MemReflex: Adaptive Flashcards for Mobile Microlearning

Darren Edge[1]
[1] Microsoft Research Asia
Beijing, China

Stephen Fitchett[1,2]
[2] University of Canterbury
New Zealand

Michael Whitney[1,3]
[3] University of North Carolina Charlotte
USA

James Landay[1,4]
[4] University of Washington
USA

darren.edge@microsoft.com, stephen.fitchett@pg.canterbury.ac.nz, mwhitne6@uncc.edu, landay@cs.washington.edu

ABSTRACT
Flashcard systems typically help students learn facts (e.g., definitions, names, and dates), relying on intense initial memoriztion with subsequent tests delayed up to days later. This approach does not exploit the short, sparse, and mobile opportunities for microlearning throughout the day, nor does it support learners who need the motivation that comes from successful study sessions. In contrast, our MemReflex system of adaptive flashcards gives fast-feedback by retesting new items in quick succession, dynamically scheduling future tests according to a model of the learner's memory. We evaluate MemReflex across three user studies. In the first two studies, we demonstrate its effectiveness for both audio and text modalities, even while walking and distracted. In the third study of second-language vocabulary learning, we show how MemReflex enhanced learner accuracy, confidence, and perceptions of control and success. Overall, the work suggests new directions for mobile microlearning and "micro activities" in general.

Author Keywords
Mobile Flashcards; Adaptive Systems; Language Learning

ACM Classification Keywords
H.5 Information interfaces and presentation: User Interfaces

General Terms
Algorithms; Design; Experimentation; Human Factors

INTRODUCTION
The mobile phone is the ideal platform for long-term learning, being portable, individual, unobtrusive, available, adaptable, persistent, and useful [22]. In particular, mobile phones can support *microlearning* [13] in fragments of free time throughout the day. Previous work in HCI has examined how flashcards can support mobile microlearning of second-language phrases presented in context, investigating *what* material should be studied *where* [12]. However, relatively little attention has been paid to *when* items should be introduced and reviewed based on *how* the learner has performed in past microlearning sessions, especially in terms of how this relates to learner motivation.

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. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MobileHCI'12, September 21–24, 2012, San Francisco, CA, USA.
Copyright 2012 ACM 978-1-4503-1105-2/12/09...$10.00.

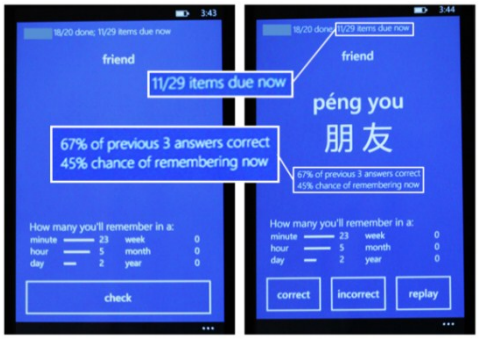

Figure 1. Adaptive flashcards. A cue (left) triggers recall of target information (right). Adaptive scheduling of cued recall tests raises response accuracies towards a goal level, e.g., 90%.

We believe mobile learning should be context-aware in a broad sense – sensitive to learner history as well as the immediate cues of time, location, and motion. We also believe that both text and audio interaction modalities are important to engage learners with different learning styles, as well as learners who move between contexts where different modalities are most appropriate (for example, it might be safer to listen via headphones when navigating busy public spaces, but politer to read from the screen in social situations where some conversation is anticipated).

This paper presents a systematic investigation of how learner *motivation* for microlearning using mobile flashcards is affected by adaptation to past performances, as well as how learner performance "on the move" is affected by the selection of interaction *modality*. The contribution to mobile HCI is a demonstration of how adaptive flashcards can support text and audio-based mobile learning even when walking, and how such adaptation helps drive learner motivation both in the moment and over the longer-term.

We begin with a literature review that motivates system features, before illustrating how existing flashcard systems do not account for the special characteristics of mobile microlearning. Next, we present the algorithm and interface design of our adaptive flashcard system, which we call MemReflex. These flashcards present cues from which learners attempt to recall the target information, with tests scheduled according to a model of the learner's memory (see Figure 1). We end with three user studies that show the effectiveness of MemReflex from the immediate to the longer-term, using either audio or text modalities, even while distracted by the demands of learning on the move.

RELATED WORK

In this section, we survey both the findings of learning research and the learning systems that aim to exploit them. Although we focus on approaches to learning that can be exploited in mobile contexts, our contribution itself is to the larger body of work on mobile learning or m-learning[1].

Learning Research

Learning is not a single process, but a hierarchy of processes reflecting progressive orders of change [2]. Zero order learning is characterized by responses without correction; first order learning by correction of errors within sets of alternatives; and second order learning by a change in the sets of alternatives or the distribution of first order learning over time (also known as "learning to learn").

Reviewing a physical flashcard is an example of first order learning: the front side of the card acts as a cue for the target on the reverse. When a learner attempts to recall the target given such a cue, the learning style is called *cued recall* and has been studied extensively in the learning literature. In contrast, our investigation of how to motivate learners to appropriate time for microlearning is a question of second order learning that has yet to be fully explored.

The testing effect

Much research into learning investigates and exploits the *testing effect* – that tests strengthen memory more than extra opportunities to study, even when mental retrieval is not accompanied by an outward response. Such test-directed learning can therefore take place in contexts where it is undesirable to produce overt responses, such as in public places. Moreover, it has been demonstrated across a variety of domains, including the learning of vocabulary in native and second languages, face–name associations, general facts, text passages, word lists, and even maps [6]. Cued recall tests of the form A→? have also been shown to enhance retention in the opposite direction B→?, as well as enhancing free recall of all cues (As) and targets (Bs) [7].

The spacing effect

The *spacing effect* is that when learning a set of items, superior retention results from multiple shorter presentations than from a single "massed" presentation. The time separating different study episodes of the same material is known as the inter-study interval or ISI. Studies of the spacing effect typically manipulate the ISI of two study episodes, and compare their effect on a later test that occurs after a fixed retention interval. In a review of 427 articles on cued recall learning, it was found that the optimal ISI increases as the retention interval increases [8]. For example, the optimum ISI for a 1-minute retention interval was less than 1 minute, whereas for retention intervals of 6 months or more it was at least 1 month. The implication is that multiple study episodes are needed for continuous retention, as shown in Figure 2.

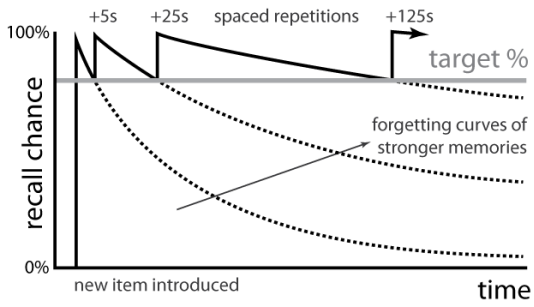

Figure 2. Adaptive spaced-repetition learning

The forgetting curve

The *forgetting curve*, discovered in 1885 [11], describes the inverse exponential nature of forgetting. In the presence of repeated, *spaced repetitions* as described above, the strength of a memory is increased, resulting in a more gradual process of forgetting. The most sophisticated psychological modeling of this process is derived from the ACT-R activation-based model of declarative memory, with each test of an item introducing a new memory trace whose decay rate is a power law function of all traces for the item at the time of the test [17]. In this model, the higher an item's activation, the smaller the effect any additional tests will have on its long-term retention.

Overlearning

Once a learner correctly recalls an item using cued recall, any further testing of the same item in the same session is described as *overlearning*. In two experiments and a review of the overlearning literature, Rohrer et al. [20] show that within a single learning session, doubling the number of tests for each item typically more than doubles the percentage of correct responses when tested again one week later. They suggest that such overlearning is necessary for short-term retention in situations such as preparing for an exam later in the day or learning foreign language vocabulary in advance of planned conversations. However, these benefits diminish for longer retention intervals.

Learning Systems

Since 1967, the dominant approach to audio language learning has been the Pimsleur System [19]. This is a series of 30-minute audio lessons in which basic vocabulary and phrases are introduced and reviewed in cued-recall fashion according to a schedule of *graduated interval recall*. This is a progressive series of exponentially expanding intervals (repetitions after 5^1s = 5 seconds, 5^2s = 25 seconds, 5^3s ≈ 2 minutes, 5^4s ≈ 10 minutes) both within and (roughly) across lessons. The advantages are that it uses deliberate overlearning to breed confidence and support immediate language use. The disadvantages of such lessons, however, are that repetitions are not scheduled in real-time, there is no adaptation to user feedback, lesson-length chunks of free time must be available, and ultimately the learner will run out of lessons. This latter difficulty is addressed by Gradint [14], which allows learners to make their own Pimsleur-like recordings using text-to-speech.

[1] See the m-learning website at http://www.m-learning.org/.

Developed at the same time as the Pimsleur system, the Leitner System [15] for physical flashcards is also based on spaced repetition. Using numbered flashcard piles P_1–P_N, each subsequent pile represents flashcards of both increasing memory strength and increasing inter-study intervals (e.g., on day D of study, test all flashcards in pile P_X where D is a multiple of X). Cards are promoted one pile if correct or else returned to the first pile for relearning. The advantage of this approach is that more difficult items are reviewed more often. The disadvantages are that there is no support for deciding when to introduce new flashcards into the first pile, and that the volumes of flashcards scheduled for review can quickly become unmanageable.

A computer-assisted approach to managing and scheduling the review of digital flashcards was popularized by the SuperMemo algorithm that targets a fixed retention rate [24]. In SuperMemo, the optimum interval I_R following a repetition R of an item is calculated as the product of I_{R-1} and the "optimum factor" for items of that "easiness" at repetition R. The table of optimum factors is updated over time, in response to achieved recall rates, aiming to converge on a long-term recall rate of 95%. Self-assessment on a 6-point scale updates the easiness of the current item and in the case of a low score, resets the repetition count of the item to be relearned. The advantage of adaptive spaced repetition in SuperMemo is that it evolves according to the learner's performance over time. The disadvantage is that it does not support initial learning, with the first test following item introduction not typically occurring for at least 5 days.

SuperMemo has inspired a whole family of algorithms, including the currently popular Anki system [1]. Anki has three distinguishing features: cards support multiple directions of cued recall between multi-attribute "facts" (e.g., the meaning, pronunciation, and appearance of a Chinese character); tests are scheduled each day at a fixed learner-specified rate; and after being tested on a card the learner selects one of four options indicating when they want to be tested on it next. These all increase learner load in terms of deciding which attributes to test, keeping up with scheduled sessions, and judging recall performance.

Finally, the FaCT system for Fact and Concept Training [18] uses the ACT-R cognitive model to predict the best item to test at any point in time, introducing new items when no existing items are near their optimum point for review. This aims to maximize the long-term recall gains of each review, rather than to keep the recall likelihood of all items above a threshold level. This system also incorporates structural models of the domains being studied; e.g., cued recall in one direction can have a calculated carryover on the probability of recall in the opposite direction. As with SuperMemo, however, the primarily visual content leads to a distinction between initial "study" trials – massed presentations of items to learn – and subsequent "drills" of cued recall tests. As such, it takes no account of shorter-term learner experience or of study for immediate use.

Summary

Cued recall learning is a powerful tool for acquiring the kind of domain knowledge that is often fundamental for participation in higher-level activities (e.g., in the domain of second language learning, vocabulary is a prerequisite for conversation). Since spaced repetition of cued recall tests enhances learner performance, cued recall is an ideal candidate for distributed sessions of microlearning.

Flashcards are the prototypical medium for cued recall, but existing systems impose time or structural constraints that do not satisfy the demands of mobile microlearning. Adaptive systems help, since matching perceived challenged to perceived skills can result in a psychological state of *flow* [10,16] as well as enhance learner motivation in the moment and over the long-term [3]. However, learner adaptation has typically focused on long-term retention rather than short-term experience, with little consideration for learners struggling to learn new items (e.g., a microlearning session could easily be dominated by failed attempts to correctly recall new items for the first time).

A more "micro" form of adaptation is thus required to facilitate positive experiences in even the shortest learning sessions, without requiring intense visual "memorization" of new items. The next section describes how we designed a system to support such adaptive mobile microlearning.

ALGORITHM DESIGN

The foundation of our algorithm is the exponential intervals (5 seconds, 25 seconds, 2 minutes, etc.) of the Pimsleur language system because of the uniform emphasis they place on short-term overlearning and long-term retention.

Our first refinement is to incorporate Leitner-style binary feedback from the user after a cue has been presented, indicating whether the target was recalled correctly or not. Such binary feedback is more suitable for lightweight mobile input than the more numerous choices used by SuperMemo and Anki and the text input used by FaCT.

Our second refinement, following SuperMemo, is to adaptively manipulate inter-test intervals according to learner performance, attempting to converge on a goal recall success rate of 90% at all stages of retention for any given item (as shown in Figure 2). In any session, given sufficient learning ability and appropriately spaced prior sessions, the learner should therefore be expected to correctly recall 90% of the items tested. This adaptive matching of challenges to demonstrated skills should help facilitate the short-term experience of flow. The corollary of this for longer-term motivation is that whenever the learner needs to, they should also be able to correctly recall and use 90% of all items they have ever studied.

Our third refinement, as in the FaCT system, is to identify opportune moments to introduce new items when there are no other items due for review. In our algorithm, such moments occur when there are no items with less than a 90% chance of being remembered at that point in time, according to our model of the learner's memory.

Our hypothesis is that this design will be successful at motivating learners through high recall accuracies whatever the learning modality, even when mobile microlearning.

Adaptive Spaced Repetition Algorithm

Our algorithm models a learner's knowledge of a set of items represented as cue-target pairs. For each item, a learnedness value l reflects the algorithm's estimate of the strength of the association of the cue-target pair at the time immediately following the learner's last response.

Exponentially-expanding inter-study intervals

The inter-study interval t_I is calculated based on l :

$$t_I = 5^{10l}$$

When the next item is required, the algorithm selects the most overdue item according to the ratio of actual time elapsed to ideal inter-study interval. If no items are due, a new item is introduced from a queue of items to learn.

For correct responses, l is incremented by a base increase value b. In our algorithm, we initially set b to 0.1 to replicate the Pimsleur intervals of 5s, 25s, 125s, etc. For incorrect responses, l is reset to this initial value of 0.1.

The human forgetting curve [11] models retention as:

$$E(recall) = e^{-t/s}$$

where t is the time that has elapsed since the last presentation of the item and s is the strength of the learner's memory for that item directly following its presentation. We can rewrite this equation to represent memory strength as a fixed scaling of the inter-study interval t_I as follows:

$$E(recall) = e^{-\frac{t}{t_I/\ln(10/9)}} = 0.9^{\frac{t}{t_I}}$$

The scale factor was chosen such that when items are tested on time ($t = t_I$), they have a 90% probability of success. Substituting another base in the final expression above adjusts the goal success rate (e.g., 0.95 would target 95%).

Adapting inter-study intervals to reach goal accuracies

The algorithm adjusts base increase values according to a learner's history across all items of similar learnedness. First, responses are grouped into buckets of size 0.1 based on their item's learnedness, each with their own base increase value. After a response to an item in a particular bucket, we compare the actual proportion of correct responses in that bucket to the goal proportion of 90%. If the correct responses for a bucket fall below this goal, it means the last increases in learnedness for items in that bucket were too large, resulting in excess forgetting. Conversely, a proportion of correct responses above this goal indicates that the last increases in learnedness were too small, resulting in insufficient forgetting and over-practice. To adapt to the learner's history following a response to an item, we therefore update the base increase value of the item's *previous bucket* in the direction and degree that would, *in retrospect*, have resulted in the 90% goal accuracy if it had been used after the last response.

To safeguard against inappropriate adaption in various instances, we set the following requirements for adaptation:

1. The item must have had at least a 60% recall probability (to adapt to forgetting, not delays between sessions).
2. The item's bucket must contain at least five responses (to provide a solid initial basis for adaptation).
3. The base increase value b of a bucket is confined by flexible bounds within the range [0, 0.2] (to avoid over-adjustment of b in the case of unusual response history).

Such stratified adjustments do not account for variations in item difficulty. We therefore added two refinements: first, "hard" items that have received multiple incorrect responses appear more frequently; second, "easy" items recalled many times before an error quickly return to their prior level if subsequent recall attempts are successful.

Simulations

We ran several stochastic discrete-event simulations of our learning algorithm. Results indicated that learning efficiency would increase with more and longer sessions, but with minimal gains after 40 repetitions per session or 4 sessions per day. Given a fixed number of daily repetitions, efficiency was highest when they were broken up into multiple sessions, supporting the strategy of microlearning.

MEMREFLEX FLASHCARDS

The interface design of our MemReflex mobile application communicates the underlying memory modeling of our novel adaptive algorithm. Figure 1 illustrates the primary interaction mechanic using second-language vocabulary taken from our third study: the learner attempts to recall the Chinese translation of "friend" before pressing "check" (left); On seeing and hearing the correct answer of "péng you", the learner indicates whether they were *correct*, *incorrect*, or would like to *replay* the audio pronunciation (right). They also see several key pieces of feedback:

1. The number of items due for review now, or else a countdown of when the next item will become due.
2. The learner's response history for the current item.
3. The estimated chance of the learner correctly recalling the current item, calculated from a memory model.
4. The estimated number of items that will be remembered in a minute, hour, day, week, month, and year.

This feedback, which updates in real-time, helps the learner understand how past interactions have helped to build up his or her memory and how this will be retained into the future. As they study, learners can see their knowledge shifting from shorter- to longer-term memory. When they are not studying, however, the same feedback shows the extent of forgetting that is expected to have occurred. Every 20 flashcard repetitions, the learner sees summary statistics of both that "microlearning" session and overall.

EVALUATION

In this section, we describe how we systematically broke down and validated the concept of adaptive flashcards for mobile microlearning. Our research questions were:

1. Do short, sparse microlearning sessions result in retention beyond the sessions themselves?
2. How does fixed progression, Pimsleur-like learning compare to adaptive learning?
3. Are both screen-text and eyes-free audio appropriate interaction modalities for mobile microlearning?
4. Does adaptive microlearning create new opportunities for short-term mobile learning, such as while walking?
5. How does mobile microlearning support motivation in the longer-term, e.g., for second language learning?

We addressed these five research questions with three user studies tackling questions 1–3, 3–4, and 1–5 respectively. In the first two studies, we use dates and ages as learning material to support controlled testing of precision recall for facts that are interesting but previously unknown. Since numbers are typically difficult to remember, reported results are potentially lower than what might be expected in non-numeric domains. In the third and final study we evaluate real use of the system for language learning.

Participants were recruited from international visitors to our lab as well as the local expatriate community. Backgrounds included engineering, design, teaching, PR, and admin.

All statistical analysis was conducted at significance level $p < 0.05$, with the Bonferroni correction applied to planned post-hoc comparisons. For clarity, we present results using the notation *sample-mean units* (*standard-deviation*) and show standard error bars on charts highlighting results. We also predefine the common measures used across studies:

1. *Items Introduced*. The total number of new items introduced to the learner by the learning algorithm.
2. *Repetition Accuracy*. The percentage of correct responses given by the learner during system use.
3. *Items Retained*. The total number of items correctly recalled on the post-test following system use.
4. *Retention Accuracy*. The percentage of items introduced during system use correctly recalled on the post-test.

STUDY 1: NON-MOBILE AUDIO MICROLEARNING

Our first study tested two hypotheses: that a generalization of the audio-only Pimsleur system could work for audio learning beyond second languages and that our adaptive algorithm could successfully adjust the spacing between repetitions to raise the accuracy of a learner's responses. For this initial study, we used a desktop rather than a mobile system so we could focus on algorithm performance and reduce variation due to differences in learning context.

Study Design

We recruited 14 participants (4 females) with a mean age of 25 for our study. Participants used a desktop application on Windows 7 to learn the years of inventions of technologies through audio-only cued recall. Sounds were synthesized using text-to-speech and on hearing the target, participants clicked a button indicating whether they had recalled it correctly. A *replay* button repeated the last sound.

Each participant used three algorithms but did not know multiple algorithms were being tested. The mapping of data to algorithm across users and algorithm order across users and sessions were counterbalanced. The algorithms were:

1. *Progressive*. Used a fixed progression of exponentially increasing intervals as in the Pimsleur system.
2. *Responsive*. As progressive, but with item relearning on incorrect responses (returning to the first interval).
3. *Adaptive*. Used adaptive intervals as described previously to raise success rates towards 90%.

Participants completed one session per weekday, resulting in 9 sessions over 11 days. Each took about 10 minutes and comprised 60 repetitions. We scheduled only 20 repetitions (around 2 or 3 minutes) per algorithm per day to induce the kind of forgetting and relearning that might be expected in real use. One week after the final session, participants were each given a post-test of all items introduced to them.

Results

Results are shown in Figure 3. We ran one-way repeated-measures ANOVA analyses over the factor of *Algorithm* (with *Progressive*, *Responsive*, and *Adaptive* levels) for the dependent measures of *Items Introduced*, *Repetition Accuracy*, *Items Retained,* and *Retention Accuracy*.

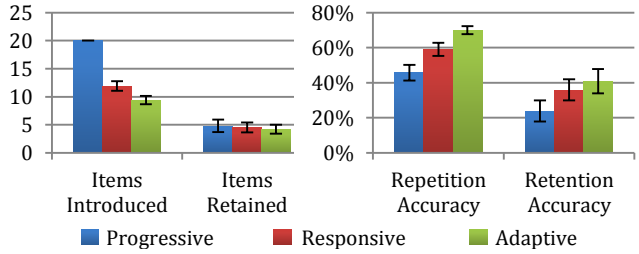

Figure 3. Main Results of Study 1

There were significant differences ($F_{2,26} = 91.8$, $p < 0.001$) in the numbers of *Items Introduced*, with means of 20 (0), 11.9 (3.2), and 9.4 (2.7) for the *Progressive*, *Responsive* and *Adaptive* algorithms respectively. There were also significant differences in *Repetition Accuracy* ($F_{2,26} = 38.0$, $p < 0.001$) with means of 45.7% (15.4), 59.0% (13.0) and 70.0% (7.9) for the *Progressive*, *Responsive* and *Adaptive* algorithms. In both cases, post-hoc tests showed pairwise differences between all three algorithms.

In the post-test 1-week after system use had ended, there were no significant differences in *Items Retained* with means of 4.8 (4.2), 4.5 (3.3) and 4.2 (3.0) for the *Progressive*, *Responsive* and *Adaptive* algorithms. There was a significant difference in *Retention Accuracy* ($F_{2,26} = 4.61$, $p < 0.05$), with means of 23.9% (20.9), 35.9% (20.9) and 40.9% (23.9). Post-hoc comparisons showed an advantage of *Adaptive* over *Progressive*.

Discussion

This first study validated our generalization of Pimsleur-like audio learning to other domains. It also demonstrated

the ability of an *Adaptive* system to raise recall success rates beyond those resulting from fixed progression along a series of expanding intervals (*Progressive*) or progression with relearning (*Responsive*). Although there was no significant difference between algorithms for the number of items recalled after the one-week retention interval, the fact that participants could precisely recall facts studied using all methods shows clear potential for the strategy of microlearning even when sessions are sparse.

STUDY 2: MOBILE LEARNING IN MOTION

An important faculty of mobile devices is their ability to support interaction "on the move" [23], and many opportunities for microlearning are likely to occur while the learner is walking. However, previous studies in mobile HCI have shown the negative effects of walking on visual selection and reading performance [21]. The trade-off between walking speed and interaction performance has also been investigated, with the finding that visual target acquisition plateaus at 40–80% of the user's preferred walking speed [4]. As an alternative to visual feedback, audio feedback while walking has been shown to result in higher interaction accuracy as well as lower mental and physical demands [23]. However, no prior work has investigated the effects of motion and modality on the effectiveness of mobile *learning*, rather than selection.

In this second study, we tested hypotheses about mobile microlearning with adaptive flashcards: that they can support high recall accuracy in both text and audio modalities, even while the learner is walking and distracted.

Study Design

We recruited 12 participants (5 females) with a mean age of 28 for our study. Participants used a Windows Phone 7 mobile application on an HTC HD7 to learn animal lifespans through either audio-only or text-only cued recall, with audio recordings provided by a native English speaker.

Data were divided into two sets balanced for complexity and paired with interaction modalities for each user. These pairings and the order of modalities between participants were counterbalanced. The two modalities were:

1. *Audio*. Participants hear audio cues and targets though headphones but see no item text on the screen.
2. *Text*. Participants see text cues and targets on the phone screen but hear no item audio.

The *Adaptive* algorithm from the first study was used for both because of its confirmed ability to control the rate of introduction of new items according to demonstrated learner performance. To support eyes-free interaction using audio, we implemented a button-free, gestural interaction style: *tap* to play the target of a cue, *flick left* if the learner could not recall or incorrectly recalled the target, *flick right* if the learner recalled the target correctly, and *flick down* to replay the last sound heard. For consistency, the same method was also used for text-only interaction.

The study began with a guided walk of an approximately 280 meter lap around an office floor, pointing out the 65 desk and room nameplates that are used in the experiment. This was followed by an introduction to learning through cued recall, using invention dates from the first study.

The first *Continuous* task for each interaction modality was to use it for 5 minutes of learning while walking in a 15m length figure-of-eight path around two pairs of adjacent chairs (a standard setup for interaction-while-walking tasks, e.g., as in [23]). Participants were instructed to focus on learning continuously, walking as fast as this would allow.

The second *Interrupted* task for each interaction modality was 5 minutes of learning while walking 280m laps around an office floor, performing the dual task of checking nameplates for a glanceable but unfamiliar property (colored dots attached to nameplates). By replicating the kind of cognitive load that might be experienced in a highly distracting environment, we could examine how this influenced interaction strategies and learning outcomes across the two modalities. Ten colored dots were added to randomly spaced nameplates, alternating between blue and green. Participants were instructed to check as many nameplates as possible in 5 minutes, always pointing out dots of a particular color and learning as continuously as this would allow. Colors were assigned randomly and switched for the second interaction modality. Each modality condition began with the learner training to proficiency in that modality. The first task, second task, and post-test then followed, each separated by a two-minute break.

Results

We ran repeated measures ANOVA analyses over the two factors of *Modality* (*Audio* and *Text*) and *Task* (*Continuous* and *Interrupted*) for each of the dependent measures *Items Introduced*, *Repetition Accuracy*, *Mean Repetition Duration*, and *Walking Speed*. *Mean Repetition Duration* was calculated by dividing the task time by the total number of repetitions, and *Walking Speed* by dividing the total distance travelled by the task time. We also used paired, two-tailed t-tests to compare the number of *Items Retained* and the *Retention Accuracy* as measured in the post-tests.

There were significant main effects of *Task* for all dependent measures. From the *Continuous* task to the *Interrupted* task, this represents fewer *Items Introduced* with $F_{1,44} = 145$, $p < 0.001$ from means of 10.8 (2.7) and 3.3 (1.5), longer *Mean Repetition Durations* with $F_{1,44} = 15.9$, $p < 0.001$ from means of 5.8s (2.1) and 8.5s (2.7), lower *Repetition Accuracy* with $F_{1,44} = 13.3$, $p < 0.001$ from means of 81.2% (9.2) and 70.5% (12.0), and slower *Walking Speed* with $F_{1,44} = 15.2$, $p < 0.001$ from means of 4.41km/h (0.64) and 3.82km/h (0.55).

For *Items Introduced*, there were no significant main effects for *Modality* or any interaction effects. For *Mean Repetition Durations*, there was a significant main effect for *Modality* with longer repetitions for *Audio* than *Text*, with $F_{1,44} =$

4.20, $p < 0.05$ from means of 7.8s (2.3) and 6.4s (3.0), but no interaction effect. For *Repetition Accuracy*, there was no significant main effect for *Modality* but there was a significant interaction effect ($F_{1,44} = 4.64$, $p < 0.05$). Post-hoc tests revealed that *Continuous Audio* learning with mean 86.3% (5.2) had significantly higher accuracy levels than *Continuous Text* learning with mean 76.2% (9.7), *Interrupted Audio* learning with mean 69.2% (14.7), and *Interrupted Text* learning with a mean of 71.8% (9.0). There was a significant main effect of *Modality* for *Walking Speed*, with *Audio* faster than *Text* ($F_{1, 44} = 14.3$, $p < 0.001$) from means of 4.40km/h (0.66) and 3.83km/h (0.55). These results best highlight the characteristics of mobile microlearning across modalities and are shown in Figure 4.

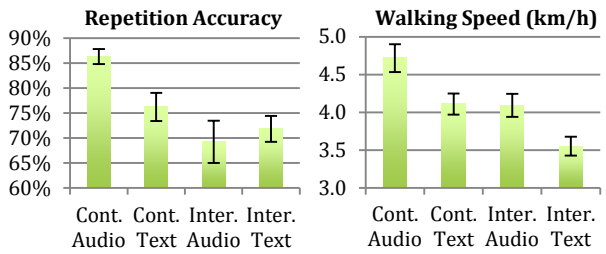

Figure 4. Main Results of Study 2

There was no significant difference in the *Items Retained* or the *Retention Accuracy* on the *Audio* and *Text* post-tests, with *Items Retained* means of 11.6 (3.7) and 11.7 (3.1) and *Retention Accuracy* means of 83.8% (19.4) and 83.6% (12.3). These accuracies represent the small amounts of forgetting below the 90% goal that would be expected during the 2-minute delay that preceded the post-tests.

Discussion

This second study demonstrated a substantial effect of interaction modality on the walking speeds of mobile learners, with eyes-free audio learning increasing walking speeds by 15% over text-based learning. The implication is that audiogestural interaction creates new opportunities for mobile learning while walking when compared with visual flashcard interaction, which typically requires many more repetitions for the same overall level of retention. This is supported by the fact that in our *Continuous* walking task, the mean walking speed with audio learning was as high as 88% of the benchmark 5km/h adult walking speed even while maintaining a mean repetition accuracy level of 86% (compared to 73% and 76% with text-based learning).

In the presence of visual distractions, audio learning maintains its advantage in terms of walking speed but loses its accuracy advantage since it becomes more difficult to regain audio context (by issuing a command to repeat the last sound) than text context (by glancing at the phone). Given that the need to pay visual attention to the environment is typically greater and more persistent than the need to listen, our *Interrupted* task was more representative of mobile navigation "in the wild". However, there is a relatively greater drop in *Repetition Accuracy* for *Audio* than for *Text* when moving from *Continuous* to *Interrupted* tasks (Figure 4), indicating a crossover point at which competing stimuli result in too much cognitive interference for productive audio learning. The conclusion is that we should allow real-time switching between modalities according to contextual demands.

Overall, the high levels of repetition and retention accuracy across both modalities in this second study suggest that adaptive flashcards can support short-term learning using either text or audio, even while attending to the demands of real-world navigation. Even in these sub-optimal learning conditions, especially when using audio-only in the absence of visual distractions, several participants remarked that they "got in the zone". Understanding how similar microlearning experiences can sustain motivation in long-term, high-value activities such as language learning is the question we address with the third and final user study.

STUDY 3: SELF-DIRECTED LANGUAGE LEARNING

Our third study tested hypotheses regarding longer-term, learner-directed system use for second language vocabulary learning: that adaptive flashcards can support high recall accuracy in this more challenging context; that such high accuracies are motivating; and that mobile microlearning is an effective strategy for second language learning.

Study Design

We recruited 12 Mandarin Chinese learners (3 females) with a mean age of 28 from our local expatriate community to participate in a 3-week long user study. Participants used a mobile application on an HTC HC7 Windows Phone 7 to learn Chinese words mined from the Web.

The vocabulary data were taken from the top 2000 two-character Chinese words online, divided into two sets balanced for frequency. To assess the value of adapting to learners' memories in ways that facilitate high recall accuracies, we compared the *Progressive* and *Adaptive* algorithms from study one. To support learner speech and exploit the fast flipping of text flashcards observed in study two, we used text flashcards with English definition cues and Chinese translation targets. To exploit focused attention to audio flashcards, we also played the pronunciation of Chinese words automatically when they appeared, as well as any time the learner pressed a button to replay the word.

Learners alternate between *Progressive* and *Adaptive* every 20 repetitions, which do not include the display of flashcard candidates whenever there are no items due. In such cases, the next most frequent word in the current set is presented as both definition and translation. The learner then chooses one of three options: to *learn* the word using the current algorithm, to indicate that they already *know* the word, or to *skip* the word if it is unappealing. At the end of each session of 20 repetitions, learners see a feedback screen of session statistics as well as overall statistics for both algorithms (simply called "A" and "B"). These are the statistics of

Items Introduced, *Test Repetitions*, and *Repetition Accuracy*. On this screen, learners give a subjective *Session Rating* of how the session felt on a scale of 1 (bad) to 5 (great). The mapping of dataset to algorithm and algorithm order were both counterbalanced across participants.

The *Progressive* and *Adaptive* interfaces shared the same basic elements: the session repetition count (out of 20), the number of items due or the time until an item will become due; the English definition cue and Chinese translation target; the item history as "X% of previous Y answers correct"; and the *check*, *correct*, *incorrect*, and *replay* buttons. The *Adaptive* interface (Figure 1) added two extra elements: an "X% chance of remembering now" estimate from the learner model and visualized estimates of how many words the learner should remember for how long.

Participants were free to use the flashcards as much or as little as they desired. We recommended that they complete at least one "A" and "B" session each weekday, or 300 repetitions per test in total, but compensation (a small gift) did not depend on this. A post-test of all items introduced and a semi-structured interview completed the study.

Results

Detailed results from study 3 are shown in Table 1. We ran two-tailed paired-sample t-tests between *Progressive* and *Adaptive* methods for the dependent measures of *Items Introduced*, *Repetition Accuracy*, *Items Retained, Retention Accuracy*, and *Mean Session Rating*.

User ID	Reps./ method	*Progressive method*					*Adaptive method*				
		II	RA	IR	IR%	MSR	II	RA	IR	IR%	MSR
P1	740	98	23	30	31	1.8	30	71	28	93	3.9
P2	640	95	14	11	12	1.1	19	64	17	89	3.1
P3	540	87	46	29	33	3.1	31	75	27	87	3.8
P4	520	75	53	41	55	3.0	22	78	21	95	3.7
P5	480	76	34	32	42	1.8	22	68	21	95	3.9
P6	400	70	42	20	29	2.3	13	63	9	69	3.3
P7	340	45	7	3	7	1.2	5	24	4	80	2.7
P8	320	53	13	8	15	1.9	9	58	6	67	2.7
P9	300	55	66	35	64	4.0	23	84	19	83	3.9
P10	260	49	29	11	22	2.0	17	68	16	94	3.7
P11	220	45	43	22	49	1.8	18	74	16	89	3.9
P12	160	32	27	13	41	2.6	7	44	7	100	2.4
mean (sd)	**410 (176)**	**65 (22)**	**33 (18)**	**21 (12)**	**32 (17)**	**2.2 (0.8)**	**18 (8)**	**64 (16)**	**16 (8)**	**87 (10)**	**3.4 (0.6)**

Table 1. Detailed Results of Study 3. II = *Items Introduced*; RA% = *Repetition Accuracy*; IR = *Items Retained*; IR% = *Retention Accuracy*; MSR = *Mean Session Rating* (1– 5)

From *Progressive* to *Adaptive*, there were significant differences in *Items Introduced* (t_{11} = 10.0, $p < 0.001$) from means of 65.0 (21.6) and 18.0 (8.4); *Repetition Accuracy* (t_{11} = 8.9, $p < 0.001$) from means of 33.1% (17.7) and 64.3% (16.3); *Retention Accuracy* (t_{11} = 11.5, $p < 0.001$) from means of 32.4% (17.1) and 86.9% (10.5); and *Mean Session Rating* (t_{11} = 4.9, $p < 0.001$) from means of 2.22 (0.84) and 3.42 (0.56). For *Items Retained*, there was almost a significant difference (t_{11} = 2.2, p = 0.054), with means of 20.9 (12.1) for *Progressive* and 15.9 (8.0) for *Adaptive*.

To summarize, the *Adaptive* method doubled learners' repetition accuracy over one quarter as many introduced items, improving the learner experience of each session. In addition, whereas *Repetition Accuracy* predicts *Retention Accuracy* for *Progressive* (means of 33% and 32%), for *Adaptive* this number increases from 64% to 87%. This again indicates stronger, more selective learning with *Adaptive*. All participants also preferred *Adaptive* and all but one (who would rather manage their own rote approach) would continue to use it for their second language learning.

To gain an insight into the possible long-term experience of using both methods, we plotted learners' session accuracies over time. Representative results are shown in Figure 5 (only the four heaviest users are shown for clarity).

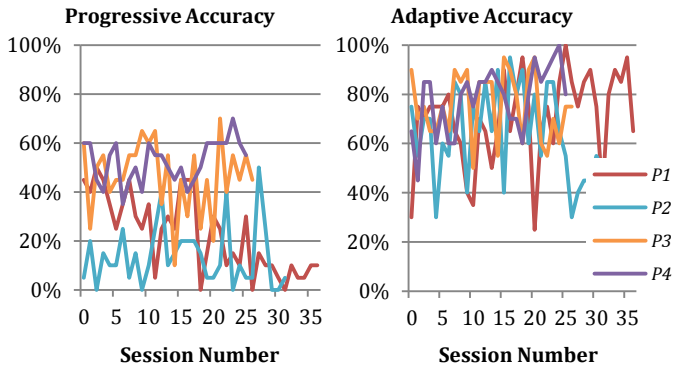

Figure 5. Accuracy over multiple sessions of 20 repetitions

Overall, *Adaptive* succeeds in raising session accuracy levels and holding them high over time. As we shall see in the next section, adapting to abilities, memories, and study schedules is critical in facilitating successful flashcard sessions and for maintaining language learning motivation.

Discussion

In the post-study interview, participants were asked to compare the *Progressive* and *Adaptive* methods in terms of learning experience in the moment and how this relates to their motivation. Participants were also asked to provide feedback on the "mobile microlearning" strategy. We frame our discussion around the themes that arose from these interviews and which provide implications for the future design of such flashcard-based learning systems.

Accuracy motivates by creating a sense of success

Participants widely associated session accuracies with feelings of success. The high accuracies of *Adaptive* sessions were described as creating "the feeling of doing something right" [*P6*], while positive feedback from session percentages amplified the "rewarding" feel of the experience itself, encouraging learners to "invest more time" [*P1*]. The additional statistics of *Adaptive* were also appreciated, with the probability of a correct answer "pushing" learners to try harder [*P5*] and the display of

"how things will stay with you in the long term" motivating learners in the moment [*P1*]. However, lower accuracies with *Progressive*, especially from many successive failures, led leaners to feel like they "don't know anything" [*P3*].

Repeating difficult items creates a feeling of control

The predictable approach to item relearning in the *Adaptive* approach gave learners a "reason" to try and remember because they understood items would "come back again soon" [*P3*]. The expectation of coming success encouraged learners to keep items "in the background" of their mind and consequently feel "in control" [*P5*]. This helped learners "remember more difficult words more easily" [*P4*]. In contrast, *Progressive* was like "seeing items and then discarding them" [*P5*], or even feeling "flooded" [*P6*].

New items should feel deserved, not forced

One common complaint with *Progressive* was that new items would be added before existing items had been learned, leading to feelings of reaching a "tipping point" and ultimately ending up "in a hole" with too many items [*P8*]. In contrast, the *Adaptive* method gave learners the ability to "hold on" to items until they could remember them [*P7*], giving learners "the right to move on" [*P9*]. The expectation of "absolutely going to remember" all *Adaptive* items was so powerful it made them feel "precious" [*P1*].

Encourage learning well, not learning more

Most participants (8/11) retained more items with *Progressive* than *Adaptive*, yet all preferred *Adaptive*. Learners thought that choosing "just a few words" that they "really want to learn" helped them to learn better [*P7*], and that "knowing everything really well is so much more important" than the total items learned [*P5*]. *Adaptive* helped learners to feel "secure and confident" in the items they knew since they had been "ingrained" [*P1*]. In contrast, one learner suggested it was just the "volume of words" that meant they learnt more with the "unpleasant" *Progressive* method [*P6*], while another described how the possibility of learning less than half of the taught words made it feel "random" and "not very goal-oriented" [*P4*].

Facilitate self-paced training of fast, automatic recall

In comparison with *Adaptive*, the *Progressive* method often felt slower, "like it went on forever" [*P1*]. It also felt more difficult and "forced" [P9] compared to the "easier way" [*P4*]. In contrast, *Adaptive* was "a lot more satisfying" because it supported learning at "the right speed" [*P3*] and resulted in "instant" recall [*P5*]. For one learner whose prior experience of language learning was to take "one or two minutes" trying to remember each word from a list, the *Adaptive* method helped him to "answer quickly" even though he "thought nothing", indicating automaticity of recall that is "especially helpful for daily life" [*P10*].

Support flexible learning that can fill any time available

Participants appreciated the "flexible" and "modular" nature of mobile microlearning that could fit into their life, preferring it to the "30 minute lessons" of Pimsleur and Rosetta Stone [*P3*]. Mobile microlearning helped learners to make "good use of time" in situations like walking, shopping, or taking a bus [*P4*], as well as in "coffee breaks" with "5 minutes to spare" [*P2*]. One learner remarked that by adding new cards every day, *Progressive* felt like "what other applications already do", which was why he stopped using those applications [*P9*]. Another learner gave the example of Anki as being "so painful" for the same reason: "if you don't study every day, you will never recover" [*P3*].

Summary

The main finding of study three is that by adapting to the learner's patterns of responses over time, we can significantly raise recall accuracies compared with graduated interval recall based on the Pimsleur intervals. These higher accuracies were found to enhance learner control, confidence, and perception of success.

OVERALL DISCUSSION

The overarching message is that microlearning can work across interaction modalities, when mobile and on the move, in ways that enhance learner motivation. The caveat is that small differences in the design of microlearning systems can preclude flexible switching between modalities (e.g., by requiring extended periods of visual attention) as well as negatively impact the learner experience when the focus is on the quantity, not the quality, of learning.

Designing for "Micro Activities"

The implications of this work can be expressed as a set of six design considerations for when transforming a user activity, such as learning, to be more "micro" in nature:

1. *Fixed* → *Mobile*. Support modality switching because user-context changes more frequently and unexpectedly.
2. *Structured* → *Streamed*. Blend sub-activities (e.g., studying, testing) because sessions can end at any time.
3. *Units* → *Bursts*. Deliver success quickly because early experiences motivate further and future use.
4. *Scripted* → *Adaptive*. Let user data drive content delivery, because users and usage develop over time.
5. *Quantity* → *Quality*. Give users control over the pace of progression, because needs vary over time and users.
6. *Scoring* → *Modeling*. Offer time-based ability estimates, because scores do not convey the effects of inaction.

Limitations and Future Directions

Neither flashcards nor microlearning are tools to be used in isolation. Future work is required to connect this fundamentally behaviorist approach to more situated, constructivist, and collaborative pedagogical methods.

For short-term learning, a promising direction for language learning is the explicit preparation of learners for upcoming conversations with native speakers. A similar approach to preparation could also support the rehearsal of material to be delivered in any kind of public speaking or presentation.

For longer-term learning, adaptive flashcards could be used to help people remember all of the interesting things they read online or in eBooks, using some kind of "clipping" functionality to capture content worthy of retention.

Expanding beyond the individual learner, we would like to derive large-scale insights from aggregate use of our adaptive learning algorithm, both in terms of parameters and content. We have already taken our first step in this direction, by incorporating our adaptive flashcard algorithm into a desktop application for language lookup and adaptive flashcard microlearning [5]. There are currently over 100,000 English as Second Language (ESL) learners using this application each day, and we will continue to evolve our system as we learn about its use in practice.

CONCLUSION

We began this paper by highlighting the opportunity for *mobile microlearning*, before introducing theories, studies, and systems based on *cued recall*, which we used to motivate the design of our MemReflex *adaptive flashcards*. Our evaluation was then set around five research questions.

The first question addressed the effectiveness of *microlearning* as a strategy. In both study one and study three, the items retained in each method suggests that this strategy is a worthwhile addition to a learner's repertoire.

The second question targeted the relative benefit of *adaptive learning* over learning with a fixed progression, and in this case the higher repetition accuracies for the adaptive approach in both study one and study three suggests that adaptation makes a substantial difference.

The third question investigated the appropriateness of audio and text as *interaction modalities* for flashcards. Study one made an initial contribution by demonstrating the effectiveness of audio flashcard learning, while study two built on this by showing equivalent retention for audio-only and text-only flashcards in mobile learning while walking.

The fourth question built on the third by seeking evidence for new *mobile learning opportunities* opened up by the flexible nature of adaptive flashcards. The quantitative results of study two and the usage pattern descriptions from study three provide compelling evidence in support of this.

The fifth and final question about the longer-term *motivational effects* of mobile microlearning was tackled solely by study three, which conclusively established the relationship between recall accuracy and learner satisfaction. By matching learning challenges to demonstrated learner skills, our adaptive flashcards were strongly preferred to a fixed progression modeled on the popular Pimsleur method.

We have only just begun to explore how technology can mediate and connect all forms of learning – from the micro to the macro, the mobile to the ubiquitous, and the personal to the social – but this paper demonstrates how mobile HCI can successfully transform a challenging activity into something more mobile, micro, and motivational in nature.

ACKNOWLEDGEMENTS

We thank all of our study participants and reviewers.

REFERENCES

1. Anki Spaced Repetition System. http://ankisrs.net/
2. Bateson, G. (1972). Steps to an ecology of mind. New York: Ballantine Books
3. Benyon, D. & Murray, D. (1993). Developing adaptive systems to fit individual aptitudes. IUI'93, 115–121
4. Bergstrom-Lehtovirta, J., Oulasvirta, A. & Brewster, S. (2011). The effects of walking speed on target acquisition on a touchscreen interface. MobileHCI '11, 143–146
5. Bing Dictionary Desktop. http://dict.bing.msn.cn/
6. Carpenter, S.K., Pashler, H. & Vul, E. (2006). What types of learning are enhanced by a cued recall test? Psychonomic Bulletin & Review, 13(5), 826–830
7. Carpenter, S.K., Pashler, H., Wixted, J.T. & Vul, E. (1998). The effects of tests on learning and forgetting. Memory & Cognition, 36(2), 438–448
8. Cepeda, N.J., Pashler, H., Vul, E. & Wixted, J.T. (2006). Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. Psychological Bulletin, 132(5)
9. Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT
10. Csíkszentmihályi, M. (1990). Flow: The Psychology of Optimal Experience, New York: Harper and Row
11. Ebbinghaus, H. (1964). Memory: A contribution to experimental psychology. Dover Publications (original 1885)
12. Edge, D., Searle, E., Chiu, K., Zhao, J. & Landay, J. (2011). MicroMandarin: Mobile Language Learning in Context. CHI 2011, 3169–3178
13. Gassler, G., Hug, T. & Glahn, C. (2004). Integrated Micro Learning – An outline of the basic method and first results. Interactive Computer Aided Learning '04
14. Gradint. http://people.pwf.cam.ac.uk/ssb22/gradint/
15. Leitner System. http://en.wikipedia.org/wiki/Leitner_System
16. Nakamura, J, & Csikszentmihayli, M. (2002). The Concept of Flow. The Handbook of Positive Psychology: Oxford University Press, 89–92
17. Pavlik, P.I. Jr. & Anderson, J.R. (2003). An ACT-R Model of the Spacing Effect. ICCM-5, 177–182
18. Pavlik, P.I. Jr., Presson, N., Dozzi, G., Wu, S.-M., MacWhinney, B. & Koedinger, K. (2007). The FaCT (fact and concept) system: A new tool linking cognitive science with educators. 29th Conf. of the Cog. Sci. Soc., Nashville, TN
19. Pimsleur, P. (1967). A memory schedule. The Modern Language Journal, 51(2), 73–75
20. Rohrer, D., Taylor, K., Pashler, H., Wixted, J.T. & Cepeda, N.J. (2005). The Effect of Overlearning on Long-Term Retention. Applied Cognitive Psychology, 19, 361–374
21. Schildbach, B. & Rukzio, E. (2010). Investigating selection and reading performance on a mobile phone while walking. MobileHCI'11, 93–102
22. Sharples, M. (2000). The Design of Personal Mobile Technologies for Lifelong Learning. Computers and Education, 34, 177–193
23. Wilson, G., Brewster, S., Halvey, M., Crossan, A. & Stewart, C. (2011). The effects of walking, feedback and control method on pressure-based interaction. MobileHCI'11 147–156
24. Wozniak, P.A. & Gorzelanczyk, E.J. (1994). Optimization of repetition spacing in the practice of learning. Acta Neurobiologiae Experimentalis, 54