4 (a, b). In general, learners preferred the screen to be as large as

(a) (b) (c)

Fig. 4. Examples of practice environments and device setups; (a) $P17^{adv}_{gt}$'s guitar practice environment; (b) $P4^{int}_{pn}$'s piano practice environment; (c) $P3^{adv}_{pn}$'s practice environment setup which works around the audio problem.

possible. Most piano learners used touchscreen devices, such as tablet and phones, since they could be placed on the piano's stand. However, touchscreen devices lack the keyboard shortcuts and are inconvenient for controlling the video navigation interface. When using laptops, learners often used the touchpad or the mouse to drag on the progress bar or control speed and used the keyboard shortcuts — space bar, left and right arrows — to skip and pause.

Learners strove to listen to both the video and the instrument at the same time, but it was difficult to properly set up audio devices with acoustic instruments. Putting on speaker while playing along with a video made it hard to hear the video. This was problematic when the instrument was not electric, since the main volume of the instrument could not be adjusted ($P20^{adv}_{pn}$). Moreover, it was often inconvenient to use speakers when there were people around ($P4^{int}_{pn}$, $P10^{int}_{gt}$). On the other hand, wearing headphones while playing along with a video made it hard to hear the instrument ($P6^{int}_{pn}$). Learners who used electric instruments, however, could completely eliminate this problem. For instance, $P3^{adv}_{pn}$ wore headphones on top of earphones to listen to the MIDI sequencer and the video, respectively, as shown in Fig. 4 (c).

4.3.4 When searching for videos, moving thumbnails provide a crude summary of visual components but not the contents of instruction. As explained in section 4.1.3, thumbnails let the learners see some visual components of the videos, such as the instructor's hands and annotation overlays. However, they were not an adequate indicator of whether a video would be helpful; learners still needed to open and watch each potentially helpful video to determine whether it suited their learning needs. For example, the learners might need to determine if they like the arrangement ($P3^{adv}_{pn}$) or how detailed the instructor explained the materials ($P9^{bgn}_{gt}$). To obtain this information, some learners watched the beginning part of the video ($P12^{int}_{gt}$, $P14^{bgn}_{pn}$); some skimmed through the whole videos and looked at the comments section, usually to see if anyone mentioned the video having or not having what it promised to teach ($P2^{bgn}_{gt}$, $P17^{adv}_{gt}$). With either strategy, checking one video at a time greatly hampered the video searching process.

4.3.5 Video interfaces fail to support navigational strategies for music learning: looping, time stamping, and skipping segments. Learners usually needed to repeat segments of a video many times for obtaining information and for playing along. This navigation behavior has also been identified for arts [31] and make-up tutorials [6]. However, with music tutorials, learners needed to lift their hands from the instruments every time they interacted with the interface and often needed to rewind to very precise timestamps. ("I have to stop it quite often. And it has to be within one second. And I have to roll it back where I recall. And one hand is on the piano." - $P3^{adv}_{pn}$) Several learners used the 5- or 10-second

skips, which were available as keyboard shortcuts, but these skips rarely matched the duration of the target segments. The mini-preview over the progress bar could often provide hints for where particular things happened in a video, especially if the video had distinct visual features for different segments. ("The problematic part is when the notes are far apart on the keyboard, so look at the synthesia bars." - P16$_{pn}^{int}$) However, the mini-preview was not precise and not every video had such visual features. Moreover, it did not work well on touchscreen, since the learners had to keep touching the screen to see the preview. The precision issue also got amplified with touchscreen device due to the "fat finger" problem [30].

5 DISCUSSION

5.1 Generalizability of the findings across learners and instruments

Despite the differences in the learners' conditions — learning styles, musical style preferences, learning goals, etc. — they shared a number of common needs and challenges. For instance, in the search-and-browse task, almost every participant looked at the thumbnails first, regardless of their instruments, learning goals, and skill levels. Our learning model (Fig. 1) also covers the range of learning processes demonstrated by our participants. However, there is also a need to accommodate the learners' individual differences. For instance, when searching for videos, most learners had additional, specific requirements to the characteristics listed in section 4.2 (e.g., P4$_{pn}^{int}$ liked when the instructor explains a chord in "snowman" order because that was how her instructor had taught her before.)

Regarding the instruments, our findings can be easily generalized to other keyboard and string instruments, which only require the two hands to play, are often played solo, and use similar forms of annotations (5-line standard notation and tabs) as references. However, instruments that have non-visual modes of instruction, such as percussion, woodwind, or brass, may require a further investigation. A survey study showed that these three categories of instruments comprised the smallest portions of music educational videos on YouTube [79]. Woodwind and brass learners have also been shown to rely heavily on their instructors, while guitar learners are most autonomous [88]. P1$_{gt}^{int}$, who had learned saxophone before, suggested this could be due to nature of the instruments, "Watching a saxophone player plays doesn't really help, because in a saxophone, you don't look at the notes, ever. You look straight. You're taught at the very beginning that you're not supposed to look at your hands at all."

5.2 The hedonic value and efficiency of a practice session are not always congruous.

Every learner told us that they played music in their leisure time "for fun". The hedonic value and the efficiency of a practice session — how quickly they can complete their targeted performances — are usually mutually beneficial. That is, being efficient makes the learners happy, and being happy gives them the intrinsic motivation for practice, which is shown to be highly correlated with efficiency [75]. However, they don't always go well together. With hedonic value as the ultimate goal, the structures of casual practice sessions are not always goal-oriented, in which case, worrying about achievement and efficiency may interfere with the pleasure of playing. For example, when P8$_{pn}^{bgn}$ learned the song "City of Stars", she watched a scene that contained the song and other scenes from the movie "La La Land" during the observed practice session "for inspiration" and to "take a break" from learning. Such behavior does not contribute to the session-level, goal-oriented progress, but it provides hedonic values and the intrinsic motivation to keep learning [46]. Therefore, it should be supported with non-tutorial resources and learning interfaces.

5.3 Reflecting on the video-based music learning process model

Compared to the regular, self-regulated and autonomous music learning process, our model in Fig. 1 differs mainly in searching and picking resources. Most songs that our learners chose for the diary studies could be performed in various arrangements. So, it was up to the learners to choose the suitable resource(s) for themselves. This is different from music learners under formal instruction, who are usually assigned to specific resource(s). Moreover, since public videos are user-generated and can be erroneous, learners may need to find more resources after they started learning, differently from how music students usually stick to their initially assigned resource(s) throughout the whole process [57].

Our learning model surprisingly shares several similarities to existing models, although our participants were casual learners and were not instructed to be effective in their learning. Similarly to guided music practice strategies in formal instructions, our learners demonstrated iterative practice process [28] and utilized metacognitive regulations mainly for goal setting, self-monitoring, and self-evaluation [7, 75]. They also obtained information in several representations, indicating that they took advantage of the online information ecology [45]. However, the other two aspects of online information ecology, information sharing and community engagement, are not evident in our data. It may be beneficial to further study how online video platforms support these aspects of a learner-centered information ecology [59].

5.4 Needs of tool support for the production of music-instruction videos

The majority of videos that our learners watched have characteristics that could be improved, such as missing annotations and incomprehensive verbal explanations; this implies a need for better video authoring tools for music instructors. Instructional video authoring is a difficult task because it contains several steps from planning to filming to editing [17]. However, many components in the proposed guidelines (section 4.2) can be semi-automated or assisted by a specialized interface to make it easier for video creators to make an effective and instructional video. Beside the need for production tools, video creators can also improve the quality of their videos by carefully considering who their target learners are and which instructional components are required. The creators may be trying to create "one-size-fit-all" videos to reach a wide range of learners [69], which means the videos would be problematic in some ways for everyone. This implies a need for a tool for recommending the appropriate videos or modifying a video that fits one learner to fit another, e.g., adding resources about basic techniques to intermediate-level videos to make them suitable for beginners.

5.5 Implications for designing video-based music learning interfaces

Our findings on the identified challenges can be largely summarized as: video navigation, search-and-browse, and metacognitive regulations. In order to explore and concretize design *implications* of our findings, we created interactive mock-ups of three proof-of-concept (PoC) interfaces.[1] We ran an informal, formative evaluation to generate further insights for each design implication with 8 music learners, drawn out of the participants from the diary study. We asked the participants to do a walkthrough of each PoC interface, then rate and comment on them by comparing against the YouTube interface.

5.5.1 Support video navigation based on the semantic layers of the video's musical segments. To address video interface navigation problems and support the need for annotations in practice sessions, we took inspiration from Liu et al.'s ConceptScape [41] and designed an interactive annotation for semantically skipping and looping segments of the video. (See Fig. 5 in Appendix B for the screenshot of the PoC interface.) Participants found the PoC interface to be very

[1] We do not claim design or systems contributions, which are beyond the scope of this paper.

helpful for both navigating the videos and cross-referencing between the sheet music and the video. ("I'd be able to find specific parts quickly and repeat the part I'm struggling with without fumbling with the technology." - P20$^{adv}_{pn}$) The challenges in implementing an automated system based on this idea lie in (1) obtaining a corresponding annotation, especially when the instructor does not publish it, and (2) synchronizing it to a video. For the first challenge, automatic polyphonic transcription can be utilized to create annotations from the video [32, 86]. For the second challenge, The video can be semantically segmented [43], then synchronized each segment to the annotation with signal processing techniques for musical performances [53] and natural language processing for verbal explanations [74].

5.5.2 Support browsing and searching for videos by surfacing their instructional characteristics. To address the difficulty in finding suitable videos, we created tags based on the guidelines from section 4.2 and surfaced some of the community's comments to help the learners identify suitable videos without having to open each video. (See Fig. 6 in Appendix B for the screenshot of the PoC interface.) The PoC interface's evaluation generated two additional design recommendations. Firstly, the results filtering functionality, which utilizes a fixed set of tags in the PoC, should allow the learners to customize the tags and specify more characteristics (P2$^{bgn}_{gt}$, P3$^{adv}_{pn}$, P20$^{adv}_{pn}$). Secondly, although performing sentiment analysis of the comments [8] helps organize the information (P8$^{bgn}_{pn}$, P19$^{adv}_{pn}$), a bad characteristic for one learner may be good for another, for example, a difficult arrangement might be bad for a beginner learner but good for an advanced one (P5$^{bgn}_{gt}$, P17$^{adv}_{gt}$, P20$^{adv}_{pn}$). Hence, learners who want factual opinions should have an option to deactivate this feature.

5.5.3 Provide guidance for effective, metacognitive regulation during and between practice sessions. To address the lack of metacognitive regulation guidance in video-based music learning, we brainstormed an interface that utilized an intelligent chatbot [22] to guide learners through the high-metacognition pathway, as shown in Fig. 1. (See Fig. 7 in Appendix B for the screenshots of the PoC interface.) The PoC interface generated mixed responses from the participants, which prompted us to revise an important, initial design implication on enhancing the learning effectiveness through metacognitive regulation. P1$^{int}_{gt}$ and P5$^{bgn}_{gt}$ commented that focusing on achieving goals and efficiency too much could reduce the "pleasure" of music learning. This stresses the importance of enhancing the hedonic values of video-based music learning experience, which is more difficult to quantify than the result-oriented effectiveness and is wholly subjective to each learner. Participants also pointed out that the chatbot seemed inflexible (P3$^{adv}_{pn}$, P5$^{bgn}_{gt}$) and "allow for little consideration on the part of the learner" (P1$^{int}_{gt}$). This shows that too much instruction can hamper the exploratory aspect of autonomous learning. Therefore, the amount of recommendations and the method of presenting them should be individualized to each learner's preferences. The ability to "hear and assess the performance" (P8$^{bgn}_{pn}$) could also be helpful in giving more appropriate recommendations [4].

5.6 Limitations and future work

Firstly, most of the participants were graduate students, so they were potentially proficient at learning. However, we made an effort to recruit participants from diverse fields of studies and prior music learning backgrounds. This was to cover many possible learners' lifestyles and learning preferences as possible and to ensure the generalizability of our findings. In the future, it may be beneficial to extend the study to other groups of participants, such as older adults, who are less tech-savvy, or learners with disabilities. Secondly, our diary study's instruction required the participants to focus on song instruction and use at least one video for every practice session. These constraints might have interfered with the participant's learning strategies, e.g., less frequent technique-centered practice sessions without videos. We accommodated for this by asking in the participants during the post-study interviews how they adjusted their learning

strategies under the study setting and how their self-evaluation of practice sessions were consequently affected. We took this information into consideration when analyzing their practice diaries and post-study interviews.

6 CONCLUSION

In this paper, we explored video-based music instrument learning in casual learners. We conducted diary studies with 20 piano and guitar learners. We studied how video resources are utilized in the self-regulated music instrument learning process, developed a set of evaluation criteria of instructional quality of music-related videos, and identified the needs and challenges that learners face with the currently available interfaces and devices. Based on our findings, we generated design implications to enhance video-based music learning. The result of the lightweight evaluation of our initial design ideas highlighted the importance of addressing individual learner's style, preferences, and goals. With the growing abundance of public video resources, our design implications can play a significant role in expanding accessibility to effective music education as well as self-development in other fields of knowledge.

REFERENCES

[1] [n.d.]. https://gb.abrsm.org/en/making-music/4-the-statistics/

[2] [n.d.]. Gallup Organization Reveals Findings of "American Attitudes Toward Making Music" Survey. https://www.namm.org/news/press-releases/gallup-organization-reveals-findings-american-atti#:~:text=Slightlymorethanonein,instrument,accordingtothesurvey

[3] 2020. 37 Mind Blowing YouTube Facts, Figures and Statistics - 2020. https://merchdope.com/youtube-stats/#:~:text=Thetotalnumberofpeople,onYoutubeeverysingleday.&text=Inanaveragemonth,8,49year-oldswatchYouTube.

[4] L. Aprameya. 2018. An Autonomous Intelligent Music Teacher. *IEEE Potentials* 37, 1 (2018), 10–14.

[5] Setareh Beheshti. 2009. Improving studio music teaching through understanding learning styles. *International Journal of Music Education* 27, 2 (2009), 107–115.

[6] Frank Bentley and Janet Murray. 2016. Understanding video rewatching experiences. In *Proceedings of the ACM international conference on interactive experiences for TV and online video*. 69–75.

[7] Carol Benton. 2014. *Thinking about thinking: Metacognition for music learning*. R&L Education.

[8] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam. 2017. Retrieving YouTube video by sentiment analysis on user comment. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 474–478.

[9] Michele Biasutti. 2015. Creativity in virtual spaces: Communication modes employed during collaborative online music composition. *Thinking skills and creativity* 17 (2015), 117–129.

[10] Monique Boekaerts and Alexander Minnaert. 1999. Self-regulation with respect to informal learning. *International journal of educational research* 31, 6 (1999), 533–544.

[11] Anne-Marie Burns, Sébastien Bel, and Caroline Traube. 2017. Learning to play the guitar at the age of interactive and collaborative Web technologies. In *Proceedings of the 14th Sound and Music Computing Conference*. 5–8.

[12] Fabio Calefato, Giuseppe Iaffaldano, Filippo Lanubile, and Federico Maiorano. 2018. Investigating Crowd Creativity in Online Music Communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 27 (Nov. 2018), 21 pages. https://doi.org/10.1145/3274296

[13] Joan G Carson and Ana Longhini. 2002. Focusing on learning styles and strategies: A diary study in an immersion setting. *Language learning* 52, 2 (2002), 401–438.

[14] Avner Caspi, Paul Gorsky, and Meira Privman. 2005. Viewing comprehension: Students' learning preferences and strategies when studying from video. *Instructional Science* 33, 1 (2005), 31–47.

[15] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to design voice based navigation for how-to videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[16] Simy Meng-Yu Chen and Robert W Howard*. 2004. Musical instrument choice and playing history in post-secondary level music students: some descriptive data, some causes and some background factors. *Music Education Research* 6, 2 (2004), 217–230.

[17] Pei-Yu Chi. 2016. *Designing Video-Based Interactive Instructions*. Ph.D. Dissertation. UC Berkeley.

[18] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 93–102.

[19] Eleonora Concina. 2019. The role of metacognitive skills in music learning and performing: theoretical features and educational implications. *Frontiers in psychology* 10 (2019).

[20] Erhan Delen, Jeffrey Liew, and Victor Willson. 2014. Effects of interactivity and instructional scaffolding on learning: Self-regulation in online video-based environments. *Computers & Education* 78 (2014), 312–320.

[21] David Dobbelstein, Gabriel Haas, and Enrico Rukzio. 2017. The effects of mobility, encumbrance, and (non-) dominant hand on interaction with smartwatches. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 90–93.

[22] Debasatwa Dutta. 2017. *Developing an Intelligent Chat-bot Tool to assist high school students for learning general knowledge subjects*. Technical Report. Georgia Institute of Technology.

[23] Thomas Goetz, Nathan C Hall, Anne C Frenzel, and Reinhard Pekrun. 2006. A hierarchical conceptualization of enjoyment in students. *Learning and Instruction* 16, 4 (2006), 323–338.

[24] Elizabeth Goodman, Mike Kuniavsky, and Moed Andrea. 2012. *Observing the User Experience A Practitioner's Guide to User Research*. Waltham, MA,USA.

[25] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 41–50.

[26] Pinar Güzel, Kadir Yildiz, Melike Esentas, and Devrim Zerengök. 2020. " Know-How" to Spend Time in Home Isolation during COVID-19; Restrictions and Recreational Activities. *International Journal of Psychology and Educational Studies* 7, 2 (2020), 122–131.

[27] Susan Hallam. 2001. The development of metacognition in musicians: Implications for education. *British Journal of Music Education* 18, 1 (2001), 27–39.

[28] John T Hart Jr. 2014. Guided metacognition in instrumental practice. *Music Educators Journal* 101, 2 (2014), 57–64.

[29] Halvdan Haugsbakken. 2014. Connecting the dots by YouTubing Tutorials. In *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2009–2018.

[30] Christian Holz and Patrick Baudisch. 2011. Understanding touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2501–2510.

[31] Jescia JH Hopper. 2016. Digitizing the easel: Student perspectives on tutorial videos in the art classroom. *Art Education* 69, 4 (2016), 23–28.

[32] José M Inesta and Carlos Pérez-Sancho. 2013. Interactive multimodal music transcription. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 211–215.

[33] Md Kariul Islam. 2020. The Effect of Music on Mental Health: An Observational Study on Fighting COVID-19 Stress with Music. *SSB Global Journal of Medical Science* 1, 01 (2020), 5–5.

[34] Kimia Kiani, George Cui, Andrea Bunt, Joanna McGrenere, and Parmit K Chilana. 2019. Beyond" One-Size-Fits-All" Understanding the Diversity in How Software Newcomers Discover and Make Use of Help Resources. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[35] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-Driven Interaction Techniques for Improving Navigation of Educational Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 563–572. https://doi.org/10.1145/2642918.2647389

[36] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 31–40.

[37] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 4017–4026.

[38] Nathan B Kruse and Kari K Veblen. 2012. Music teaching and learning online: Considering YouTube instructional videos. *Journal of Music, Technology & Education* 5, 1 (2012), 77–87.

[39] Larry Lagerstrom, Petr Johanes, and Mr Umnouy Ponsukcharoen. 2015. The myth of the six-minute rule: Student engagement with online videos. In *Proceedings of the American Society for Engineering Education*. 14–17.

[40] Yi-Chieh Lee, Wen-Chieh Lin, Fu-Yin Cherng, Hao-Chuan Wang, Ching-Ying Sung, and Jung-Tai King. 2015. Using time-anchored peer comments to enhance social interaction in online educational videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 689–698.

[41] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. ConceptScape: Collaborative Concept Mapping for Video Learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173961

[42] Jennifer A Livingston. 2003. Metacognition: An Overview. (2003).

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[44] Marion Long, Susan Hallam, Andrea Creech, Helena Gaunt, and Linnhe Robertson. 2012. Do prior experience, gender, or level of study influence music students' perspectives on master classes? *Psychology of Music* 40, 6 (2012), 683–699.

[45] Chee-Kit Looi. 2001. Enhancing learning ecology on the Internet. *Journal of Computer Assisted Learning* 17, 1 (2001), 13–20.

[46] Dorothy Lucardie. 2014. The impact of fun and enjoyment on adult's learning. *Procedia-Social and Behavioral Sciences* 142 (2014), 439–446.

[47] Anastassia McNulty and Bojan Lazarevic. 2012. Best practices in using video technology to promote second language acquisition. *Teaching English with technology* 12, 3 (2012), 49–61.

[48] Graham J McPhail. 2013. Developing student autonomy in the one-to-one music lesson. *International Journal of Music Education* 31, 2 (2013), 160–172.

[49] Gary E McPherson and John McCormick. 1999. Motivational and self-regulated learning components of musical practice. *Bulletin of the Council for Research in Music Education* (1999), 98–102.

[50] Jennifer Mishra. 2004. A model of musical memory. In *Proceedings of the 8th International Conference on Music Perception and Cognition. Adelaide, Australia: Causal Productions*. 74–86.

[51] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. 2013. NoteVideo: Facilitating Navigation of Blackboard-Style Lecture Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1139–1148. https://doi.org/10.1145/2470654.2466147

[52] Matt Morain and Jason Swarts. 2012. YouTutorial: A framework for assessing instructional online video. *Technical communication quarterly* 21, 1 (2012), 6–24.

[53] Meinard Müller. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.

[54] Alexander Ng, Stephen A. Brewster, and John H. Williamson. 2014. Investigating the Effects of Encumbrance on One- and Two- Handed Interactions with Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1981–1990. https://doi.org/10.1145/2556288.2557312

[55] Cuong Nguyen and Feng Liu. 2015. Making software tutorial video responsive. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1565–1568.

[56] Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2012. Video Summagator: An Interface for Video Summarization and Navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 647–650. https://doi.org/10.1145/2207676.2207767

[57] Siw Nielsen. 2001. Self-regulating learning strategies in instrumental music practice. *Music education research* 3, 2 (2001), 155–167.

[58] Siw G Nielsen. 1999. Learning strategies in instrumental music practice. *British Journal of Music Education* 16, 3 (1999), 275–291.

[59] SA O'Neill. 2014. Mind the gap: Transforming music engagement through learner-centred informal music learning. *The Recorder: Journal of the Ontario Music Educators' Association* 56, 2 (2014), 18–22.

[60] Caroline Palmer and Rosalee K Meyer. 2000. Conceptual and motor learning in music performance. *Psychological Science* 11, 1 (2000), 63–68.

[61] Rosie Perkins and Aaron Williamon. 2014. Learning to make music in older adulthood: A mixed-methods exploration of impacts on wellbeing. *Psychology of Music* 42, 4 (2014), 550–567.

[62] Paul R Pintrich. 1995. Understanding self-regulated learning. *New directions for teaching and learning* 1995, 63 (1995), 3–12.

[63] Maddy Shaw Roberts. 2018. 10 reasons you should take up a musical instrument. https://www.classicfm.com/discover-music/reasons-to-play-a-musical-instrument/

[64] Katja Rogers, Amrei Röhlig, Matthias Weing, Jan Gugenheimer, Bastian Könings, Melina Klepsch, Florian Schaub, Enrico Rukzio, Tina Seufert, and Michael Weber. 2014. Piano: Faster piano learning with interactive projection. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*. 149–158.

[65] Abdulhadi Shoufan. 2019. What motivates university students to like or dislike an educational online video? A sentimental framework. *Computers & education* 134 (2019), 132–144.

[66] Abdulhadi Shoufan and Fatma Mohamed. 2017. On the likes and dislikes of youtube's educational videos: A quantitative study. In *Proceedings of the 18th annual conference on information technology education*. 127–132.

[67] Sunghyun Song, Jeong-ki Hong, Ian Oakley, Jun Dong Cho, and Andrea Bianchi. 2015. Automatically adjusting the speed of e-learning videos. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1451–1456.

[68] Nicholas Michael Stefanic. 2014. Creativity-based music learning: Modeling the process and learning outcomes in a massive open online course. (2014).

[69] Craig Stewart, Alexandra I Cristea, Tim Brailsford, and Helen Ashman. 2005. 'Authoring once, delivering many': creating reusable adaptive courseware. (2005).

[70] Jason Swarts. 2012. New modes of help: Best practices for instructional video. *Technical Communication* 59, 3 (2012), 195–206.

[71] Yoshinari Takegawa, Tsutomu Terada, and Masahiko Tsukamoto. 2008. A piano learning support system considering rhythm. *Journal* 32, 1 (2008), 9–25.

[72] Cristen Torrey, Elizabeth F. Churchill, and David W. McDonald. 2009. Learning How: The Search for Craft Knowledge on the Internet. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1371–1380. https://doi.org/10.1145/1518701.1518908

[73] Anh Truong and Maneesh Agrawala. 2019. A Tool for Navigating and Editing 360 Video of Social Conversations into Shareable Highlights. In *Proceedings of Graphics Interface 2019* (Kingston, Ontario) *(GI 2019)*. Canadian Information Processing Society, 9. https://doi.org/10.20380/GI2019.14

[74] Adrian Ulges, Christian Schulze, Markus Koch, and Thomas M Breuel. 2010. Learning automatic concept detectors from online video. *Computer vision and Image understanding* 114, 4 (2010), 429–438.

[75] Wynnpaul Varela, Philip C Abrami, and Rena Upitis. 2016. Self-regulation and music learning: A systematic review. *Psychology of Music* 44, 1 (2016), 55–74.

[76] Janice Waldron. 2013. User-generated content, YouTube and participatory culture on the Web: Music learning and teaching in two contrasting online communities. *Music Education Research* 15, 3 (2013), 257–274.

[77] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 405–416.

[78] Jennifer A Whitaker and Evelyn K Orman. 2014. A content analysis of responses to music teaching videos on YouTube. *ISME Commission on Research* (2014), 330.

[79] Jennifer A Whitaker, Evelyn K Orman, and Cornelia Yarbrough. 2014. Characteristics of "music education" videos posted on YouTube. *Update: Applications of Research in Music Education* 33, 1 (2014), 49–56.

[80] Jennifer A Whitaker, Evelyn K Orman, and Cornelia Yarbrough. 2016. The effect of selected parameters on perceptions of a music education video posted on YouTube. *International perspectives on research in music education* (2016), 255–272.

[81] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.

[82] Ruth Wright and Panagiotis Kanellopoulos. 2010. Informal music learning, improvisation and teacher education. *British Journal of Music Education* 27, 1 (2010), 71.

[83] Xiao Xiao, Paula Aguilera, Jonathan Williams, and Hiroshi Ishii. 2013. MirrorFugue iii: conjuring the recorded pianist.. In *CHI Extended Abstracts*. Citeseer, 2891–2892.

[84] Weilong Yang and Zhensong Qian. 2011. Understanding the Characteristics of Category-Specific YouTube Videos. *Entstanden im Rahmen eines Informatikseminars an der kanadischen Simon Fraser University* (2011).

[85] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S. Fels. 2019. "Can You Believe [1:21]?!": Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300719

[86] Kazuki Yazawa, Katsutoshi Itoyama, and Hiroshi G Okuno. 2014. Automatic transcription of guitar tablature from audio signals in accordance with player's proficiency. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3122–3126.

[87] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 5372–5384.

[88] Katie Zhukov. 2015. Exploring the role of technology in instrumental skill development of Australian higher education music students. *Australian Journal of Music Education* 23, 2 (2015).

A PARTICIPANT INFORMATION

Table 3. Demographics and instrument skill levels information of all participants when they began the diary study.

No.	Gender	Age	Occupation	Instrument	Skill level
P2	Male	28	Graduate student - CS	Guitar	Beginner
P5	Male	27	Graduate student - Economics	Guitar	Beginner
P7	Female	27	Research Assistant - Electrical and Computer Engineering	Guitar	Beginner
P9	Female	30	Graduate student - Psychology	Guitar	Beginner
P1	Male	34	Graduate student - Educational Studies	Guitar	Intermediate
P10	Male	27	Graduate student - Geography	Guitar	Intermediate
P12	Male	20	Undergraduate student - Civil Engineering	Guitar	Intermediate
P15	Male	44	Software developer	Guitar	Advanced
P17	Male	25	Self-employed	Guitar	Advanced
P18	Male	21	Undergraduate student - CS	Guitar	Advanced
P8	Female	31	Graduate student - MBA	Piano	Beginner
P11	Female	25	Graduate student - Science Education	Piano	Beginner
P13	Female	25	Graduate student - Pharmaceutical Sciences	Piano	Beginner
P14	Female	62	Flight attendant	Piano	Beginner
P4	Female	23	Graduate student - English	Piano	Intermediate
P6	Male	25	Graduate student - Linguistics	Piano	Intermediate
P16	Female	22	Undergraduate student - English	Piano	Intermediate
P3	Male	55	University lecturer - CS	Piano	Advanced
P19	Male	43	Business owner	Piano	Advanced
P20	Female	21	Undergraduate student - CS	Piano	Advanced

B PROTOTYPE INTERFACES

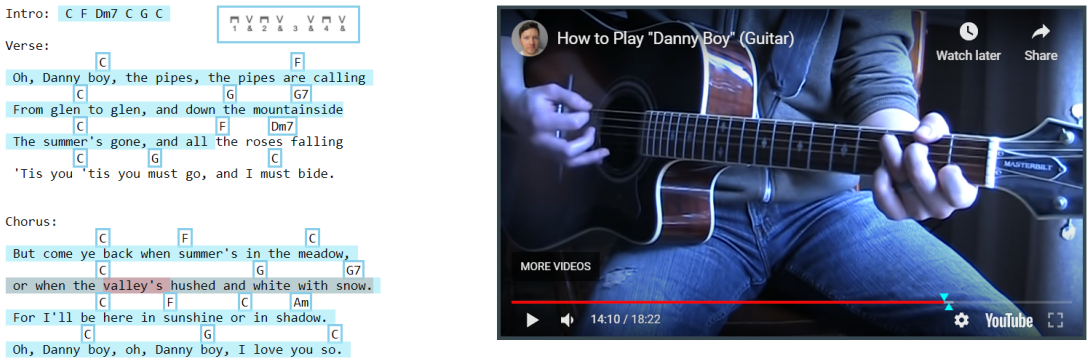


Fig. 5. The blue color indicates the interactive areas on the interface. The learner can highlight a region of the score and go directly to explanation or demonstration of that part in the video and automatically repeat the segment. The learner can also navigate to specific contents by clicking on related objects on the annotations, such as chord names, strumming patterns, or musical symbols.

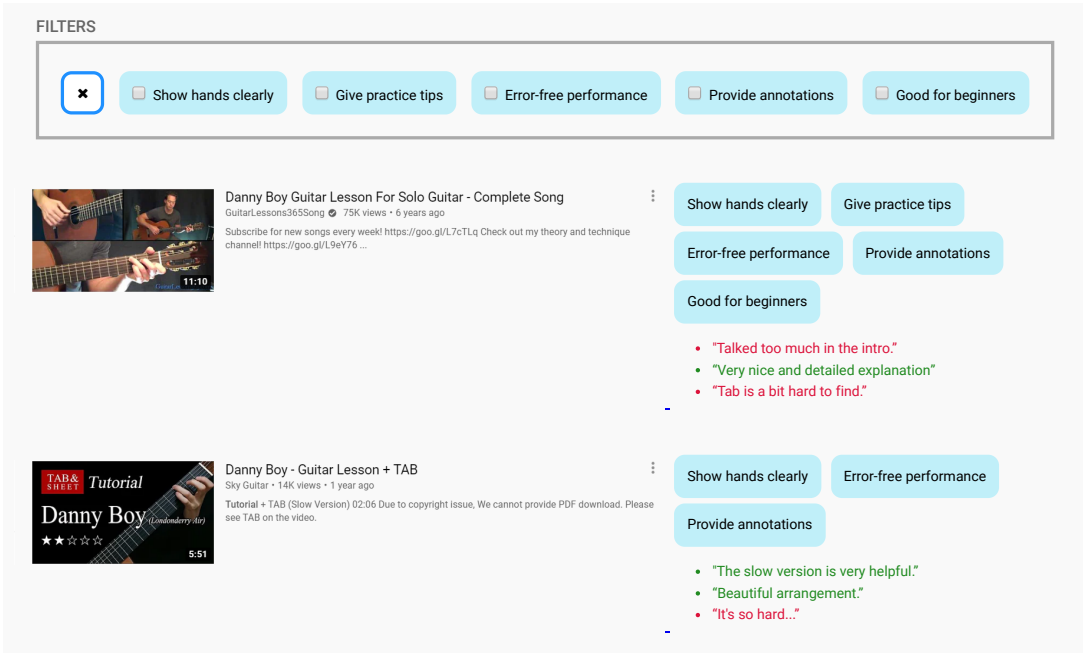
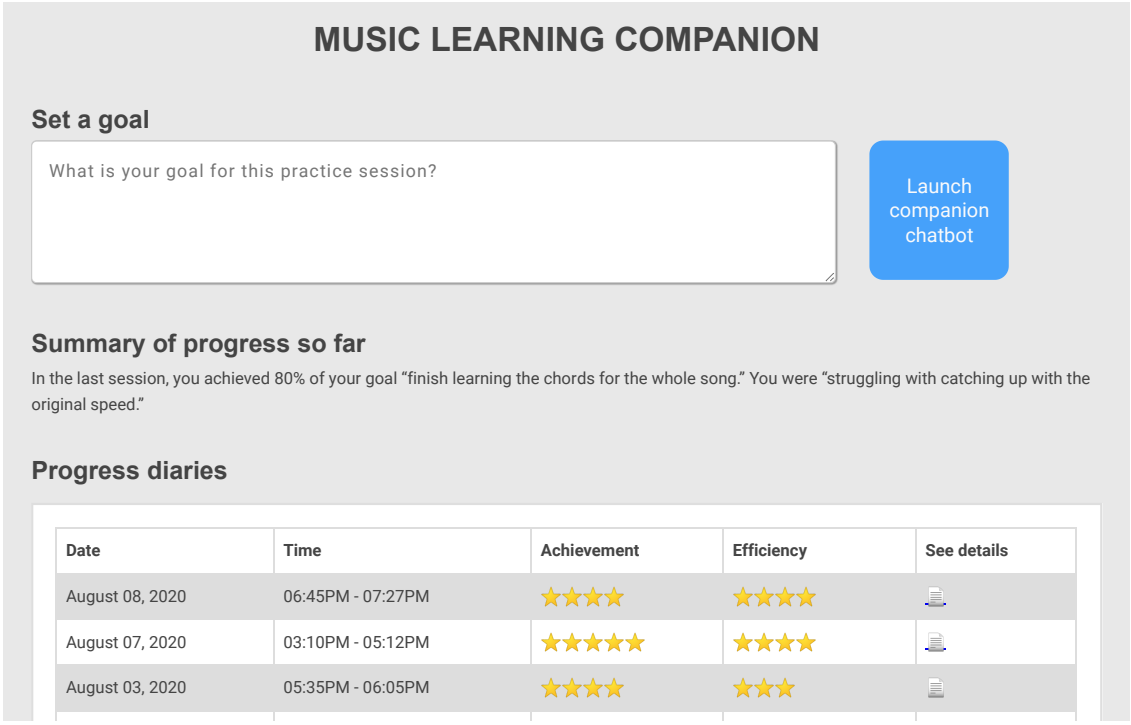
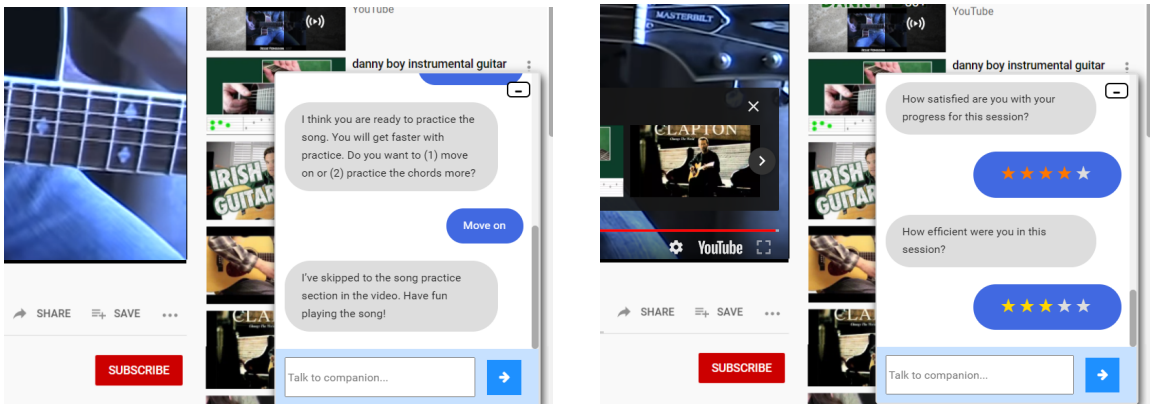


Fig. 6. Learners can choose the tags they would like to include with their search query to filter their search results using the checkboxes in the FILTERS box. Each search result includes a set of tags and a selection of helpful comments from other learners, color-coded based on the sentiment of the comment, green for positive and red for negative.



(a)



(b)

Fig. 7. (a) The interface first asks learners to set their goals at the beginning of each session. (b-left) the chatbot then takes a proactive role in guiding the learner towards their goal by suggesting videos for them to use, providing supplementary materials if they get stuck, and suggesting practice strategies. (b-right) At the end of the session, the chatbot prompts the learner to evaluate their session performance and keep track of their progress.