Designing CAST: A Computer-Assisted Shadowing Trainer for Self-Regulated Foreign Language Listening Practice

Mohi Reza
mohireza@alumni.ubc.ca
University of British Columbia
Vancouver, Canada

Dongwook Yoon
yoon@cs.ubc.ca
University of British Columbia
Vancouver, Canada

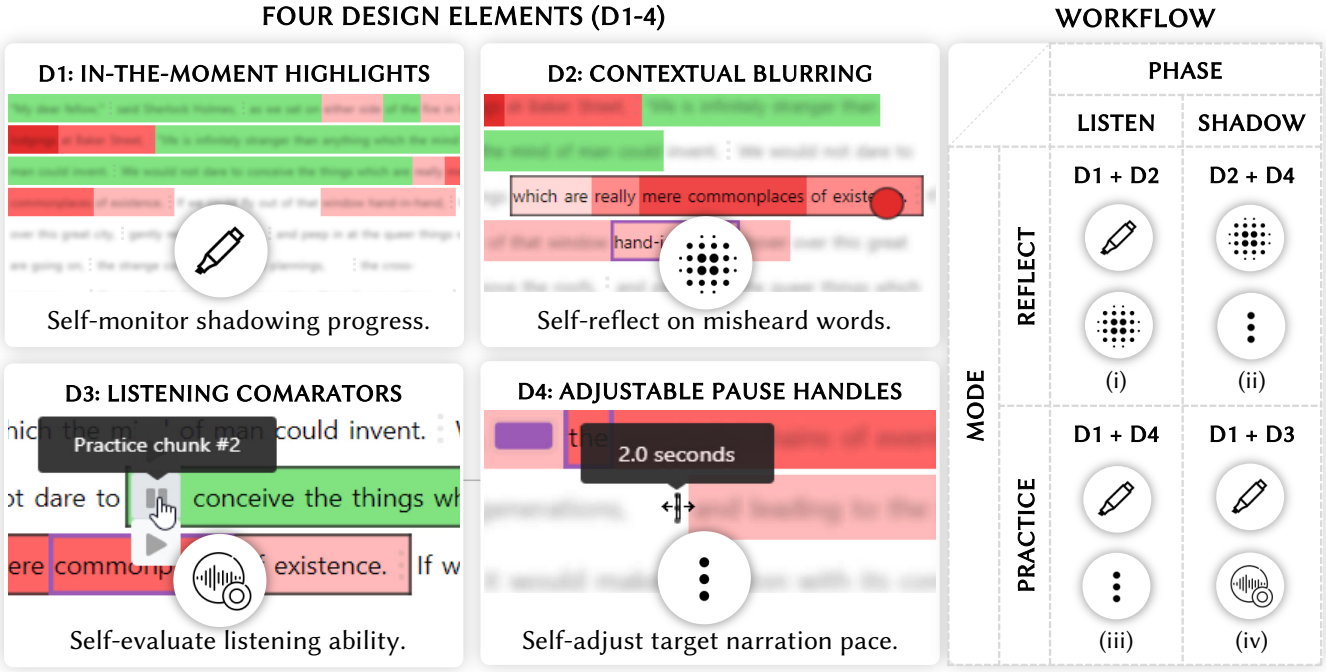

Figure 1: The Self-Regulated Shadowing (SRS) process with CAST: (i) While listening, learners track their progress with D1 by noticing misheard words from the transcript with D2. (ii) While shadowing, they get timely support from the transcript with D2, and short breaks with D4 to make their learning experience less overwhelming. (iii) After listening, learners review difficult chunks using D1, and adjust pause lengths between them using D4. (iv) After shadowing, learners spot-check practice recordings with D1 and D4.

ABSTRACT

Shadowing, i.e., listening to recorded native speech and simultaneously vocalizing the words, is a popular language-learning technique that is known to improve listening skills. However, despite strong evidence for its efficacy as a listening exercise, existing shadowing systems do not adequately support listening-focused practice, especially in self-regulated learning environments with no external feedback. To bridge this gap, we introduce **Computer-Assisted Shadowing Trainer (CAST)**, a shadowing system that makes self-regulation easy and effective through four novel design elements — (i) *in-the-moment highlights* for tracking and visualizing progress, (ii) *contextual blurring* for inducing self-reflection on misheard words, (iii) *self-listening comparators* for post-practice self-evaluation, and (iv) *adjustable pause-handles* for self-paced practice. We base CAST on a formative user study (N=15) that provides fresh empirical grounds on the *needs* and *challenges* of shadowers. We validate our design through a summative evaluation (N=12) that shows learners can successfully self-regulate their shadowing practice with CAST while retaining focus on listening.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445190

CCS CONCEPTS

• **Applied computing → Computer-assisted instruction**; • **Human-centered computing → Interactive systems and tools**; *Empirical studies in HCI*.

KEYWORDS

computer-assisted language learning, self-regulated learning, multimedia learning, speech shadowing, audio and speech interfaces

ACM Reference Format:
Mohi Reza and Dongwook Yoon. 2021. Designing CAST: A Computer-Assisted Shadowing Trainer for Self-Regulated Foreign Language Listening Practice. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445190

1 INTRODUCTION

Speech shadowing, i.e., listening to some target audio and *immediately* vocalizing the words [29], is a popular language-learning technique that is known to be effective for listening skill development [17]. Unlike written text, where the boundaries between words are clearly delineated, speech is a transient concoction of phonemes, strung together in a continuous stream of sounds. While *native speakers* can effortlessly disentangle these phonemes into words, *non-native speakers* have a much harder time. This is where shadowing helps with listening — it sharpens the phoneme perception skills of non-native speakers [20, p. 47], thereby improving their ability to extract words from sounds.

However, existing software tools for shadowing do not provide adequate support for *listening* practice, because mainstream usage of the technique is fixated on *speaking* skill development. The difference between shadowing for *listening* and shadowing for *speaking* is significant. The former targets *bottom-up listening skills* [21, 46], i.e., the ability to recognize words from phonemes, whereas the latter targets aspects of *oral proficiency* such as pronunciation, accent, and intonation, which are tangential to listening skill development. A good example of a recent shadowing system from the HCI community that takes the latter approach is *WithYou* [56], a *speech-tutoring* system that automatically adjusts audio playback and difficulty level by comparing "a learner's *speech and pronunciation*", with a "speech template to determine if a learner's performance is good or not" [56]. Off-the-shelf shadowing apps (e.g., [12, 26, 39]) share a similar focus on speaking skill development.

The major barrier to the development of new software tools is a lack of understanding in the field around the specific *needs and challenges* associated with listening-focused shadowing practice. This gap is critical and somewhat unexpected, given that the background literature on the effectiveness of shadowing for listening practice is *more substantive* than for speaking practice [21, p. 390]. This *does not* undermine the usefulness of speaking-focused shadowing systems, because speaking skills are important, and such systems may catalyze future research efforts on shadowing for speaking. However, this *does* signify a clear need for the development of shadowing systems that focus on listening skill development.

To develop a system for listening-focused shadowing, we first bridged the gap in our understanding of learner needs and challenges by conducting a formative user study with 15 English as a Second Language (ESL) students, and found *self-regulation* to be the major stumbling block for listening-focused shadowing practice. We drew from a rich body of literature on Self-Regulated Learning (SRL) theory to ground our findings, and shaped the process for listening-focused shadowing around Zimmerman's SRL cycle [58]. We found that aspects of shadowing practice tied to SRL, such as, *monitoring* listening ability, *reflecting* on misheard portions of the target audio, *self-evaluating* shadowing performance, and increasing the overall *self-awareness* during practice were areas where learners needed most support. Supporting these aspects proved to be particularly challenging for shadowing because the activity requires heavy multi-tasking (i.e. listening and vocalizing the words *at the same time*), which overwhelms the learners with high cognitive load [18].

The *transcript*, i.e. the written form of the target audio, proved to be a potential source of support for learners because we found that reading misheard words *after* listening to them enhanced the learner's ability to reflect on their mistakes. However, we also found that using the transcript led to what we describe as the *text-dependency problem* — firstly, when given access to the full transcript, learners were tempted to read words *before* listening to them. This behaviour diminished their opportunity to notice misheard words because they already saw them in writing. Secondly, reading from the transcript shifted focus *away* from listening, thereby, hampering the main learning goal of listening-skill development. This second observation is tied to prior experimental work on *selective attention* and reading while listening, which shows that our "ability to read and to listen concurrently is limited by the availability of both general and task-specific processing capacity" [31]. Our core challenge, then, was to design a system that supports the learner's ability to self-regulate their shadowing practice using the transcript, while retaining a strong focus on listening.

To address this challenge, we designed **Computer-Assisted Shadowing Trainer (CAST)**, a listening-focused shadowing system that enhances the Self-Regulated Shadowing (SRS) process, and works well in situations with no external feedback. Through iterative design, we developed an ensemble of four novel design elements that work together to make SRS easy and effective (see Figure 1): (i) *in-the-moment highlights*, i.e., light-weight text-highlighting interactions over a blurred transcript for progress tracking, (ii) *contextual blurring*, i.e., blurring and deblurring selective parts of the transcript to resolve the text-dependency problem, (iii) *listening comparators*, i.e., shadowing recordings interlaced with the transcript for post-practice self-evaluation, and (iv) *adjustable pause-handles*, i.e., strategically positioned draggable handles that can be adjusted to introduce short breaks between difficult chunks without altering playback speed. We validated our design through a summative evaluation study (N = 12) that provides evidence in support of the efficacy of CAST as a self-regulated shadowing tool for listening skill development.

In this work, we contribute: (i) CAST, the first self-regulated shadowing system for foreign language listening practice, fresh empirical insights on learner needs and challenges associated with listening-focused shadowing, upon which we base our design approach, and results from a summative evaluation that validates our design.

2 RELATED WORK

Our work has been informed by SRL theory, previous education technologies in HCI concerned with self-regulated learning, the rich body of shadowing literature from language pedagogy, cognitive psychology, and simultaneous interpreter training, and speech-based interfaces that use visual representations of audio to overcome its linear nature.

2.1 Shadowing is Rooted in Listening

The background literature on shadowing reveals that the technique has important applications in two different, albeit connected, disciplines — *cognitive psychology* and *simultaneous interpretation*. While in this work, we are primarily interested in *language pedagogy*, tracing the rich history of the technique back to those two disciplines gives us useful insights on why shadowing is effective as a listening exercise.

Starting as early as the 1950s, shadowing was used by cognitive psychologists to study selective attention [6]. A classic example of this is the application of shadowing in *auditory attention experiments* [8] on Moray's cocktail party effect [34] — why are we so good at tuning into a single conversation amidst a cacophony of background voices?

In those experiements, learners were given a dichotic listening task involving two different audio streams, one in each ear, and asked to shadow only one of those streams. For the stream they shadowed, participants were unable to recall its contents, i.e., they focused on the *sounds* of the words, *not their meaning* [20]. For the other stream, participants were completely oblivious to its message, and did not notice even when its language was altered mid way from English to German [8], i.e. they focused *solely on the shadowed stream*. These results hint at the power of shadowing as *focusing technique* that forces learners to pay close attention to the sounds of a single audio stream.

Shadowing found its second home among simultaneous interpreters [27, 29], i.e., those who translate between languages in real-time. The technique became a precursory exercise that helped trainee interpreters practice timing, listening, and short-term memory skills [35], which can also benefit language learners because shadowing improves phoneme-perception skills [17]. Here, timing refers to the latency between heard and reproduced speech.

Building on insights from cognitive psychology and simultaneous interpreter training, researchers from a Japanese EFL pedagogy context [19, 46–48] spearheaded efforts in shaping the shadowing technique into a language learning exercise for bottom-up listening-practice. Since then, because of a growing global interest in shadowing, the results of those efforts have been made accessible to a wider international audience [17, 20].

While few preliminary studies have looked at the potential impact of shadowing on aspects of speech such as pronunciation [32], intonation [23] and oral fluency [53], the research on the impact of shadowing on listening skill development is more substantial [20, p. 390]. This makes our focus on listening-skill development with CAST well-aligned with pre-existing shadowing research.

2.2 Existing Shadowing Systems

While the theoretical underpinnings of shadowing as a listening exercise are well-understood [20, p. 9], existing shadowing systems do not adequately address listening-focused shadowing, because popular usage of the technique remains fixated on speaking practice.

A quick search for off-the-shelf shadowing apps brings to light the imbalanced focus on speaking over listening. For example, downloadable shadowing apps such as [12, 26, 39] all focus solely on improving English *speaking* skills: [12] describes shadowing as "training for English fluency", and "the best way to improve English speaking", and [26] frames it as a technique for learning how to "speak like a native by improving your pronunciation, rhythm, and intonation."

A recent and noteworthy shadowing system stemming from the HCI community is *WithYou* [56], which uses "context-dependent speech recognition" to automatically adjust the audio playback and the difficulty of a "native speech template" when learners fail to shadow smoothly, thereby supporting them when they face difficulties, and helping them improve their speaking skills. We distinguish CAST from these existing systems by noting its strong focus on listening-skill development.

2.3 Enhancing Self-Regulated Learning

If we zoom out from the specific learning context of listening-focused shadowing, we can situate CAST within a broader array of *systematic interventions* for enhancing SRL. We are motivated by previous SRL literature in favour of the notion that the "students' self-regulatory competence can be enhanced through systematic interventions"[43]. What strings together these interventions with CAST is their shared conceptual framework, as described by various SRL models [38], two of which are of particular interest to us, namely, the Pintrich [40] and Zimmermann [57] model.

Pintrich's model comprises four phases: (i) forethought, planning and activation, (ii) monitoring, (iii) control, and (iv) reaction and reflection [40]. These phases are highly flexible, and only "specifies the possible range of activities" for SRL, and "does not necessitate them", nor does it "presume that the phases are linearly ordered" [42]. Therefore, when designing CAST, we considered which aspects from these phases need support in our specific learning context by analyzing empirical findings from our exploration of learner needs. For guidance on ordering, we turned to Zimmerman's model, and shaped our SRS process around its cyclical phases, namely, (i) forethought, (ii) performance, and (iii) self-reflection [57].

2.4 Designing for Self-Reflection

Of the three phases in Zimmerman's model, self-reflection is noteworthy because we are concerned with enhancing the learner's ability to reflect on misheard words from the target audio. Supporting reflective practice through design has been of particular interest to HCI researchers for some time now [4, 14]. We can broadly classify the various approaches for inducing self-reflection under *prompting* (e.g., [7, 44, 51], and *visualization* (e.g., [9, 16, 49]).

In the first classification, learners self-reflect by responding to prompts that concretize their thinking. For example, a learner may be asked to explain their solution to a math problem [51], or to

answer reflective questions while watching educational videos [44]. This approach works well only in situations where *interrupting* the learner is okay. We must also make the a priori assumption that the learner is able to *recall* how their practice went when interrupted. With shadowing, prompting is unsuitable because it induces heavy cognitive load [45], making interruptions *during* practice far too obtrusive, and prompts *after* practice ineffective due to reduced immediate recall [5].

In the second classification, learners use *information visualization* to glean insights from their experience, and in doing so, become more self-aware of their learning process. For example, a learner may reflect on how they spend their time by using charts and time-logs [16]. Visualization becomes especially helpful in situations where moving information from the learner's memory to an external form enables them to see new patterns. In the context of shadowing practice, the transcript can be used as a visual counterpart to the audio. However, as we shall discuss in Section 3, using the transcript for listening-focused shadowing comes with many caveats that we address through design.

2.5 Visual Representation for Audio

From the perspective of interaction design, the transcript-based speech navigation features in CAST have their roots in early HCI systems such as SpeechSkimmer [3] and SCANMail [50] and more recent systems such as RichReview [54], TypeTalker [2] and Skimmer [25]. These systems overcome the transient, un-skimmable nature of audio using *visual representations of sound* such as audio transcripts [25, 50], threaded wave-forms [54], and captions [50, 54]. Early systems favoured non-transcript visualizations such as wave-forms and binary representations [22] because automatic transcript generation was not yet practical. Since then, computer-generated audio transcriptions have become inexpensive and accurate, and so we opt for transcript-driven audio representation in CAST.

Our design approach adds new interaction techniques for multimedia navigation and consumption, such as blurring and revealing parts of the transcript to guide the user's attention, as detailed in Section 4.2, and interlacing multiple audio segments with a comparator for in-situ reflection on one's own performance, as detailed in Section 4.3.

3 EXPLORING LEARNER NEEDS

To understand the *needs and challenges* of language learners associated with listening-focused shadowing practice, we conducted a formative need-finding study where we explored how language learners practiced shadowing using a representative audio player and document viewer, and looked at how their needs and challenges were connected with two popular forms of shadowing instruction, namely, *video-based* and *in-person* instruction.

3.1 Method

We conducted semi-structured interviews with our participants after providing them with two shadowing tasks.

3.1.1 Participants . Our participants consisted of 15 international students aged 18 to 24 (12 women and 3 men), who had taken ESL lessons within the last two years. We screened them for first-language (L1) variety so that our findings weren't tied to specific L1 traits. Our participants spoke Chinese (Mandarin, Cantonese, and other local dialects), Korean, Russian, Ukranian, Hindi, and Arabic as L1, and their English proficiency levels ranged from A2 (beginner) to C1 (advanced) on the CEFR scale [37], with B2 (intermediate) being the most common. These levels were assigned and cross-validated by two native English speakers based on a two-minute recorded conversation at the beginning of each interview. Responses on prior familiarity with shadowing ranged from *definitely not* ($13.\overline{33}\%$) and probably not ($26.\overline{66}\%$), to *probably yes* (50%) and *definitely yes* (20%).

3.1.2 Tasks . The first task simulated a scenario where participants discovered shadowing from a video, and practiced on their own. The second task simulated a scenario where they had access to one-on-one guidance from an instructor. By asking them to practice shadowing twice, first with video-only instruction, and then with individualized guidance from an instructor, we were able to identify which of the needs and challenges were intrinsic to the technique, and which were tied to the form of guidance.

3.1.3 Materials . We used *Arthur the Rat* [1], a standard passage in native British English [28], as shadowing material. To ensure equal difficulty levels for both tasks, we divided the passage into two halves of equal length of around 160 words each, and used one half for each shadowing task, counterbalancing the order in which they were presented. As instructions for the first task, we used a highly popular video on shadowing [36]. This video is representative of what learners may typically find when doing online searches on shadowing[1].

3.1.4 Procedure . We began the study by demonstrating how the tools worked, and asked participants to try them out to make sure they were comfortable with using them. Then, they completed the first shadowing task after watching the video. For the second task, we provided in-person guidance on the shadowing process by going over each step based on a script adapted from shadowing instructions in [20]. Then, participants completed the second shadowing task. We recorded the computer-screen and audio during both tasks, and made observation notes from a distance. Finally, with the shadowing experience fresh in their mind, we conducted a follow-up interview where we unpacked our participants' needs and challenges. The entire study took approximately one hour to complete.

The lead investigator conducted the whole study. During pilots, it became apparent that shadow learners tend to be self-conscious and anxious when others are present. To remedy their self-consciousness, we decided not to outnumber the participant during the study. To minimize potential demand characteristics, all study procedures, including the specific instructions given to the learner, followed a script that was validated by the second author.

3.1.5 Analysis . Our data consisted of interview transcripts and task observation notes. To deduce a set of requirements for CAST that addresses real-world learner needs and challenges, we coded and analyzed the data using reflexive thematic analysis [10] through

[1]While the video's title mentioned speaking practice only, the content covered the role of listening in shadowing, and includes statements such as "[shadowing] trains your ear to *listen very very carefully*", and that it helps the learner to become "very good at hearing".

an inductive-deductive lens, using the rich theory on listening-focused shadowing as a pre-existing code that guided our interpretations.

3.2 Findings

Our findings (F1-F5) indicate the need for a tool that enhances the learner's ability to self-regulate their shadowing practice. We order them considering both their prevalence in our data, and our judgement on their importance.

F1. Learners want to practice alone. When asked to describe the ideal environment for shadowing, 9 participants made 20 references on their desire for mental and physical space, which is tied to the need to practice alone. This was due to two interrelated factors, namely, *self-consciousness* and the *inability to concentrate* on shadowing in front of others, especially their instructor. P9, P12 and P13 felt uncomfortable "speaking in front of people". The issue of self-consciousness has been observed in related HCI sytems which require ESL students to speak, e.g., [55], but was exacerbated in our context because the shadowed speech produced by our participants was often garbled and unintelligible, making them feel even more self-conscious. P16 mentioned that the *act of imitating someone* was "kind of embarrassing". Isolation offered participants the *freedom to make mistakes*, which they valued. When alone, P4 felt that they were "...free to talk aloud...to make mistakes...to miss words...and to say them incorrectly". P14 felt that shadowing in front of an instructor "can be so messed up", because being observed made them "feel pressured". The power imbalance typical in student-instructor relationships is therefore an important sub-factor that makes self-regulated learning an attractive choice because that negates the need for external observation.

F2. Learners have a hard time self-regulating their practice. 9 participants made 19 references to the sense of overwhelm they felt during practice, which hampered their ability to self-regulate. When practicing alone, the learner's ability to self-regulate their practice becomes important due to "high levels of learner autonomy and low levels of teacher presence" [52]. However, we found that self-regulation didn't come easy for learners, as they were prone to making poor strategic choices during practice, even when provided with video-instructions on how to shadow. The video instructed the participants to (i) listen very carefully to the audio, then (ii) shadow with the transcript, and then, (iii) shadow *without the transcript*. P7 and P10 remained quiet during the *entire* practice session, missing the basic requirement that words must be vocalized during shadowing. P7 thought that vocalization wasn't necessary and P10 only quietly moved their mouth. P1 never practiced without the transcript, even though step (iii) in the video required them to do so. P3 spent considerable time reading the text *before* playing the audio, getting the order of steps wrong. With in-person guidance in the second task, based on observed performance from the first task, participants made better choices during practice. However, providing such guidance requires observation from an instructor during practice, which is in direct conflict with their need to practice alone (F1). We interpret this observation on not following instructions as an issue of self-regulation because it indicates a misalignment between the learner's strategic choices during practice (e.g. remaining quiet, reading before listening), and their learning goal (i.e. listening improvement) due to inadequate self-monitoring. Learning strategy formation, goal setting, and self-monitoring are all key processes in self-regulated learning [56], and therefore, such misalignment can be framed as a problem of self-regulation.

F3. Learners tend to read from the transcript *before* listening to the target audio. 15 participants made 55 references to their desire to read before listening. Listening-focused shadowing requires learners to rely on their ears. However, audio-only shadowing is cognitively demanding [24]. When given access to the transcript, we found that participants were tempted to use *reading* instead of *listening* as a shortcut shadowing strategy, which hampered their listening practice. This observation formed the basis of the *text-dependency problem* as discussed in Section 1. P9 "tried to *read the words* at the same time as the audio...", while focusing primarily on the text. P11 thought "reading the text and *then* saying things" made the exercise easier, because with the text gone, they could only *partially* recall what was there before. We can glean three important patterns from these comments. First, participants relied on the text because reading felt easier than audio-only shadowing. Second, as confirmed by additional comments from P6 and P14, participants were reluctant to remove the transcript because they were trying to *memorize* the the material in advance. Third, participants turned to the transcript as a source of support. The text-dependency problem became more pronounced when participants couldn't keep up with the narrator's pace, because shadowing became a difficult game of playing catch-up, as described by P10: "...the audio just keeps carrying on, and I have to catch-up...but that's very hard."

F4. Reading *soon after* listening helps learners reflect on their mistakes. 6 participants made 9 references to the phenomenon of the text leading to self-reflection on mistakes. Checking the transcript *soon after* listening to a difficult portion led to *aha moments*, where participants realized what they misheard. For example, P11 misheard "hole" as "home" during the listening step, and only realized this when they checked the transcript while shadowing that part. P1, P10, and P12 had a similar experience, as summarized by P9: "...I recognized so many things, so many words that I *thought* I understood, but it turned out to be a different word."

Our data offers insights on exactly *when* those aha-moments occur. Participants (P11, P13, P14) noticed misheard words if they checked the text *soon after* the audio reached that point in the passage. Reviewing the text *before* listening to the target audio diminished their opportunity to mishear something, since they had already seen the word in writing. Reviewing the text *after* listening only worked if the learner did not have to search through the text. P14 reported that it took too long to "search through the text" when looking for a misheard word. By the time the learner located the misheard word, if the audio had moved much further into the passage, their opportunity to notice misheard words was reduced. Reviewing the text *soon after* mishearing something maximized their chances for self-reflection — that is where the learning happened. Therefore, the timing window for self-reflection is small, and precise time-synchronized stimuli from the audio *and* the transcript is key to effective transcript-induced self-reflection.

F5. Learners tend to skip difficult parts without revisiting them. Our participants (P8-P11) skipped difficult parts when overwhelmed, and did not consistently revisit them in successive rounds because it was difficult for them to recall which parts they skipped once they were done with a round. This negatively impacted their learning because those parts contained words they were most likely to mishear. The extent of overwhelm was unequally distributed. We see this in P9's comment,"in general, I didn't feel like the pace was too fast, except when things became really unfamiliar." Therefore, learners need support for tracking skipped parts, so that they can revisit them with consistency.

4 DESIGNING CAST

We triangulated our findings with self-regulated learning theory [38, 40, 57] multimedia learning theory [33], and existing empirical work on shadowing [6, 8, 20, 34], to define a set of design requirements for CAST. These consist of the overarching requirement (R0) for enabling SRS, and four supporting requirements (R1-4), that when fulfilled, resolve R0. The first two columns in Figure 2 provide an overview of our requirements, and how they bridge our user findings and design elements.

R0. Provide structured guidance on the self-regulated shadowing process. This is the overarching requirement that encompasses all other requirements. While all of our findings point towards R0, the first two are of particular relevance because these indicate that learners want to practice alone (F1), but are unable to self-regulate their practice without added support (F2). R0 positions CAST as an intervention for enhancing self-regulated learning.

R1. Maximize self-reflection on mistakes without inducing text-dependency. This requirement seeks to resolve the conflict between the learner's ability to use to the transcript to reflect on misheard words (F4) and the text-dependency problem (F3). It positions the transcript as a self-reflection device for SRS, but requires us to address the pitfalls of giving full-access to the transcript during practice.

R2. Enable tracking of misheard portions with minimal cognitive overload. This requirement is based on our observation that learners are inclined to skip difficult parts while shadowing, without always remembering to return to them in successive rounds of practice (F5). To fulfill R2, the learner must be able to systematically track misheard words in an external, easy-to-visualize format, so that they can return to those words later rounds without needing to remember everything in their head. The high cognitive load of shadowing [18] makes R2 challenging to fulfill because any additional tasks that learners must do for tracking must not be burdensome to them.

R3. Enable post-practice self-evaluation without requiring practice recall. Frequent self-evaluation is an important self-regulated learning strategy, especially when it "conveys information that students may not acquire on their own" [41]. For shadowing, this information consists of the specific parts of the passage where the learners faced difficulties when speaking along in the Shadow phase. Both F3 (text-dependency) and F5 (skipping difficult parts without returning to them) are symptomatic of the learner's inability to recall and pinpoint specific parts that need additional practice. R3 seeks an easy self-evaluation mechanism that makes such pinpointing possible.

R4. Minimize overwhelm during practice without lowering playback speed. The self-regulated learner "would not tend to satisfice or minimise effort...", and "...would not be confused or overwhelmed by learning tasks" [13]. However, contrary to this description, shadowing often overwhelms the learner (F5), and when overwhelmed, learners seek shortcuts to minimize effort (F4). This requirement, therefore, seeks to reduce overwhelm during practice.

We do not lower playback speed because shadowing improves listening skills by requiring the learner to meet the target narrator's pace [6]. R4 has been identified and addressed in *WithYou* using speech recognition-based dynamic pause modification [56]. We explore a self-regulated variant of the pause-modification idea to see if learners are able to adjust pauses by themselves without relying on an external mechanism that makes these decisions on their behalf.

4.1 Structuring the Self-Regulated Shadowing Process

We manifest the overarching requirement (R0) in our design by structuring the SRS process into two phases: Listen and Shadow[2], structuring each phase around two interleaved modes: Practice and Reflect (see Figure 3). Playing the target audio triggers Practice, whereas pausing triggers Reflect, because we see pauses as opportune moments for self-reflection and forethought.

CAST offers four core design elements (D1-4, see Figure 1) that bind and support the four phase-mode combinations: In-the-moment Highlights (D1) are mnemonic devices for annotating, tracking, and visualizing easy/difficult parts of the passage. Contextual blurring (D2) selectively blurs and reveals parts of the transcript to minimize text-dependency through assisted self-control. Self-listening comparators (D3) preserve recorded chunks of shadowing practice that are overlaid on the transcript to allow learners to make close comparisons between the transcript and self-recordings. Adjustable pause-handles (D4) are used to introduce short breaks between punctuated chunks and sentences.

(1) In Listen-Practice, the learner listens to the target audio, and highlights difficult portions (D1). In doing so, they form a visual map of areas where they anticipate difficulty during Shadow-Practice, leaving them better prepared.
(2) In Listen-Reflect, they pause to review their highlights from Listen-Practice, and introduce short breaks between difficult chunks by adjusting pause-handles (D4).
(3) In Shadow-Practice, they focus solely on shadowing because any other activities during Shadow-Practice will distract them and impede their learning. CAST quietly preserves their practice in the background in the form of self-listening comparators (D3).
(4) In Shadow-Reflect, they engage in post-practice self-evaluation, by using the self-listening comparators (D3) created during Shadow-Practice in combination with in-the-moment highlights (D1), to listen to themselves and update their visual map with highlights.

[2]Note on terminology: In Section 4, we reserve "Shadowing" for the technique as a whole, and "Shadow" for the specific phase.

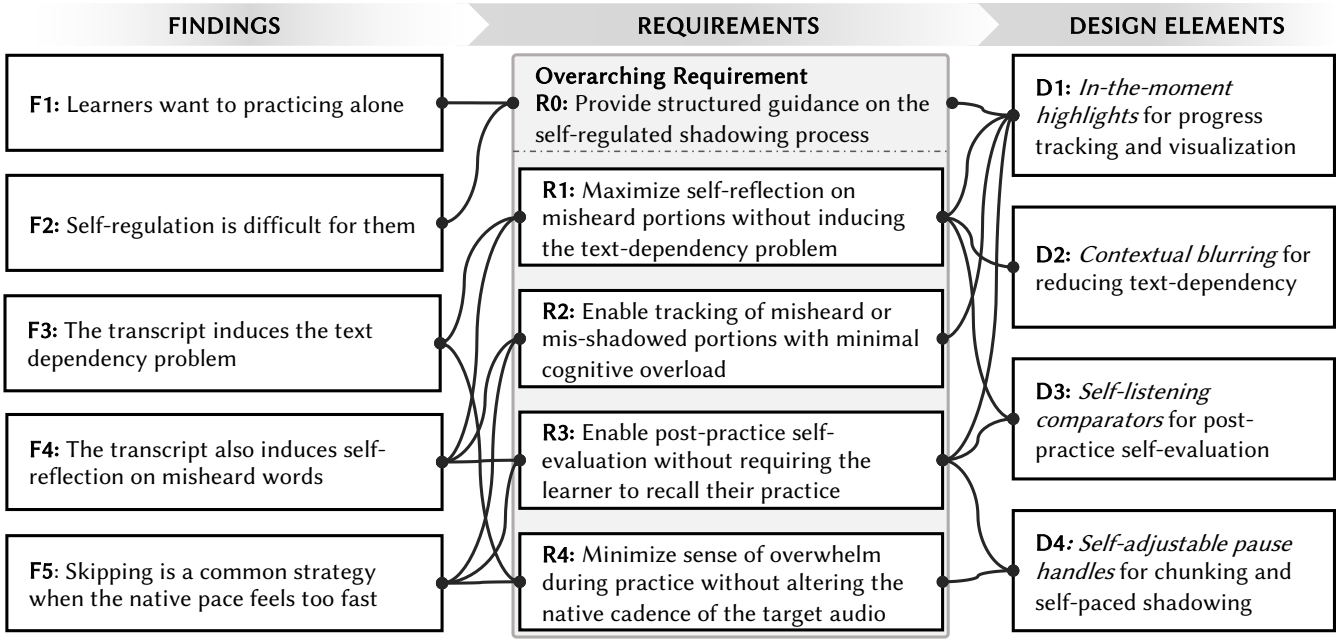

Figure 2: How the findings from our exploration of learner needs translate into requirements and design elements of CAST

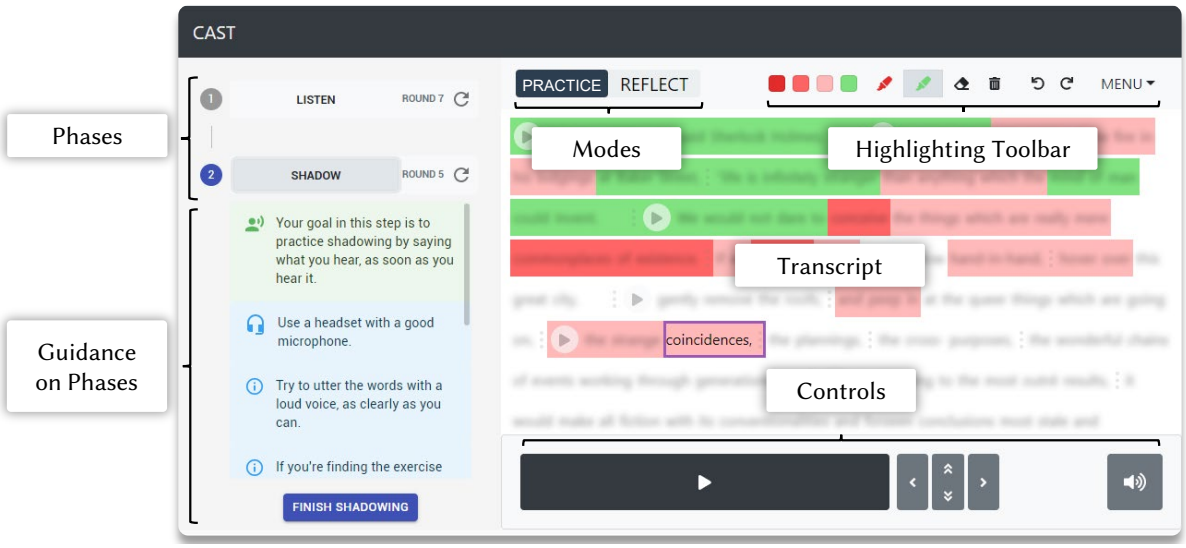

Figure 3: The layout design in CAST is structured around phases and modes. The user interacts with the transcript and the target audio using a set of controls and a highlighting toolbar.

4.2 Resolving text-dependency with Contextual Blurring

We maximize self-reflection while minimizing text-dependency (R1), and partially enable tracking of misheard words (R2) by combining highlights (D1) with contextual blurring (D2). With D1, learners track misheard words as they listen by highlighting them. Learners highlight words after they hear something they don't understand, i.e. the word they want to mark falls before the current point in the audio. Therefore, the audio is automatically paused the moment the learner begins highlighting, and resumed when they finish. This allows the learner to go back and check the transcript to see if they misheard something. They can make simultaneous references between the transcript and the audio by clicking on the word to connect the text with the sounds.

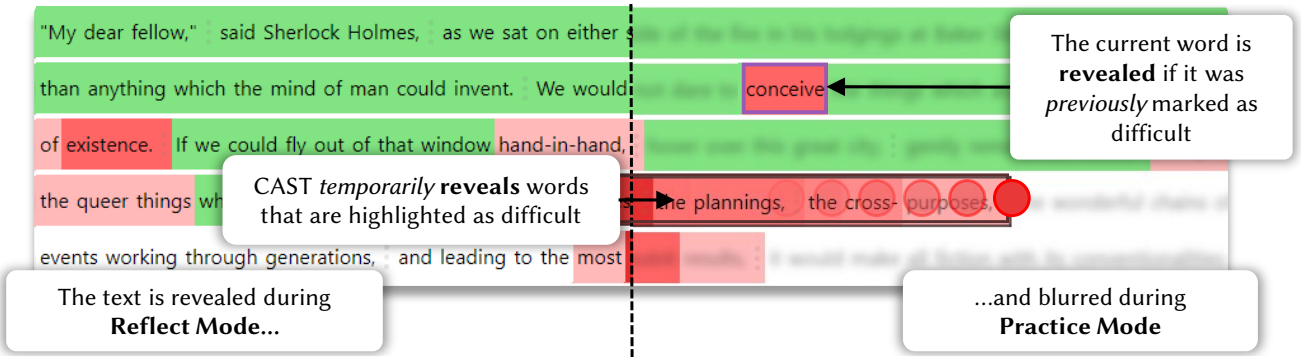

Figure 4: *Contextual blurring* and *in-the-moment highlights* work together to solve the text-dependency problem.

D1 on its own does not induce self-reflection on misheard words because of the text-dependency problem (If learners read as they listen, or read in advance, they do not notice misheard words). We resolve this problem by combining D1 with D2. To retain focus on listening, D1 blurs the transcript whenever the target audio is played. To induce self-reflection on misheard words during listening, D1 works together with D2 to reveal difficult parts while they are being highlighted by the learner, i.e. the moment the learner lets go of the highlight, it transitions back to being blurred. If the learner highlights something and holds on, the text slowly transitions back to being blurred to prevent learners from fixating on reading a specific part for too long, and to encourage them to continue listening. Learners know where to mark even when the text is blurred by marking in relation to a time-synchronized moving word marker.

4.3 Tracking Mis-shadowed Words with Self-Listening Comparators

We complete the tracking process (R2), and enable post-practice self-evaluation (R3) by introducing self-listening comparators (D3). R2 is partially supported by D1 and D2 because the highlighting process afforded by them is doable during listening, but not shadowing. Early on in the iterative design process, we tested the idea of doing highlights while shadowing, but this proved to be too challenging for learners, likely due to the significantly higher cognitive load for shadowing compared to listening [24]. Therefore, to fully support R2, we enable tracking of mis-shadowed words in addition to misheard words by introducing self-listening comparators (D3). D3 works in the background to record the learner whenever they shadow, and inserts these recordings into the transcript in a manner that allows for tight comparisons between the recordings and the transcript. When the learner plays a recording, the time-synchronized word marker moves at the rate of the target audio, to allow learners to observe any difference between their rate and the target rate.

There are subtle, but notable differences between the way D1 works in Listen-Practice and Shadow-Reflect. In the latter, the audio *isn't* pause during highlighting because learners listen to *themselves*, as opposed to the target audio. This enables them to spot-check the current word and update their highlights as they listen. During Listen-Practice, the learner uses D1 to map out words they perceive as misheard. In Shadow-Reflect, when they use D1 in combination with D2, they are confronted with how they actually shadowed. The D1 mapping process in Listen-Practice helps with forethought before shadowing, whereas the mapping process in Listen-Reflect helps with post-practice self-evaluation of listening ability after shadowing. Updating the map by iterating through successive rounds of Listen-Practice and Shadow-Reflect allows the learner to adapt their practice through self-monitoring progress because the map reflects their current state, and signifies where to focus in the next round, and encourages them to keep practicing until they can confidently mark everything in green.

4.4 Reducing Overwhelm and Fixation with Adjustable Pause-Handles

To reduce overwhelm during practice (R4), we introduce adjustable pause-handles (D4) that integrate nicely with the existing workflow (see Figure 1). Pause-handles are placed between punctuation marks and periods in the transcript because pausing at those points does not alter the native cadence. Learners use the visual map resulting from D1 to decide where to pause. Pausing before a difficult chunk reduces overwhelm by providing learners with a short break, whereas pausing after a difficult chunk prevents them from fixating on previous difficulties by grounding them back to the present moment. The current pause is signified by a pulsating marker that encourages learners to breathe.

5 EVALUATING CAST

We wanted to test whether the inclusion of our design elements positively impacted the learner's ability to self-regulate their shadowing practice using the transcript while retaining a strong focus on listening.

5.1 Method

To validate our design, we conducted a summative evaluation using a baseline interface as a reference. This baseline included features that are typically found in media players and document viewers, i.e., play/pause button, volume control, slider for audio navigation, and the ability to view the transcript. We removed tangential differences between the study conditions that could potentially confound our results by maintaining the same overall visual layout of the common

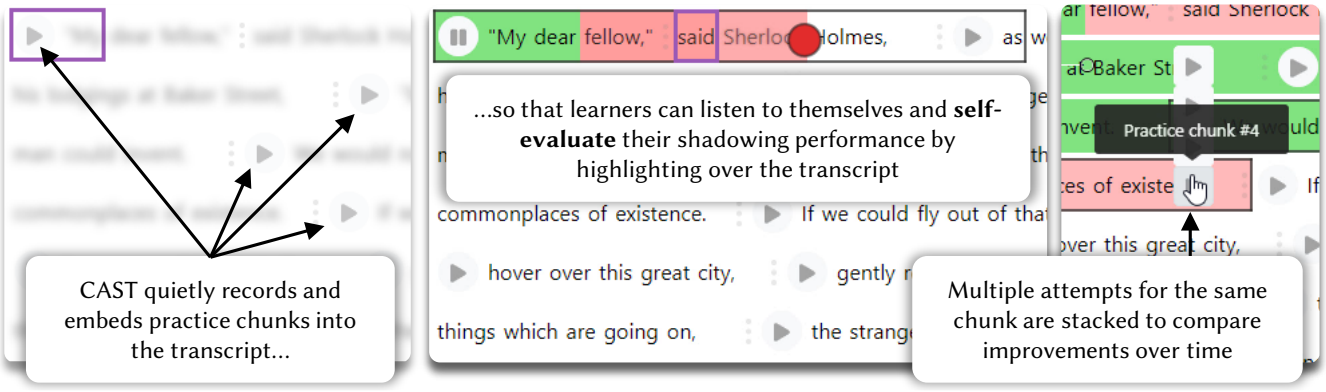

Figure 5: Self-listening comparators support post-practice self-evaluation through comparisons between self-recordings and the text.

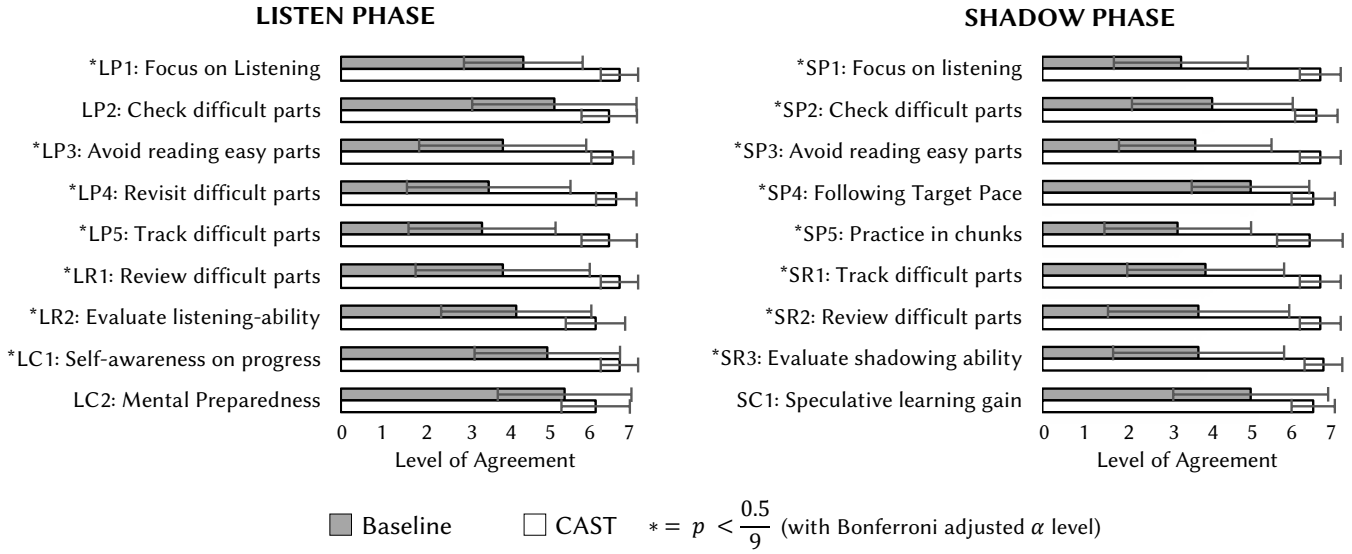

Figure 6: CAST enhances the learner's ability to self-regulate their listening-focused shadowing practice using the transcript.

UI elements (the placement and dimension of buttons and text, font size, and color) in both interfaces.

We designed our measures to cover three user-experience dimensions tied to both self-regulation and shadowing: (i) the relative attention between listening and reading (4 measures from Figure 6: SP1 & 3, LP1 & 3), (ii) the ability to identify, reflect upon, and correct shadowing mistakes through self-reflection and self-evaluation without feeling overwhelmed (11 measures from Figure 6, all excluding LC1 & 2 and SC1), and (iii) mental preparedness and speculative listening improvement (3 measures from Figure 6: LC1 & 2 and SC1).

We ran 4 pilots to iterate on our measures and used the feedback we received to ensure that our participants understood the questions. Our primary focus was to investigate the comparative advantages of CAST over baseline. We asked the same set of questions for each condition, and used one-item measures as opposed to a multi-item scale. The simplicity afforded by the one-item measures made it easier for participants to compare the the conditions.

5.1.1 Participants. The inclusion criteria for participants was the same as our need-finding study. We recruited a new batch of 12 ESL students aged 18 to 34 (7 women, 5 men) through purposive sampling to ensure that they spoke a variety of L1s. We did not include participants from our previous study to counter potential demand characteristics.

5.1.2 Tasks. Each participant finished two shadowing tasks, one with baseline and the other with CAST. The features of CAST work in tandem. For example, the highlights (D1) are taken into account by the contextual blurring feature (D2), and pause-handles (D4) delimit recorded chunks created by the comparators (D3). Therefore, we designed our tasks to give participants a sense of how the each

interface worked as a whole, instead of presenting each feature in isolation. Interface order was fully counterbalanced.

5.1.3 Materials. To provide an ecologically valid learning experience, we used four real-world articles as shadowing material. These articles were on two different topics (Science and Movies), to minimize the chance of domain interest affecting shadowing performance. The order of the topic was counterbalanced. We chose these articles based on relevance to the learner's real life, word variety, and unfamiliarity (i.e., passages that the participants did not know in advance).

5.1.4 Procedure. We conducted the study remotely over a 1.5 hour, recorded video-conference call. Doing the study online helped us reach international ESL participants, and enabled us to simulate the experience of practicing alone as closely as possible. We began the study by introducing the self-regulated shadowing process, and encouraged participants to reflect on their mistakes for *both* tasks for a fair comparison. For each task, participants completed a Likert-scale based questionnaire once after finishing the Listen phase, and once after finishing Shadow. We used this questionnaire to gain insights on where the learners were focused during each phase, and to see whether they were able to self-regulate their learning through self-monitoring, self-evaluation, self-reflection and self-pacing. We concluded the evaluation with a 10 minute semi-structured interview.

5.1.5 Analysis. For paired-comparisons, we opted for a Wilcoxon signed-rank test. We were interested in testing whether CAST offered significant improvements over baseline, and so we chose a one-tailed test with $H_0 : B > C$. To minimize chances of committing a Type I error, we applied the Bonferroni correction ($\alpha = \frac{.05}{9} = .00\overline{55}$ because there were 9 tests for each phase).

5.2 Results

> *"The experience [with CAST] is pretty amazing, actually. With the first version [baseline], I had some difficulties with keeping track of where I am, and to find the hard parts. The text was making it very difficult to focus on the audio. But with the second version [CAST], it was very convenient ... I especially liked the ability to track parts ... and because the text was blurred, I could focus on the audio ... also, the 'double-highlighting' feature, where I could mark difficult words in deeper shades of red, helped me practice those parts more than once ... as I am not a native speaker, I couldn't keep up with the pace [with baseline], so dividing the passage into chunks [with pause handles] was pretty amazing."* — P10

The overall response to CAST, as exemplified by P10's comment, was largely positive, with 15 out of the 18 indicators in our Likert-scale questionnaire (see Figure 6) showing statistically significant improvements over baseline ($p < \frac{.05}{9}$, $d > 1$) in terms of the learner's ability to focus on listening, and to self-regulate their shadowing practice.

In the following sections, L and S refers to the Listen and the Shadow phase, whereas P and R refers to the Practice and Reflect mode. C refers to the stage after phase completion. For example, LP1 refers to the first question about the learner's experience during Listen-Practice, whereas LC1 refers to the first question after competing the Listen phase.

CAST improves the learner's ability to focus on listening (LP1, SP1): We note a significant improvement in the learner's ability to focus on the audio during both listening (LP1: $p < .001$, $d = 1.896$) and shadowing (SP1: $p < .001$, $d = 1.996$). This is because contextual blurring in CAST was very well received, and learners appreciated the ability to use the transcript without feeling distracted by the text, which was a recurring issue with baseline.

> *"I think the first one [CAST] is much better because I can focus on listening more than reading. In the second one [baseline], I feel like I am reading the text but I am not hearing what the speaker is saying"* — P14

> *"...when the text becomes blurred you're not distracted by the other words"* — P9

CAST improved the ability of learners to check difficult parts during shadowing, and enabled them to avoid unintentional glances at surrounding text (LP2, LP3, SP2, SP3): Since in baseline, the transcript is always visible, and the Listen phase does not require too much effort, checking difficult parts while listening was doable with both versions (LP2: $p = .044$, $d = .542$).

However, without a moving word marker and contextual blurring, participants had to rely on skimming to find difficult parts with baseline, which was cognitively demanding for them. In CAST, such skimming is not necessary, and hence we see a notable improvement in the learner's ability to check difficult parts while shadowing (SP2: $p = .001$, $d = 1.161$). Furthermore, without contextual blurring, checking difficult parts forced participants to make unintentional glances at surrounding portions of the passage, even when they wanted to avoid reading those parts and to focus on listening. Once again, CAST resolved this issue with contextual blurring (LP3, SP3: $p < .001$, $d > 1$)

CAST makes it easy to track, review, and read difficult parts (LP4, LP5, LR1, SR1, SR2): In-the-moment highlights made tracking difficult parts during listening practice (LP5: $p < .001$, d = 1.526) and shadowing reflection (SR1: $p < .001$, d = 1.287) effective. With all the difficult parts highlighted over the blurred transcript, participants could easily use the moving word marker and transcript-driven audio navigation features to revisit (LP4: $p < .001$, $d = 1.38$), review (LR1: $p < .001$, $d = 1.259$), and redo (SR2: $p < .001$, $d = 1.26$) those parts until they mastered them.

CAST enables and enhances post-practice self-evaluation (LR2, SR3): When learners pause to reflect on their listening and shadowing ability, having a visual map of areas to focus significantly improves their ability to evaluate how well they were able to listen (LR2: $p < .001$, $d = 1.108$).

Comments from our participants confirm that they cannot easily remember how well they were able to shadow, nor can they do in-the-moment highlights during shadowing.

> *"You can't remember what you spoke...that's why [self-evaluating with baseline] wasn't good."* — P5

> *"It is too much to mark and shadow at the same time."* — P13

Therefore, our evaluation results confirm that combining in-the-moment highlights with self-listening comparators makes post-practice self-evaluation possible and effective (SR3: $p < .001$, $d = 1.467$).

Pause handles enable learners to match the target pace by make chunking significantly easier, and prevent them from fixating on hard words (SP5, SP4): By adjusting the pause handles, learners found it significantly easier to break down the passage into meaningful chunks with CAST (SP5: $p < .001$, $d = 1.627$). One of the reasons why we designed these pause handles was to enable learners to match the target pace without altering the native speed of audio, and we can confirm that pause handles achieve this purpose. (SP4: $p < .001$, $d = 1.141$) In addition to matching the target pace, comments from P4 and P9 indicated that the pause handles provided an unexpected additional benefit — it stopped them from fixating on difficult words. While shadowing, when learners come across a difficult word, thinking too hard about their past shadowing performance can adversely impact their future performance and learning. The pause handles introduce small breaks that give learners a moment to reflect on past performance and move on.

CAST heightens self-awareness on progress but does not impact learner's confidence level before shadowing (LC1, LC2): The visual mapping process supported by in-the-moment highlights give learners a clear and complete idea of all the hard and easy parts of the passage (LC1: $p < .002$, $d = 1.055$), thereby heightening their self-awareness on progress. However, this did not impact how mentally prepared they felt to begin shadowing after completing the listening phase (LC2: $p = .010$, $d = .777$). While we do not have data on the specifics of why mental preparedness was not impacted, we can say that knowing which areas need more work may not make learners feel better about their shadowing ability.

Learning gain remains an open question (SC1): For the given duration of practice (approximately 15 minutes for each phase), and the single session over which participants used the two interfaces, the difference between self-reported pre and post-task learning gains was not statistically significant. The original p-value for the speculative learning gain showed only a weak trend (SC1, $p = .009$, d = .797), and there's the possibility that Bonferroni adjustment may have induced a Type II error.

Comments from participants (P1, P2, P4, P9) showed the promise of longer-term learning gain with CAST over baseline:

> *"[With CAST], I know where I am not doing well...If I can clearly identify where I'm struggling with, I can repeat it to make sure I can do it better next time."* — P1

It is also worth noting that previous shadowing studies concerned with learning gain typically span multiple sessions and involve a large number of participants (see examples of shadowing research on page 25 of [20]). Therefore, long-term learning gain with CAST remains an open question, and can form the basis of a future study of that nature.

6 DISCUSSION

In this section, we reflect on the generalizability and implications of our findings.

6.1 The Practice-Reflect Model for Multimedia-based Self-Regulated Learning

In CAST, we introduce the practice-reflect model, where we dynamically adjust content representation (blur/unblur) depending on the phase, (listen/shadow) and mode (practice/reflect) of the learning context (shadowing). The purpose of this model is to guide the learner through the metacognitive processes necessary for self-regulated learning. We can apply this model to different multimedia-based learning contexts by defining the phases and modes specific to those contexts. For example, if the goal is to learn how to solve a math problem, the supporting content can be a step-by-step solution to the problem. A self-regulated system based on our practice-reflect model can reveal parts of the solution only in contexts where doing so will make the learner self-reflect on their mistakes. Akin to the transcript-dependency problem, providing unconstrained access to the entire solution in advance will make the learner too reliant on it, and therefore, this model is useful here.

This model is applicable even in contexts where the learning goal conflicts with shadowing. Say we want to learn the Iliad by heart (a famous epic poem with 15,693 lines [30]). The learning goal between shadowing and memorization are flipped — in the former, memorization is a vice because it removes the need for learners to rely on listening to decode words, in the latter, memorization is the goal. Instead of beginning with a blurred document, we may begin reading from an unblurred document, and highlight parts where we feel confident to blur them. Using what we know about memory retention over time, we can figure out the contexts where revealing blurred parts can help the learner. For example, we can apply an algorithm that uses the classic Ebbinghaus forgetting curve [11] as a basis. This curve suggests that we tend to continually halve our "memory of newly learned knowledge in a matter of days or weeks" unless we "actively review the learned material" [15]. Therefore, revealing portions that require review based on that curve can help us define the contexts in contextual blurring.

6.2 Applying the CAST Design Approach to Other Languages

> *"I'm Chinese and for us, we don't get praised for doing well, we just want to correct all of our mistakes."* — *P1*

Returning to our localized learning context of foreign-language listening practice, we chose English because it is of interest to a very large group of language learners. However, neither shadowing, nor self-regulated learning are exclusive to English pedagogy, and therefore, the overarching design concepts embodied within CAST can be generalized for the acquisition of other foreign languages. Most of the components can be used as-is, with little to no modification. For example, contextual blurring is applicable as long as the language has a written script that is supported by the computer. The same can be said about the process of self-evaluation through comparisons between the self-listening comparators and the transcript. Some of the features require additional forethought. For example, if pause handles are to be used, we must reconsider what constitutes a meaningful chunk in the target language because punctuation marks such as commas and periods are not universal. Furthermore, culture can influence design in unexpected, albeit significant ways, and culture is inextricably linked with language. For example, P1

from our evaluation study avoided marking parts as complete, and focused solely on identifying parts that she *couldn't* do yet. When asked why, she noted that in her culture, it is commendable to focus on areas of improvement rather than areas of achievement, and her cultural lens shaped how she used in-the-moment highlights.

6.3 Differentiating between Consumption and Learning

> *"[With baseline], it's very useful if you just want to hear a story or say you are on the bus...but it won't help you to learn English or practice my listening...[wtih CAST] it's interactive, and you're spending more time, and you are involved in learning..."—P9*
>
> *"...going through all parts until I could highlight everything in green...I felt like going over it again and again...but with the other one [baseline], there's no progress to be made..."—P11*

We can understand why CAST enhanced self-regulated learning by considering a fundamental distinction between tools designed for consumption, and tools designed for learning. When the learner uses a media player and document viewer for shadowing, it is easy for them to consume the audio and the text, but it is not necessarily easy for them to engage with the content in a manner that makes them reflect on their consumption. From this perspective, we can view the features offered by CAST as mechanisms for engaging and interacting with the material in a structured manner that helps them transition from the role of a content consumer to the role of a content learner. This is reflected in the comments from P9 and P11 on why they engaged more deeply with the shadowing material using CAST.

7 FUTURE WORK

7.1 Supporting Self-Regulated Speaking Practice

While the CAST design process focused on listening-focused shadowing, some of our findings and design elements are applicable to speaking-focused shadowing. Findings F1 (practicing alone), F2 (need for self-regulation) and F5 (skipping hard parts), remain applicable, as they are tied *how* shadowing is done. D1 (in-the moment highlights) and D4 (adjustable pause-handles) remain useful for tracking and chunking. Self-listening Comparators (D4) must be modified into Speaking Comparators by changing the standard of comparison from the transcript to the target narration. Comparing two audio streams can be tricky, however, because audio is linear. D2 (contextual blurring) can still help with speaking since closely matching the sounds in the target audio requires careful listening. Validating, modifying and extending the design elements of CAST for other learning contexts such as speaking is a potential avenue for future work.

7.2 Component-wise Validation and Long-term Learning Gain

Our evaluation allowed for a *holistic* validation of our design approach, but how learners interact with our design components individually, and the nature of any potential learning gain resulting from long term use remain open questions. Therefore, studies on component-wise validation and long-term learning gain can also form the basis for future work.

8 CONCLUSION

In this work, we introduced CAST, a novel shadowing-based language learning system for self-regulated listening practice. We explored the needs and challenges of learners through a formative user study with ESL students (N=15), and found that they want to practice alone, but are unable to self-regulate their shadowing practice. We also found that the transcript induces self-reflection on misheard words. We used these findings to develop an ensemble of four design elements that include contextual blurring, in-the-moment highlighting, listening comparators, and adjustable pause handles. We validated our design elements through a summative evaluation study (N=12), that showed learners were successfully able to self-regulate their listening-focused shadowing with CAST.

9 ACKNOWLEDGEMENTS

We would like to thank Dr. Bryan Gick and Dr. Strang Burton from UBC Linguistics for their helpful feedback and guidance on the project, Misuzu Kazama from UBC Asian Studies for providing insights on second-language acquisition from an instructor's perspective, and Fatimah Mahmood from UBC Vantage College for help with student recruitment. This work was partially supported by the NSERC Discovery Grant and CREATE programs as well as the generous gift from Adobe Research.

REFERENCES

[1] David Abercrombie. 1967. *Elements of General Phonetics.* Edinburgh University Press, Edinburgh. http://www.jstor.org/stable/10.3366/j.ctvxcrw9t

[2] Ian Arawjo, Dongwook Yoon, and François Guimbretière. 2017. TypeTalker: A Speech Synthesis-Based Multi-Modal Commenting System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1970–1981. https://doi.org/10.1145/2998181.2998260

[3] Barry Arons. 1997. SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 4, 1 (1997), 3–38.

[4] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the Use of Reflection in Interactive System Design. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[5] Jill Boucher. 1981. Immediate free recall in early childhood autism: Another point of behavioural similarity with the amnesic syndrome. *British Journal of Psychology* 72, 2 (1981), 211–215.

[6] Nicholas Bovee and Jeff Stewart. 2009. The utility of shadowing. In *JALT 2008 Conference Proceedings. Tokyo: JALT*. The Japan Association for Language Teaching, Tokyo, Japan, 888–900.

[7] Hui Chen, Agnieszka Ciborowska, and Kostadin Damevski. 2019. Using Automated Prompts for Student Reflection on Computer Security Concepts. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (Aberdeen, Scotland Uk) *(ITiCSE '19)*. Association for Computing Machinery, New York, NY, USA, 506–512. https://doi.org/10.1145/3304221.3319731

[8] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

[9] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[10] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (2017), 297–298. https://doi.org/10.1080/17439760.2016.1262613 arXiv:https://doi.org/10.1080/17439760.2016.1262613

[11] Hermann Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences* 20, 4 (2013), 155.

[12] Elfiz Media. 2019. *Shadowing - English Speaking Exercise.* Bilteksoftware. https://play.google.com/store/apps/details?id=com.pinholesoftware.shadowing&hl=en

[13] Christina J Evans, John R Kirby, and Leandre R Fabrigar. 2003. Approaches to learning, need for cognition, and strategic flexibility among university students. *British Journal of Educational Psychology* 73, 4 (2003), 507–528.

[14] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on Reflection: Framing a Design Landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 216–223. https://doi.org/10.1145/1952222.1952269

[15] Sandie Gay, Michelle Bishop, and Stuart Sutherland. 2016. Chapter 8 - Teaching Genetics and Genomics for Social and Lay Professionals. In *Genomics and Society*, Dhavendra Kumar and Ruth Chadwick (Eds.). Academic Press, Oxford, 147 – 164. https://doi.org/10.1016/B978-0-12-420195-8.00008-2

[16] Sten Govaerts, Katrien Verbert, Erik Duval, and Abelardo Pardo. 2012. The Student Activity Meter for Awareness and Self-Reflection. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 869–884. https://doi.org/10.1145/2212776.2212860

[17] Yo Hamada. 2012. An effective way to improve listening skills through shadowing. *The language teacher* 36, 1 (2012), 3–10.

[18] Yo Hamada. 2014. The effectiveness of pre-and post-shadowing in improving listening comprehension skills. *The Language Teacher* 38, 1 (2014), 3–10.

[19] Yo Hamada. 2016. Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. *Language Teaching Research* 20, 1 (2016), 35–52. https://doi.org/10.1177/1362168815597504 arXiv:https://doi.org/10.1177/1362168815597504

[20] Yo Hamada. 2016. *Teaching EFL Learners Shadowing for Listening: Developing learners' bottom-up skills.* Routledge, Abingdon, UK.

[21] Yo Hamada. 2019. Shadowing: What is It? How to Use It. Where Will It Go? *RELC Journal* 50, 3 (2019), 386–393.

[22] Debby Hindus, Chris Schmandt, and Chris Horner. 1993. Capturing, structuring, and representing ubiquitous audio. *ACM Transactions on Information Systems (TOIS)* 11, 4 (1993), 376–400.

[23] Kun-Ting Hsieh, Da-Hui Dong, and Li-Yi Wang. 2013. A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics* 11, 2 (2013), 43–65.

[24] Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology* 48, 3 (1995), 598–612.

[25] Taslim Arefin Khan, Dongwook Yoon, and Joanna McGrenere. 2020. Designing an Eyes-Reduced Document Skimming App for Situational Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376641

[26] Miri Kim. 2020. *English Shadowing: TEDICT.* Apple App Store. https://apps.apple.com/us/app/english-shadowing-tedict/id1455961007

[27] Ingrid Kurz. 1992. 'Shadowing'exercises in interpreter training. In *Teaching translation and interpreting*. John Benjamins, Amsterdam, 245.

[28] Peter Ladefoged. 2005. The Complete Story of Arthur the Rat in a British Accent. http://www.phonetics.ucla.edu/course/transcription%20exercises/peter.htm. Accessed on 2020-13-08.

[29] Sylvie Lambert. 1992. Shadowing. *Meta: Journal des traducteurs/Meta: Translators' Journal* 37, 2 (1992), 263–273.

[30] Richmond Lattimore, Leonard Baskin, et al. 1962. *The Iliad of Homer.* Cambridge University Press Archive, Cambridge.

[31] Maryanne Martin. 1977. Reading while listening: A linear model of selective attention. *Journal of Verbal Learning and Verbal Behavior* 16, 4 (1977), 453–463.

[32] Rob Martinsen, Cherice Montgomery, and Véronique Willardson. 2017. The effectiveness of video-based shadowing and tracking pronunciation exercises for foreign language learners. *Foreign Language Annals* 50, 4 (2017), 661–680.

[33] Richard E. Mayer. 2009. *Multimedia Learning* (2 ed.). Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511811678

[34] Neville Moray. 1959. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology* 11, 1 (1959), 56–60.

[35] Tim Murphey. 2001. Exploring conversational shadowing. *Language teaching research* 5, 2 (2001), 128–155.

[36] Julian Northbrook. 2013. English Speaking Practice | How to improve your English Speaking and Fluency: SHADOWING. https://www.youtube.com/watch?v=GVWFGIyNswI.

[37] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge University Press, Cambridge, UK.

[38] Ernesto Panadero. 2017. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology* 8 (2017), 422.

[39] Picup Inc. 2017. *Shadowing - English Speaking Exercise.* Apple App Store. https://apps.apple.com/us/app/shadowing-english-speaking-exercise/id1182789540

[40] Paul R Pintrich and Moshe Zeidner. 2000. *Handbook of self-regulation.* Academic Press, San Diego, California.

[41] Dale H Schunk. 1996. *Self-Evaluation and Self-Regulated Learning.* Technical Report. Graduate School and University Center, City University of New York, New York.

[42] Dale H Schunk. 2005. Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational psychologist* 40, 2 (2005), 85–94.

[43] Dale H Schunk and Peggy A Ertmer. 2000. Self-regulation and academic learning: Self-efficacy enhancing interventions. In *Handbook of self-regulation.* Academic Press, San Diego, California, 631–649.

[44] Hyungyu Shin, Eun-Young Ko, Joseph Jay Williams, and Juho Kim. 2018. Understanding the Effect of In-Video Prompting on Learners and Instructors. 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/10.1145/3173574.3173893

[45] Hideki Sumiyoshi. 2019. The Effect of Shadowing: Exploring the Speed Variety of Model Audio and Sound Recognition Ability in the Japanese as a Foreign Language Context. *Electronic Journal of Foreign Language Teaching* 16, 1 (2019), 8.

[46] Ken Tamai. 1992. The effect of follow-up on listening comprehension. *STEP Bulletin* 4, 1 (1992), 48–62.

[47] K Tamai. 1997. Shadowing no koka to chokai process ni okeru ichizuke [The effectiveness of shadowing and its position in the listening process]. *Current English Studies* 36 (1997), 105–116.

[48] Ken Tamai. 2005. Listening shidoho to shite no shadowing no koka ni kansuru kenkyu [Research on the effect of shadowing as a listening instruction method].

[49] Alice Thudt, Uta Hinrichs, Samuel Huron, and Sheelagh Carpendale. 2018. Self-Reflection and Personal Physicalization Construction. 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, 1–13. https://doi.org/10.1145/3173574.3173728

[50] Steve Whittaker, Julia Hirschberg, Brian Amento, Litza Stark, Michiel Bacchiani, Philip Isenhour, Larry Stead, Gary Zamchick, and Aaron Rosenberg. 2002. SCANMail: A Voicemail Interface That Makes Speech Browsable, Readable and Searchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 275–282. https://doi.org/10.1145/503376.503426

[51] Joseph Jay Williams, Tania Lombrozo, Anne Hsu, Bernd Huber, and Juho Kim. 2016. Revising Learner Misconceptions Without Feedback: Prompting for Reflection on Anomalies. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 470–474. https://doi.org/10.1145/2858036.2858361

[52] Jacqueline Wong, Martine Baars, Dan Davis, Tim Van Der Zee, Geert-Jan Houben, and Fred Paas. 2019. Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human–Computer Interaction* 35, 4-5 (2019), 356–373.

[53] Fereshteh Yavari and Sajad Shafiee. 2019. Effects of Shadowing and Tracking on Intermediate EFL Learners' Oral Fluency. *International Journal of Instruction* 12, 1 (2019), 869–884.

[54] Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review. 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, 481–490. https://doi.org/10.1145/2642918.2647390

[55] Dongwook Yoon, Nicholas Chen, Bernie Randles, Amy Cheatle, Corinna E. Löckenhoff, Steven J. Jackson, Abigail Sellen, and François Guimbretière. 2016. RichReview++: Deployment of a Collaborative Multi-Modal Annotation System for Instructor Feedback and Peer Discussion. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 195–205. https://doi.org/10.1145/2818048.2819951

[56] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2020. WithYou: Automated Adaptive Speech Tutoring With Context-Dependent Speech Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376322

[57] Barry J Zimmerman. 2000. Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation.* Academic Press, San Diego, California, 13–39.

[58] Barry J Zimmerman. 2002. Becoming a self-regulated learner: An overview. *Theory into practice* 41, 2 (2002), 67. https://doi.org/10.1207/s15430421tip4102_2