Complementing Text Entry Evaluations with a Composition Task

KEITH VERTANEN, Montana Tech
PER OLA KRISTENSSON, University of St Andrews

A common methodology for evaluating text entry methods is to ask participants to transcribe a predefined set of memorable sentences or phrases. In this article, we explore if we can complement the conventional transcription task with a more externally valid composition task. In a series of large-scale crowdsourced experiments, we found that participants could consistently and rapidly invent high quality and creative compositions with only modest reductions in entry rates. Based on our series of experiments, we provide a best-practice procedure for using composition tasks in text entry evaluations. This includes a judging protocol which can be performed either by the experimenters or by crowdsourced workers on a microtask market. We evaluated our composition task procedure using a text entry method unfamiliar to participants. Our empirical results show that the composition task can serve as a valid complementary text entry evaluation method.

Categories and Subject Descriptors: H.5.2 [**User Interfaces**]: Input Devices and Strategies

General Terms: Human Factors, Experimentation

Additional Key Words and Phrases: Text entry evaluation, composition, transcription, crowdsourcing

ACM Reference Format:
Keith Vertanen and Per Ola Kristensson. 2014. Complementing text entry evaluations with a composition task. ACM Trans. Comput.-Hum. Interact. 21, 2, Article 8 (February 2014), 33 pages.
DOI: http://dx.doi.org/10.1145/2555691

1. INTRODUCTION

Effective text entry methods are crucial for pleasant, fluent, and efficient use of many of the computer systems that surround us. Due to various requirements, such as the small form factor of mobile devices, or a user's limited motor abilities, the pervasive full-sized QWERTY keyboard may not always be a feasible input device. As a result, a wide array of text entry methods have been designed and evaluated using a variety of input modalities, such as single-switches, keypads, touchscreens, eye-trackers, accelerometers, and joysticks. For surveys of text entry methods, see MacKenzie and Soukoreff [2002], Zhai et al. [2005], Kristensson [2009], and Dunlop and Masters [2009].

Similar to other user interface techniques, text entry methods need to be evaluated in order for us to better understand and improve them. In this paper we show how short composition style tasks can be used to help evaluate text entry methods. Such composition tasks can complement the traditional transcription task used in text entry evaluations. With the exception of a speech recognition study by Karat et al. [1999], composition tasks have rarely been used in text entry evaluations.

This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/H027408/1) and the Scottish Informatics and Computer Science Alliance.
Authors' addresses: K. Vertanen, Department of Computer Science, Montana Tech, 1300 West Park Street, Butte, Montana 59701; email: kvertanen@mtech.edu; P. O. Kristensson, School of Computer Science, University of St Andrews, KY16 9SX, St Andrews, UK; email: pok@st-andrews.ac.uk.

DOI: http://dx.doi.org/10.1145/2555691

1.1. The Transcription Task in Text Entry Evaluations

The established procedure when comparing two text entry methods is to conduct a controlled experiment. The setup typically involves recruiting 6–20 participants in a within-subjects multisession experiment in which the text entry method is an independent variable and entry rate and error rate are dependent variables (e.g., Broderick and MacKay [2009], Castellucci and MacKenzie [2008], Clarkson et al. [2005], Kristensson and Denby [2009], Lyons et al. [2006], MacKenzie and Zhang [1999], and Wobbrock et al. [2007], see also the surveys MacKenzie and Soukoreff [2002], Zhai et al. [2005], and Kristensson [2009]). In these text entry evaluation experiments, participants are instructed to transcribe a memorable phrase or sentence as "quickly and as accurately as possible."

Researchers have continually refined the transcription task. In particular, the choice of phrase set has been extensively discussed in the literature. Before 2003, researchers used ad-hoc text sources for their transcription task, such as sentences drawn from a Western novel [Karat et al. 1999] and phrases from a Linux operating system [Isokoski and Raisamo 2000]. MacKenzie and Soukoreff [2003] criticized this practice, pointing out two limitations. First, researchers often did not specify which exact phrases they used. This meant that many text entry studies could not be accurately reproduced. Second, the stimulus phrases should be memorable. This is important to avoid participants switching their visual attention back and forth between the stimulus phrase and the text entry method. To help the situation, MacKenzie and Soukoreff [2003] released a publicly available phrase set consisting of 500 phrases. MacKenzie and Soukoreff [2003] used short idioms, adages, and clichés in hopes of making the stimuli text easy for participants to remember.

Since the release of this phrase set, researchers have also developed alternative phrase sets for specialist applications. Kano et al. [2006] created a phrase set for children. Vertanen and Kristensson [2011a] created a phrase set for Augmentative and Alternative Communication (AAC) by using messages suggested by AAC specialists. To aid the design of specialist phrase sets, Paek and Hsu [2011] proposed a statistical method to generate representative n-gram phrases from text sources. They also released a 4-gram phrase set based on the Enron email dataset for use in text entry evaluations.

Recently, we released a phrase set for mobile text entry evaluations [Vertanen and Kristensson 2011b]. This phrase set has three advantages compared to previous sets. First, it consists of genuine mobile emails extracted from the publicly available Enron email corpus [Klimt and Yang 2004]. Second, a subset of the phrases are validated as indeed being memorable. Third, the individual phrases contain additional metadata, such as how fast and how accurately participants typed the phrases on full-sized keyboards. In Kristensson and Vertanen [2012a], we compared some of the aforementioned phrase sets in two large-scale crowdsourced text entry experiments. We found that the phrase sets released by MacKenzie and Soukoreff [2003], Paek and Hsu [2011], and Vertanen and Kristensson [2011b] resulted in very similar entry and error rates for both the familiar full-sized QWERTY keyboard and for the unfamiliar ATOMIK optimized on-screen keyboard. This result suggests that researchers using transcription tasks can use the most suitable phrase set among these three for their application. For example, mobile text entry method evaluations should use the mobile email phrase set [Vertanen and Kristensson 2011b] because it consists of memorable phrases taken from actual mobile email messages.

Another research question that has been investigated for transcription tasks is how to present the stimulus phrase. The phrase can either remain visible or be hidden during the writing phase. By hiding the stimulus phrase, participants are forced to

memorize the phrase before attempting to write it. As pointed out by MacKenzie and Soukoreff [2003], using memorable phrases should reduce participants' tendency to shift attention between the stimulus phrase and the text entry method. By forcing users to memorize phrases beforehand, this effect can be mitigated. Assuming attention shifts between the stimulus phrase and the text entry method do occur, a plausible hypothesis is that hiding the stimulus phrase during the writing phase will increase the entry rate. To validate this hypothesis, Soukoreff and MacKenzie [2003] found that hiding the stimulus phrase during writing did in fact yield a slight but statistically significant increase in entry rate. We did a similar study using two large-scale crowdsourced text entry experiments [Kristensson and Vertanen 2012a]. We found a similar slight but statistically significant increase in entry rate for both the familiar full-sized QWERTY keyboard and for an unfamiliar ATOMIK optimized on-screen keyboard. We found that this increase in entry rate came at the cost of statistically significant longer task times. In other words, participants wrote faster when they memorized phrases beforehand. However, participants spent more time on each experimental task, presumably due to time spent memorizing each phrase before starting to write.

Finally, Isokoski and Linden [2004] investigated how participants' language proficiency affects transcription task performance. Isokoski and Linden [2004] let 16 native Finnish speakers write phrases in Finnish and English. They found that participants had a 16% lower entry rate when they wrote English phrases compared to Finnish phrases. The difference was statistically significant. This result suggests that transcription tasks are language-sensitive and this raises two issues. First, as pointed out by Isokoski and Linden [2004], study heterogeneity *is* affected by a participant's native language. Second, minimizing study heterogeneity when conducting experiments with participants who are nonnative speakers will be difficult. We either have to accept performance differences across studies carried out among nonnative speakers by using a standard phrase set, or we have to develop language-specific phrase sets. The latter solution is unlikely to generate directly comparable results.

1.2. Complementing the Transcription Task

The previous review makes it clear that the transcription task is firmly entrenched as the *de facto* research methodology for text entry experiments. The advantage of the transcription task is that it strengthens the internal validity of the experiment by reducing variance among participants from at least three sources. First, it ensures all participants write the same text. This removes the variance that might occur due to participants writing widely varying texts. As an extreme example, the amount of attention to spelling, grammar, and use of vocabulary varies greatly between writing a legal document versus writing a short message to a friend. Second, a transcription task does not require participants to think of something to write, a process which demands additional cognitive processing time. If this processing time occurs in the middle of the entry of a phrase, it will increase the variance in measured entry rates. Third, as long as stimuli are short and memorable, participants can internalize the stimuli before they start writing. This means that once participants start copying the stimuli, they do not have to devote visual attention to the text they are transcribing.

The downside of using a transcription task is primarily its low external validity. In actual practice, users compose original text—they do not transcribe text that was flashed to them a second earlier. Unless we know what individual users actually want to write, we risk testing text entry methods on inappropriate stimuli. This is particularly problematic if the text entry method uses a dictionary or a statistical language model. If the language model is well adapted to the stimuli used for the text entry evaluation but ill suited for users' actual writing then the transcription task could lead to misleadingly positive results. The converse is also true: A text entry method may draw on a language

model that is wel suited for what end users are actually writing (e.g., mobile emails) but may be ill suited for stimuli used in a transcription task, such as the phrase set by MacKenzie and Soukoreff [2003], which mainly consists of short memorable idioms, adages, and clichés. In this case, the transcription task might lead to misleadingly negative results.

The external validity of the transcription task is also a problem when evaluating certain types of text entry interfaces, such as Dasher [Ward and MacKay 2002]. Dasher is an interface that enables users to write by navigating a dynamically changing scene that constantly proposes the most likely letters based on previously written text. Dasher demands constant visual attention from the user and is designed to support users writing a series of related sentences and paragraphs (as in actual writing). A transcription task thus interferes with writing using Dasher because the user may need to periodically shift visual attention between reading the stimulus phrase and driving the Dasher interface.

The transcription task can also interfere with the ability to test a text entry interface in realistic settings. In our study of the Parakeet mobile speech recognition interface [Vertanen and Kristensson 2009], we had participants speaking and correcting phrases while walking around outdoors. Due to the limited screen space on the small mobile device, a participant who forgot the stimulus phrase had to take an explicit interface action in order to refer back to the phrase. We contend this required additional cognitive resources from participants who were already heavily loaded. We believe a Parakeet study based on a composition task may yield entry rates more representative of what one could really expect from text entry using the interface "in the wild."

In some regards, a transcription task may also have a lower internal validity than a composition task. First, all users do not have the same writing styles and preferences. These factors might vary according to users' level of education or cultural factors. Despite this, in a transcription task, all participants are forced to write the text they are provided. If participants do not know how to spell certain words, it may affect both their entry and error rate. Second, if participants are used to writing in a different genre or style, then their cognitive effort to adopt the task may be higher. These two factors are likely to vary among individuals and could confound the results of a study.

1.3. Using Composition Tasks in Text Entry Experiments

While the transcription task is definitely useful and often good practice, as we just discussed, it is not without its flaws. We believe other aspects of text entry can be illuminated by complementing the transcription task with a composition task. However, a researcher contemplating the use of a composition task to evaluate a new text entry method may have a number of concerns:

(1) The composition task may be too slow and too variable to provide reliable estimates about a text entry method's performance potential.
(2) The cognitive overhead associated with inventing compositions may increase variance among participants and, therefore, decrease the internal validity of the experiment by an unacceptable degree.
(3) Participants may find it hard to think about things to write and may spend a large proportion of the experiment planning what to write instead of actually using the text entry method.
(4) There may be no good way to determine the error rate of compositions because there is no reference text.
(5) Participants may lack creativity, resulting in compositions that are short and simplistic.

In this article, we hope to alleviate these concerns and show how the composition task can be a useful additional method in the text entry evaluation toolbox. Our goal is not to replace the traditional transcription task but to provide researchers with an additional option that can help better measure anticipated real-world text entry performance.

1.4. Article Organization

To show that a composition task can be workable for text entry evaluation, we describe a series of experiments. In the first experiment, we compare the transcription task with four different composition tasks. The purpose of this experiment is to explore if composition tasks are at all suitable for text entry evaluation, and if so, which composition task is the most effective. In the second experiment, we explore the most promising composition task and investigate how to fine tune the procedure to ensure participants write high-quality text. Both of these experiments use the full-sized QWERTY keyboard as the text entry method. In the third experiment, we evaluate our proposed composition task and compare it against a transcription task for a text entry method unfamiliar to participants: an optimized on-screen keyboard. In the last experiment, we investigate using crowdsourcing to efficiently judge compositions. Finally, we discuss limitations, implications, open issues, provide a best-practice procedure for using composition tasks in text entry evaluations, and thereafter conclude.

2. EXPERIMENT 1: EXPLORING COMPOSITION TASKS

We conducted a within-subjects experiment in which users were given five different kinds of writing tasks. These tasks included a conventional transcription task, two guided composition tasks, and two freeform composition tasks. The goals of this experiment were (1) to evaluate different composition tasks to see how participants reacted, (2) to explore the efficiency of collecting text entry data using composition, and (3) to examine the quality of the compositions.

2.1. Method

We wanted to collect data from a large number of users across a broad cross-section of the world's population. To this end, we designed a web-based experiment that could be performed on the popular crowdsourcing website Amazon Mechanical Turk. In our first experiment, participants entered text into a standard web textbox using their full-sized keyboard. The web page was instrumented to record the timestamp of every keystroke in the textbox. Each participant did five different types of writing tasks and wrote 10 entries in each condition. One condition was transcription, two were guided composition, and two were freeform composition:

- COPY — Participants typed a provided sentence. The sentences were taken from a corpus of email messages written by Enron employees on their BlackBerry mobile devices [Vertanen and Kristensson 2011b].
- REPLY — Participants were given a mobile message and asked to reply to it. The messages were taken from messages written by other workers who were asked to invent a message they might send from a mobile device.
- SITUATION — Participants were given a specific mobile situation and asked to invent a message to send from a mobile device. The situations were invented by the authors.
- COMPOSE — Participants were asked to invent a freeform mobile message.
- AID — Participants were asked to imagine they were unable to speak or type (due to a medical condition or accident). They were told to invent a communication as if they were using a special communication aid that talked for them. This

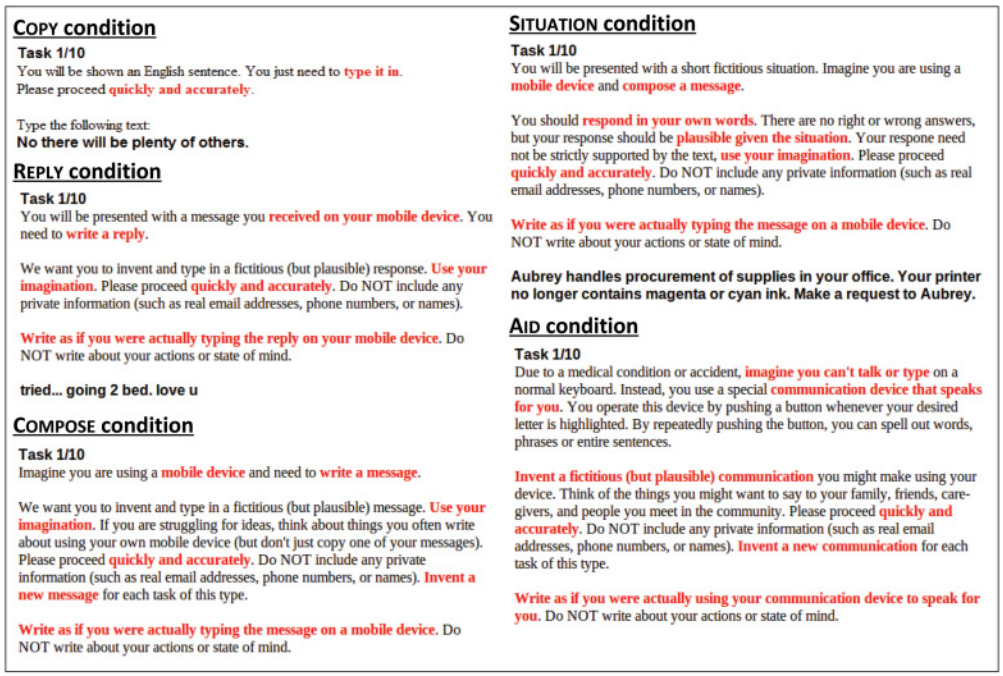


Fig. 1. The instructions given to participants in the five conditions in Experiment 1. Each condition consisted of 10 tasks and each task asked participants to perform a composition or a transcription. The screenshots show the complete instructions provided to the participants for each condition.

Table I. Example Stimuli from the Two Guided Composition Conditions in Experiment 1

Condition	Stimulus text shown to participant
REPLY	The bus is late again, and I'm getting very irritated!
	Hey i got 2/3 but not d minus sign.
	Wanna go to the park?
	I hope that you are okay.
SITUATION	Your housemate has been sick for the last week. You are currently shopping downtown. See if he requires anything.
	Your friend Carol is at the local sandwich shop. Request a ham and cheese baguette.
	Your co-worker Carl will be 29 years old today. Send him a greeting.
	Your friend is picking you up at the airport but you are still waiting for your bags. Inform your friend of the situation.

condition was designed to investigate how additional constraints affect a freeform composition task. As we will later show in this article, this condition resulted in shorter compositions.

Figure 1 shows the instructions given to our crowdsourced participants in each condition. In each condition, participants completed 10 text entry tasks. In two of the conditions, REPLY and SITUATION, participants were shown an existing message or situation. Some examples of the stimuli shown in these two conditions are given in Table I. There were a total of 200 possible messages in the REPLY condition and 73 possible situations in the SITUATION condition. In each condition, participants received 10 messages or situations selected at random. All our stimulus texts are provided in the appendix.

The order each participant encountered the five conditions was randomized. We collected a variety of information before the first condition. We asked participants for their sex, country, age, English proficiency, typing ability, and computer type (e.g., laptop). We also asked for the same information after the last condition, but this time we partially permuted the form. We eliminated participants that had two or more discrepancies in the provided data. The Human Intelligence Task (HIT) was priced at $1.00, and the only limitation was that workers needed to have a 95% accepted HIT rate.[1]

2.2. Participants and Data Filtering

We had 200 participants complete the HIT. It took 39 hours for all the HITs to be completed. Participants took on average 27 minutes to do the experiment. Participants were required to complete the experiment within 2 hours of accepting the HIT. We eliminated a total of 19 participants who had participated in several pilot experiments, who entered garbage text, or who provided inconsistent information about themselves. This left us with 181 participants who made a total of 9,050 entries. We eliminated 20 entries in which participants had entered a single character.

The top self-reported countries were the United States (64%) and India (18%). Our sample was fairly gender balanced, with 53% female and 47% male. The majority of participants were aged 20–34 (70%). English proficiency was reported most often as native (71%), followed by advanced (20%). Most participants reported using a laptop computer (55%) or a desktop computer (45%). A full report of the demographic data appears in the appendix.

In some cases, we found participants paused for a long time between keypresses during a particular entry task or between different tasks in the experiment. For example, one participant waited over 30 minutes between two keypresses. This is presumably due to the participant being interrupted or taking a break midexperiment. We removed any individual entry task in which the participant paused for more than 1 minute between keypresses, or paused for more than 1 minute before the first keypress. This filtering eliminated only a small number of entries (108 out of 9,030). This filtering prevented infrequent long pauses from skewing our calculated statistics.

2.3. Results

Overall, we were pleasantly surprised by the quality and creativity the participants exhibited in their compositions (Table II). For each entry, we measured the entry rate in words per minute (wpm). The time was measured from a participant's first keypress to his or her last keypress. We used the standard convention of defining a word as five consecutive characters (including spaces). As might be expected, COPY had the fastest entry rate of 59 wpm (sd 21). The composition condition entry rates were: REPLY 50 wpm (sd 21), COMPOSE 48 wpm (sd 18), AID 44 wpm (sd 19), and SITUATION 38 wpm (sd 14). An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ revealed a statistically significant difference between the conditions ($p < 0.0001$, $\eta_p^2 = 0.307$, $F_{4,720} = 79.767$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the REPLY and COMPOSE conditions.

Entry rates were variable in every condition (Figure 2). This may reflect the diverse English and typing abilities of participants. One possible explanation for the lower

[1]A Human Intelligence Task (HIT) is a microtask that is bid out on a microtask market, in this case Amazon Mechanical Turk. A microtask market is an online market place that connects requesters and workers. Requesters bid out work and workers carry out work. Requesters pay a fixed amount of money to workers in return for workers completing microtasks. The accepted HIT rate of a worker is the percentage of HITs a worker has carried out that has been accepted by requesters.

Table II. Examples of Text Generated by Participants in the Four Composition Conditions in Experiment 1

Condition	Text generated by participant
Reply	They're in the pantry next to the Cheerios.
	Hai is cute! ^.^ and definitely, what time?
	It's really nice out, you should come here lol!
	We saw Inception - it was AWESOME
Situation	Let's watch Casablanca on TV
	can i get a ham and cheese baguette?
	Hey buddy good luck at the game tomorrow.
	I can go tomorrow but only if its sunny.
Compose	lol, that's funny. :D :D :D
	Where are we for bridge this week?
	When do you get back from fencing?
	I have to go to Laura's baby shower tomorrow.
Aid	Where do you want to go out for dinner?
	Can you get me a tea
	I am thirsty.
	Want to play a game?

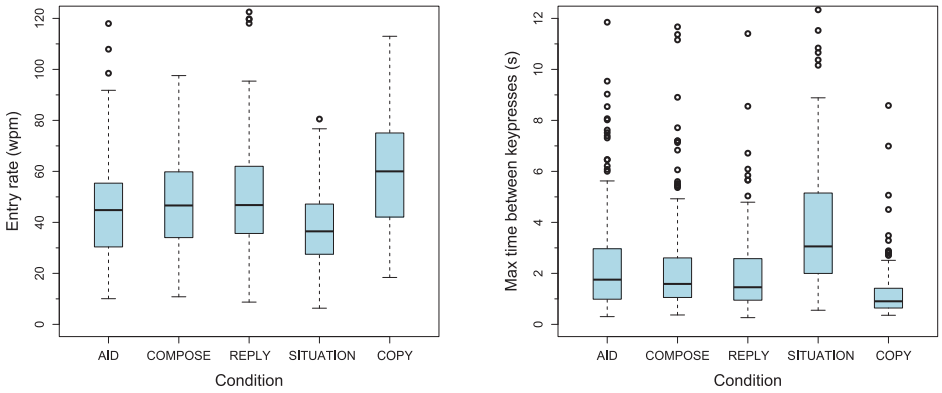


Fig. 2. Box and whisker plots of each participant's mean entry rate (left) and each participant's mean maximum time between keypresses (right) in Experiment 1.

composition entry rates is that participants were pausing to think in the middle of entries. To investigate this, we calculated the maximum time gap between keypresses in each entry. Indeed, the maximum gap was longer for composition than for transcription: Copy 1.2 s (sd 1.0), Reply 2.1 s (sd 2.0), Compose 2.3 s (sd 2.2), Aid 2.5 s (sd 2.5), and Situation 4.1 s (sd 3.4). The maximum time between keys was variable (Figure 2, right). An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ on the log-transformed time intervals revealed a statistically significant difference between the conditions ($p < 0.0001$, $\eta_p^2 = 0.397$, $F_{4,720} = 118.501$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the Aid, Reply, and Compose conditions.

The number of characters written was highly dependent on the condition: Situation 66 characters (sd 28), Copy 48 characters (sd 9), Compose 38 characters (sd 25), Reply 26 characters (sd 14), and Aid 26 characters (sd 24). The number of characters written in some of the conditions also exhibited high variability (Figure 3, left). However, as we have previously shown, the variability in entry rate was similar between composition

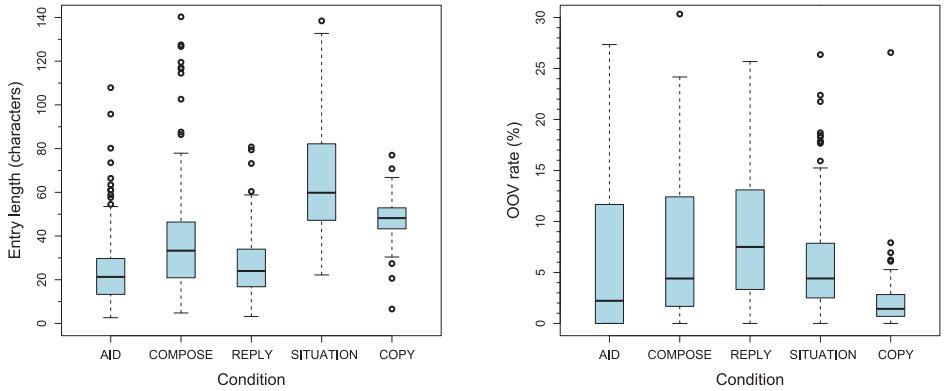


Fig. 3. Box and whisker plots of each participant's mean characters per entry (left) and each participant's mean OOV rate (right) in Experiment 1.

and transcription tasks. In other words, the measure of entry rate normalizes the variable amount of text users write in a composition task.

The hypothetical text entry constraints implied by the AID condition did in fact result in participants writing shorter freeform entries compared to the COMPOSE condition. This shows that different task instructions can influence writing behavior. An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ revealed a statistically significant difference in the number of characters written between the conditions ($p < 0.0001$, $\eta_p^2 = 0.466$, $F_{4,720} = 156.791$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the AID and REPLY conditions.

We noticed a prevalence of texting abbreviations and emoticons in the compositions. Our participants were asked to imagine they were using a mobile device and they did so with gusto. We calculated an Out-Of-Vocabulary (OOV) rate with respect to a 64K lexicon of words. The 64K lexicon used the most common words in an email corpus that also appeared in a large human-edited dictionary.[2] The OOV rate was high in all conditions aside from COPY (Figure 3, right). The means and standard deviations of the conditions were: COPY 2.0% (sd 2.4%), SITUATION 6.4% (sd 6.4%), AID 9.5% (sd 16.5%), COMPOSE 9.7% (sd 12.9%), and REPLY 9.9% (sd 9.9%). An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ revealed a statistically significant difference between the conditions ($p < 0.0001$, $\eta_p^2 = 0.119$, $F_{4,720} = 24.271$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the AID, REPLY, and COMPOSE conditions.

A concern about composition style entry tasks is that it may take users a long time to invent a composition before text entry can even begin. This time is not reflected in our previously calculated entry rates. It is plausible different conditions required varying amounts of cognitive overhead before writing could commence. We, therefore, measured the time participants spent between each entry in a condition. A concern with this measure is that it may include instances of participants taking a break mid-condition. However, there is no reason to suspect such breaks would occur more often in any particular condition. The mean times between entries in each condition were: COPY 3.0 s (sd 1.8), REPLY 6.6 s (sd 3.0), COMPOSE 7.1 s (sd 4.3), AID 8.1 s (sd 5.1), and SITUATION 10.6 s (sd 5.0). Figure 4 (left) shows that participants did appear to be spending longer before starting to enter text in the composition conditions compared

[2]http://keithv.com/software/composition

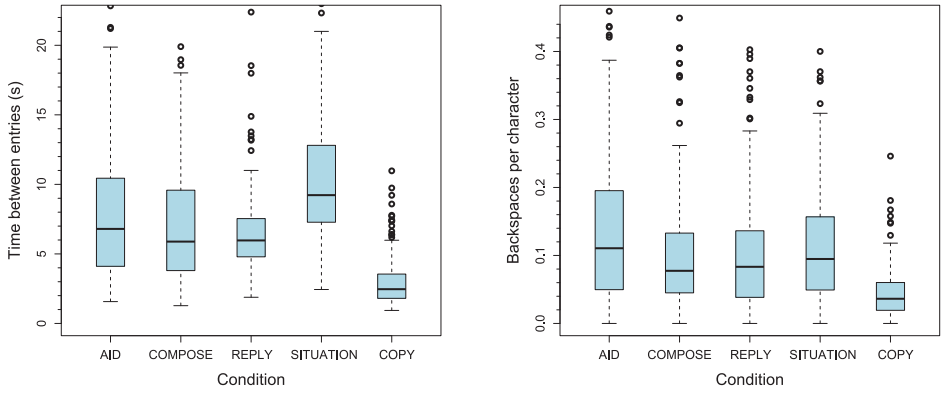


Fig. 4. Box and whisker plots of each participant's mean time spent between entries (left) and each participant's mean number of backspaces per character in the final entry (right).

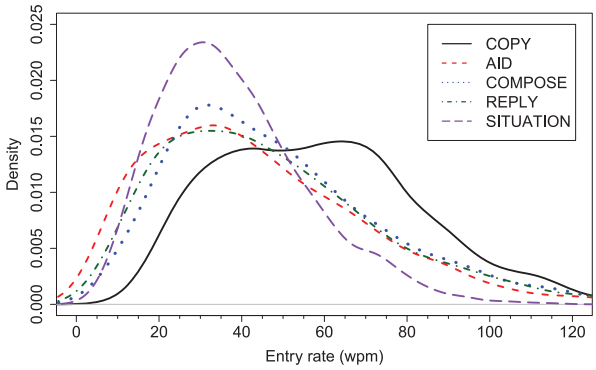


Fig. 5. Distribution of entry rates in Experiment 1.

to the COPY condition. An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ on the log-transformed time intervals revealed a statistically significant difference between the conditions ($p < 0.0001$, $\eta_p^2 = 0.617$, $F_{4,720} = 290.583$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the AID and REPLY conditions and between the REPLY and COMPOSE conditions.

It is likely composition entries would be more tenuous and subject to revision than transcribed entries. We calculated the total number of times users hit backspace divided by the number of characters in the final output. The mean number of backspaces per output character were: COPY 0.04 (sd 0.04), REPLY 0.11 (sd 0.11), COMPOSE 0.11 (sd 0.11), SITUATION 0.11 (sd 0.08), and AID 0.15 (sd 0.14). Figure 4 (right) shows that composition tasks experienced a much higher degree of editing compared to transcription tasks. An omnibus repeated measures analysis of variance test at significance level $\alpha = 0.05$ revealed a statistically significant difference between the conditions ($p < 0.0001$, $\eta_p^2 = 0.174$, $F_{4,720} = 37.916$). Pair-wise Bonferroni-corrected post hoc tests showed that all pair-wise differences were statistically significant except between the REPLY, SITUATION, and COMPOSE conditions.

The entry rate distribution of every condition is shown in Figure 5. The density plots in Figure 5 are kernel density estimates using a Gaussian kernel. As can be seen, the entry rates for the transcription condition (Figure 5, solid line) were often much faster

than the entry rates for the composition conditions (Figure 5, dotted lines). All the four composition conditions had very similar entry rate distributions.

In summary, the first experiment showed that participants' performance varied depending on the type of composition task they were asked to complete. Taking all the data from the experiment into account, the COMPOSE task appeared to be the most promising one. COMPOSE was among the fastest in term of entry rate and had the second longest compositions. While SITUATION produced longer compositions, it also had much longer pauses before and during entries. Another advantage of COMPOSE is that it requires no stimuli, as participants simply invent freeform compositions. However, a problem with the COMPOSE task was that participants frequently used SMS-style abbreviations, emoticons, and so on. This is not too surprising given we asked participants to pretend they were sending a message using a mobile device. Unfortunately, such text can be hard to judge whether it is correct or not. In the next experiment, we therefore explore how to fine tune the COMPOSE task to obtain text more suitable for error rate measurement.

3. EXPERIMENT 2: FINE-TUNING THE COMPOSITION TASK

Given our findings from the first experiment, we conducted a second experiment to further explore using a freeform short composition task similar to the COMPOSE condition. In this experiment, we tested (a) whether we could influence the style of text participants composed and (b) whether we could alter when participants spent time planning their compositions.

3.1. Method

We hypothesized it was possible to increase entry rates by reducing the amount of time participants spent formulating text midentry. We investigated this via a between-subjects experiment in which half the participants received the instruction: "Think carefully about what you intend to write before you start. You may want to say your intended message to yourself before you start typing." The other half of participants received no such instruction. As in Experiment 1, we used Amazon Mechanical Turk for this experiment.

In order to discourage texting abbreviations and emoticons, we added the following instruction to both conditions: "Please write complete sentences with good grammar and spelling. Do NOT use texting abbreviations or slang."

Each participant wrote a total of ten compositions. The HIT was priced at $0.40, and it was open to all workers with 95% or more accepted HITs. As in Experiment 1, we asked participants for their sex, country, age, English ability, typing ability, and computer type, both before and after the experiment.

3.2. Participants and Data Filtering

Overall, we had 105 participants complete Experiment 2. We eliminated a total of 11 participants who had participated in pilot experiments, participants who entered garbage text, or participants who provided inconsistent information. This left us with 94 participants, 47 who received the "Think carefully" instruction and 47 who did not. The experiment took participants about 7 minutes to complete. The information collected about each participant was similar to Experiment 1 (see the appendix).

As in Experiment 1, we occasionally observed participants taking breaks during the experiment. We eliminated any entry task where the participant paused for more than 1 minute during text entry or where the participant paused for more than 1 minute between compositions. This filtering eliminated only a small number of the 940 entries, 12 from the "Think carefully" condition and 8 from the no instruction condition.

Table III. Some Example Compositions from Participants in Experiment 2

Do you want to get coffee tomorrow?
What do you think the Calgary Flames need to change so they can start winning games consistently?
The weather is going to be really icy tonight, you might want to salt down your driveway.
We are running late.
Stan and I are going out for Thai, then then bowling. If anyone wants in, ping me.
How are you feeling?
Anyone know of a good restaurant on West Honeysuckle Lane?
Could you pick up some bread on the way home? I just realized we're out.
Do you remember the name of that book?
Call me when you get this please!

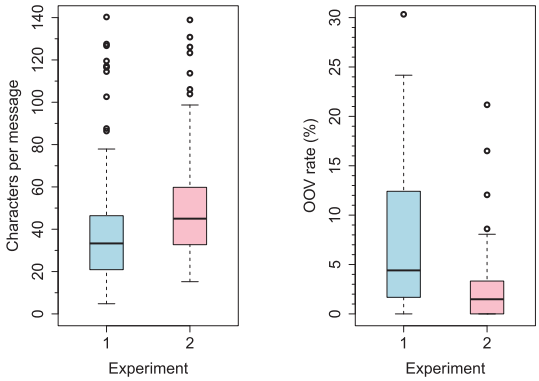


Fig. 6. Box and whisker plots of each participant's mean characters per entry (left) and each participant's mean out-of-vocabulary rate (right) between Experiment 1 (COMPOSE) and Experiment 2.

3.3. Results

An inspection of the resulting compositions showed that participants did appear to be much better in terms of using complete sentences, full words, and participants mostly avoided texting specific language (see Table III for some examples). The length of entries had similar variability in both experiments (Figure 6 left). Participants' mean entry length in characters increased to 52 characters (sd 27) in Experiment 2 from 38 characters (sd 25) in the COMPOSE condition of Experiment 1. An analysis of variance at significance level $\alpha = 0.05$ revealed that this difference was statistically significant ($p < 0.0001$, $\eta_p^2 = 0.062$, $F_{1,273} = 18.050$).

The OOV rate was also reduced to 2.3% (sd 3.4) in Experiment 2 compared to 9.7% (sd 12.9) in Experiment 1. The OOV rate was much less variable in Experiment 2 (Figure 6, right). An analysis of variance at significance level $\alpha = 0.05$ revealed that this difference was statistically significant ($p < 0.0001$, $\eta_p^2 = 0.097$, $F_{1,273} = 29.372$).

Overall, we found that the "Think carefully" instruction had little effect on participants' behavior. The mean participant entry rate was 41 wpm (sd 15) when given the instruction and 45 wpm (sd 15) when they did not receive the instruction (Figure 7, left). An analysis of variance at significance level $\alpha = 0.05$ revealed that this difference was not statistically significant ($p = 0.232$, $\eta_p^2 = 0.015$, $F_{1,92} = 1.445$).

The maximum delay we saw between keypresses was 2.8 s (sd 2.0) when given the instruction and 2.9 s (sd 2.0) without the instruction (Figure 7, right). An analysis of variance at significance level $\alpha = 0.05$ on the log-transformed time intervals revealed that this difference was not statistically significant ($p = 0.809$, $\eta_p^2 = 0.001$, $F_{1,92} = 0.059$).

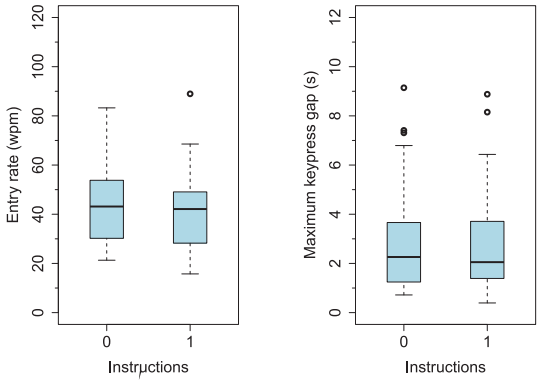


Fig. 7. Box and whisker plots of each participant's mean entry rate (left) and each participant's mean maximum pause time between keypresses (right) depending on if participants received the "Think carefully" instruction (1) or not (0).

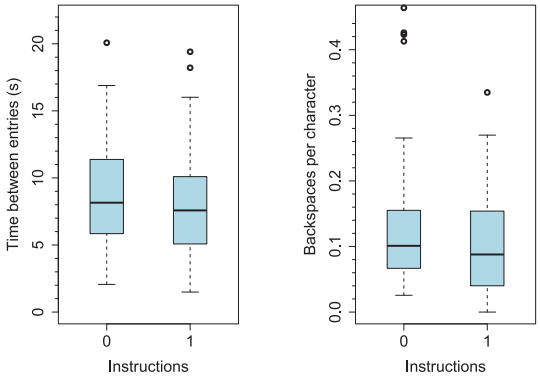


Fig. 8. Box and whisker plots of each participant's mean time spent between entries (left) and each participant's mean backspaces per output character (right) depending on if participants received the "Think carefully" instruction (1) or not (0).

The time participants spent between entries was 8.7 s (sd 4.2) when given the instruction and 8.2 s (sd 4.9) without the instruction (Figure 8, left). An analysis of variance at significance level $\alpha = 0.05$ on the log-transformed time intervals revealed that this difference was not statistically significant ($p = 0.468$, $\eta_p^2 = 0.006$, $F_{1,92} = 0.531$).

Editing behavior was similar in the two conditions. The number of backspaces per output character was 0.13 (sd 0.15) with the instruction and 0.13 (sd 0.11) without (Figure 8, right). An analysis of variance at significance level $\alpha = 0.05$ revealed that this was not statistically significant ($p = 0.899$, $\eta_p^2 = 0.0$, $F_{1,92} = 0.016$).

In summary, the second experiment showed that by changing the instructions given to participants, we could usually obtain compositions in good English that avoided SMS-style language. It also showed that encouraging participants to think about what to compose before they started to write had no effect on entry rate. This may be because thinking is inherently interleaved with writing or it may be that the specific instruction caused participants to be more careful and thus type more slowly.

We also learned that by providing carefully worded instructions, it is feasible to have participants write short compositions of good length and quality. However, so far we have only tested the composition task on the familiar full-sized keyboard. To verify that

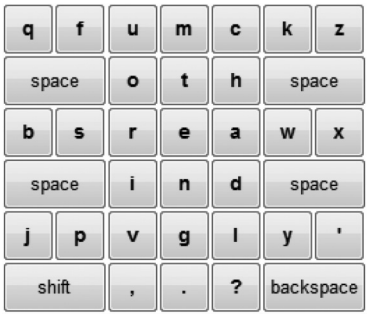


Fig. 9. The modified version of the OPTI keyboard used in Experiment 3.

the composition task also works when evaluating an unfamiliar text entry method we conducted a third experiment.

4. EXPERIMENT 3: EVALUATING THE COMPOSITION TASK

We conducted an evaluation to demonstrate how short freeform composition can be used in a practical text entry evaluation using a novel text entry interface. Additionally, we wanted to obtain data about the subjective taskload difference between transcription and composition tasks, and to solicit qualitative feedback about each task type.

4.1. Method

For our entry interface, we used the optimized keyboard layout OPTI presented by MacKenzie and Zhang [1999]. This keyboard was designed to minimize the amount of movement required when a single pointer (e.g., a finger or pen) is used to select frequently occurring letter combinations. We designed a web-based version of the OPTI keyboard. Using OPTI with a mouse will obviously be slower than with a stylus (which OPTI was designed for). However, in this experiment, we are not concerned with absolute performance. Instead, our goal is to compare the relative performance difference between the transcription and the composition task using an interface unfamiliar to participants.

We modified OPTI to include keys for apostrophe, period, comma, and question mark. We also added a backspace key and a shift key (Figure 9). When users pressed the shift key, all letters on the keyboard changed to uppercase and the next letter would be output in uppercase.

To measure the participants' subjective workload, we used the NASA task load index [Hart and Stavenland 1988]. NASA-TLX asks participants to rate six aspects of task difficulty: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants then make 15 pairwise decisions about which aspect in each pair was perceived as more important. We developed an online version of the NASA-TLX index for use in web-based experiments, which we have made available to other researchers.[3]

Experiment 3 used a within-subjects experimental design with two conditions. In the COPY condition, participants transcribed sentences using the OPTI keyboard. In the COMPOSITION condition, participants invented their own sentences. The experiment consisted of the following parts:

[3]http://keithv.com/software/nasatlx

Participant information	As in Experiment 1 and 2, we collected information about the participant's sex, country, age, English proficiency, typing ability, and computer type.
COPY session	Participants transcribed prompted sentences for 10 minutes. Sentences were drawn from messages written by Enron employees on their BlackBerry mobile devices [Vertanen and Kristensson 2011b]. We only used sentences with characters available on our OPTI keyboard. We chose 200 sentences with between 4 and 10 words. Participants encountered the sentences in random order.
COPY taskload	Participants completed a NASA-TLX survey about the previous transcription session.
COPY feedback	We asked for optional "Comments about the on-screen keyboard in the last session" and "Comments about the text entry task (copying the provided sentences)."
COMPOSITION session	Participants invented and typed compositions for 10 minutes.
COMPOSITION taskload	Participants completed a NASA-TLX about the previous composition session.
COMPOSITION feedback	We asked for optional "Comments about the on-screen keyboard in the last session" and "Comments about the text entry task (inventing your own messages)".
Participant information 2	We presented a permuted version of the initial questionnaire.

The order of the COPY and COMPOSITION tasks was balanced. For this experiment, we paid $2.00 per HIT. Workers were required to have 95% or more accepted HITs. Before starting the experiment, participants were told we required two 10-minute periods of uninterrupted text entry.

In our previous two experiments, we occasionally saw non-US participants who appeared to be copying text from news headlines and other sources rather than creating novel compositions. The quality of English in the compositions was also sometimes poor. We felt these poor compositions might make it difficult to obtain an accurate error measure. We, therefore, limited this experiment to workers from the United States.

4.2. Participants and Data Filtering

Overall, we had a total of 51 participants complete Experiment 3. We eliminated one participant who did not follow the instructions in the COMPOSITION condition. None of the participants had more than one discrepancy in their two sets of information. On average, the experiment took 31 minutes to complete. Participants were eager to do this experiment and all instances were taken within 70 minutes.

We found 72% of participants self-reported as female. All but one participant self-reported as a native speaker of English. See the appendix for more details about the information collected from participants.

As before, we filtered any tasks in which a participant paused for more than a minute before, or during, text entry. This filtering eliminated only a small number of the 1,192 entries, 10 from the COPY condition and 20 from the COMPOSITION condition.

4.3. Results

The mean participant entry rate was 8.5 wpm (sd 2.1) in COMPOSITION and 8.8 wpm (sd 2.0) in COPY (Figure 10 left). A repeated measures analysis of variance at significance

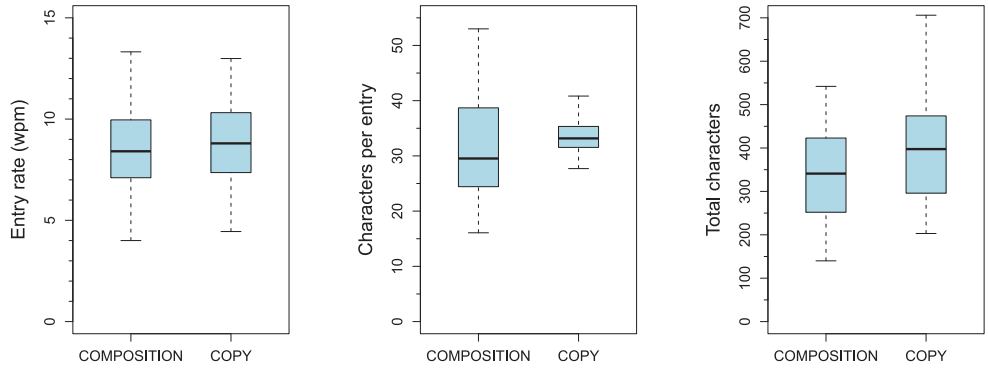


Fig. 10. Box and whisker plots for each participant's mean entry rate (left), each participant's mean characters per entry (middle), and each participant's mean total characters written in each condition (right).

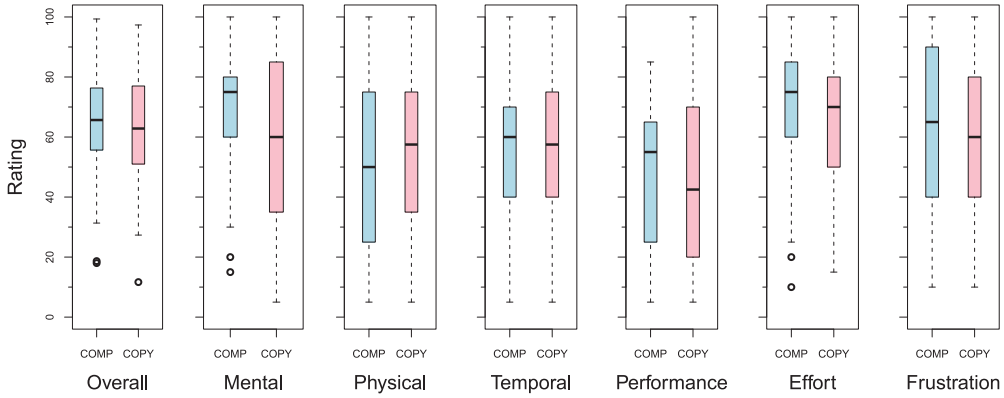


Fig. 11. Box and whisker plots of overall and individual NASA-TLX ratings for COMPOSITION (blue) and COPY (pink).

level $\alpha = 0.05$ revealed that this difference was not statistically significant ($p = 0.117$, $\eta_p^2 = 0.050$, $F_{1,49} = 2.552$).

In the COMPOSITION condition, participants wrote an average of 39 characters per entry (sd 33). In the COPY condition, participants wrote an average of 33 characters per entry (sd 3). A repeated measures analysis of variance at significance level $\alpha = 0.05$ revealed that this difference was not statistically significant ($p = 0.282$, $\eta_p^2 = 0.024$, $F_{1,49} = 1.186$). Figure 10 (middle) show entry length was quite variable in the COMPOSITION condition.

Since each condition had a fixed 10-minute writing period, we also computed the sum of characters entered during each condition. Participants wrote an average of 339 characters (sd 104) in COMPOSITION and 394 characters (sd 111) in COPY (Figure 10, right). Thus, the composition task did cause a modest reduction in total text entry activity by our participants. A repeated measures analysis of variance at significance level $\alpha = 0.05$ revealed that this difference was statistically significant ($p < 0.0001$, $\eta_p^2 = 0.327$, $F_{1,49} = 23.823$).

We found the overall NASA Task Load Index was similar for both conditions, with a median participant index of 66 in COMPOSITION and 63 in COPY. Friedman's test revealed that the difference was not statistically significant ($\chi^2 = 0.510$, $df = 1$, $p = 0.475$). As shown in Figure 11, there were no clear rating differences between the conditions.

Table IV. User Comments About the COMPOSITION Task

it wasn't that hard, but i'm sure my sentences were boring
Inventing my own sentences was actually fun. I was just not sure how long my sentences should be.
That was a large part of the mental effort – trying to diligently create new messages one after the other.
the hardest part was to have imaginary conversations with no one.
Would have been more comfortable learning with sentences given to me than having to think up the messages.
Conversation with myself is always fun.
I just wrote what came into my mind just like I always do.
I just texted random messages about the weather, my dog, etc. It was not that difficult to keep coming up with short sentences/thoughts.
It was a little odd to be typing full words for a message I was pretending to be a task.
It was easier to make a conversation than I thought it would be.
It wasn't as easy as I thought it would be. It took a lot more thinking than I thought it would.
This was an easier task than the first...I didn't have to compare my text to a preexisting template but could write whatever I chose.

Table V. User Comments About the COPY Task

I didn't like copying it- SO TEDIOUS to me!!
It took away some of the thought and I could concentrate on learning the keys.
it was easier having a sentence to type rather than thinking of one
It was nice having something to copy from.
Less mentally stressful, but a little more boring.
This one was a lot easier than coming up with sentences.

Many participants provided feedback in the open comments section after each condition. Many comments were about details of the OPTI keyboard layout which we will not discuss here. We also got numerous comments specifically addressing the composition task (Table IV) and the transcription task (Table V). Opinions were mixed; some participants preferred composition while others preferred transcription.

The experiment showed there was no statistically significant difference in entry rate between the composition task and the transcription task. However, we suggest caution when interpreting this result. First, a failure to reject the null hypothesis does not imply the means are identical. Rather, it implies that there is insufficient evidence to exclude the possibility that the means differ due to chance. Second, we exposed participants to an unfamiliar text entry method for only a brief period of time, which may have resulted in a floor effect.

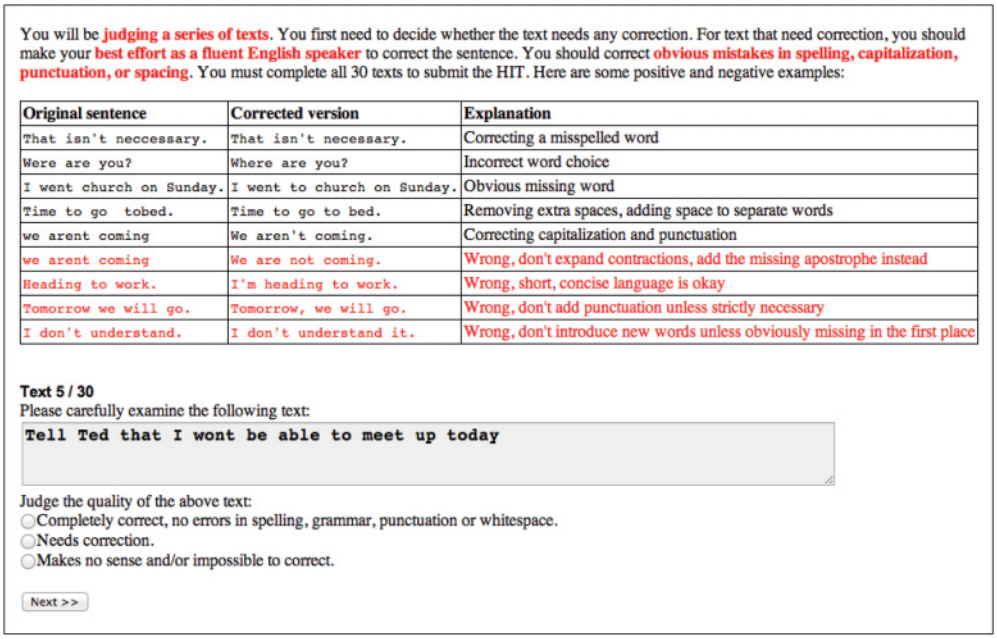
5. EXPERIMENT 4: JUDGING COMPOSITIONS

One possible concern of a composition entry task is how to measure the error rate. In this section, we explain how we used crowdsourcing to provide human judging of the compositions written in Experiment 3. We have made the code we used for this process publicly available[4].

5.1. Method

Our approach was to have 10 workers on Amazon Mechanical Turk judge and correct each composition. Workers did this in sets of 30 compositions for a payment of $0.20. Workers were required to have a 95% accepted HIT rate. We limited the HIT to workers located in the United States.

[4]http://keithv.com/software/judging


You will be judging a series of texts. You first need to decide whether the text needs any correction. For text that need correction, you should make your best effort as a fluent English speaker to correct the sentence. You should correct obvious mistakes in spelling, capitalization, punctuation, or spacing. You must complete all 30 texts to submit the HIT. Here are some positive and negative examples:

Original sentence	Corrected version	Explanation
That isn't neccessary.	That isn't necessary.	Correcting a misspelled word
Were are you?	Where are you?	Incorrect word choice
I went church on Sunday.	I went to church on Sunday.	Obvious missing word
Time to go tobed.	Time to go to bed.	Removing extra spaces, adding space to separate words
we arent coming	We aren't coming.	Correcting capitalization and punctuation
we arent coming	We are not coming.	Wrong, don't expand contractions, add the missing apostrophe instead
Heading to work.	I'm heading to work.	Wrong, short, concise language is okay
Tomorrow we will go.	Tomorrow, we will go.	Wrong, don't add punctuation unless strictly necessary
I don't understand.	I don't understand it.	Wrong, don't introduce new words unless obviously missing in the first place

Text 5 / 30
Please carefully examine the following text:
Tell Ted that I wont be able to meet up today
Judge the quality of the above text:
Completely correct, no errors in spelling, grammar, punctuation or whitespace.
Needs correction.
Makes no sense and/or impossible to correct.
Next >>

Fig. 12. The task used to judge and correct compositions. If the worker selected "Needs correction," a second text entry area appeared allowing the user to edit the original text.

Table VI. Examples of Some of the Injected Errors and the Corrections We Considered Acceptable

Sentence with injected error	Acceptable correction(s)
How is you new job working out?	How is your new job working out?
you're going to fail imediately if you do.	You're going to fail immediately if you do.
I saw yg.ou at the meetin	I saw you at the meeting.
Call me when you get chance.	Call me when you get a chance.
	Call me when you get the chance.
are you free next friday for dinner?	Are you free next Friday for dinner?

We asked workers to rate each composition on a 3-point scale: 2 = completely correct, 1 = needs correction, or 0 = uncorrectable. If a composition was scored a 1, the worker was instructed to edit the text to correct obvious errors in spelling, grammar, punctuation, or whitespace. The exact instructions we gave workers evolved over the course of several pilot experiments. Figure 12 shows our final instructions and the judging interface.

In order to help ascertain how good each worker was at spotting errors, in each set of 30 compositions we injected 10 sentences with a known set of likely corrections. This allowed us to estimate workers' judging accuracy and eliminate workers suspected of doing a poor job due to inattention, poor language skills, and so on. Table VI shows some examples of the injected errors and the corrections we considered acceptable.

To measure errors, we used Character Error Rate (CER). CER is the number of character insertions, deletions, and substitutions required to transform the participant's original text into the judge's corrected text, divided by the number of characters in the judge's text. If a sentence was judged completely correct, the CER was taken as 0%. If a sentence was judged uncorrectable, the CER was taken as 100%. If correctable, we took the judge's correction as the reference and computed the CER of the composition.

For each composition, we define the *judged CER* as the median of all the judges' error rates for that composition. We used the median because the distribution of the

Table VII. Judged CER, Original Composition and the Judges' Corrections

CER (%)	Original composition followed by correction(s)	Number judges
1.6	Please don't forgee about our appointment tomorrow afternoon.	
	Please don't forget about our appointment tomorrow afternoon.	7
2.9	You want to go to the movies todby?	
	You want to go to the movies today?	6
	Do you want to go to the movies today?	1
5.3	The weather is kind of teribble today.	
	The weather is kind of terrible today.	7
	(judged correct)	1
9.7	i am tired and i hate the snow	
	I am tired and I hate the snow.	7
	I'm tired; I hate the snow.	1
	I am tired, and I hate the snow.	1
12.0	i wil l be there in a min	
	I will be there in a minute.	3
	I will be there in a min.	3
	I will be there in a min	1
100.0	fetRealy stick	
	(judged uncorrectable)	6
	Feet realy stink	1
	Felt really sick.	1

judges' error rates for a particular sentence tends to be asymmetric. This is because the majority of judges tend to agree on one particular correction while a minority decide on some other correction.

5.2. Results

Seventy-two unique workers took part in the experiment. We rejected one worker who submitted the same corrections for multiple compositions. It took workers, on average, 6 minutes to judge and correct a set of 30 sentences. Workers had to get 70% of known corrections exactly correct in order to be included in the pool of judges. This reduced our pool to 60 unique judges. For each of the 591 compositions, we had between 7 and 10 judges.

Judges rated 79% of compositions as completely correct, 20% as needing correction, and only 1% as being uncorrectable. The mean judged CER over all compositions was 1.5% (sd 6.6). We found that 81% of compositions had a judged CER of 0%. Table VII shows some examples of compositions and the judges' corrections. As can be seen, in most cases, the majority of judges provided a very sensible correction of the composition. However, in some cases judges made subjective corrections to punctuation and wording despite our instructions not to do this.

6. DISCUSSION

Experiments 3 and 4 showed that we can test a novel text entry method using a composition style task. Our participants wrote sensible short messages with relatively few errors. Further, we demonstrated that these compositions could be judged, either by the experimenters, or via a crowdsourcing judging protocol.

As alluded to earlier we do not advocate replacing the transcription task with a composition task. Rather, we propose complementing the methodological toolbox for text entry evaluations with composition. We have demonstrated composition can provide remarkably stable results for a wide variety of participants.

To be able to test many conditions and variations on a wide sample from the population, we crowdsourced our experiments on the Amazon Mechanical Turk microtask market. This crowdsourcing approach was inspired by previous successful

crowdsourced HCI evaluations (e.g., Kittur et al. [2008] and Heer and Bostock [2010]). While crowdsourcing does not, for obvious reasons, enable us to have full control over an experiment, other researchers have successfully carried out a wide array of tasks and experiments using crowdsourcing. A particularly striking example is the demonstration by Heer and Bostock [2010] that crowdsourcing can replicate several research results obtained from the graphical perception literature.

An open research question is how composition task performance evolves in longitudinal experiments over several sessions. One hypothesis is that participants become increasingly competent at composing text as a function of practice, which might reduce the variance among participants and increase composition task entry rates in comparison to transcription tasks. However, more research is required to shed light on this question.

Another open issue is how well the composition task models what users actually write in practice. Our most effective composition task was to let participants compose short messages. However, in actual use, participants may want to write essays, longer emails, and other documents. There may, therefore, be alternative composition tasks we have not considered that could also be effective in discerning participants' composition entry rates.

Last, it is worth emphasizing that not all text entry evaluations use a baseline text entry method as a control condition (e.g., Kristensson and Vertanen [2012b]). When the purpose of an evaluation is merely to get a sense of the capabilities of a text entry method rather than showing it performs significantly better than an established baseline, the composition task may be more informative. For example, participants composing their own text messages may uncover unexpected deficiencies, such as an inability to input certain characters or words.

7. RECOMMENDED COMPOSITION TASK METHOD

We compared a variety of different tasks that elicit freeform compositions from participants. We found that providing participants with a simple instruction of creating a short message in the domain of interest was successful in getting participants to quickly invent and compose text. It does not appear necessary to provide participants with a specific situation or message in order to help them invent a message.

For example, in our COMPOSE condition in Experiment 1, we used the instruction: "Imagine you are using a mobile device and need to write a message. We want you to invent and type in a fictitious (but plausible) message. Use your imagination. If you are struggling for ideas, think about things you often write about using your own mobile device."

In order to make it easier to estimate the error rate of participants' compositions, it is advisable to add the instruction: "Please write complete sentences with good grammar and spelling. Do NOT use texting abbreviations or slang." We found in Experiment 2 that participants responded to this instruction and avoided the use of abbreviations.

To estimate the error rate of compositions, we recommend having multiple people judge each composition. Judges should be instructed to correct obvious mistakes using their knowledge as a fluent speaker of the language (see Figure 12). The judges' corrected compositions can then be aggregated into a *judged CER* using the median of each individual judge's CER. As we demonstrated in Experiment 4, it is possible to crowdsource the judging process.

8. CONCLUSIONS

In this article, we have proposed a new methodology to add to the text entry evaluation toolbox: the composition task. While the transcription task has higher internal validity, the composition task has higher external validity. Thus, these tasks occupy

different points in the tradeoff between internal and external validity. Which task is more appropriate depends on the research questions we want to answer.

Do we want to understand how text entry methods behave when they are used in situations close to actual practice, that is, when users are writing their own messages and emails? When answering this research question, we have shown that the composition task is a remarkably reliable method. First, participants write compositions that are original (Experiment 1, 2, and 3). Yet, at the same time, these compositions can be sufficiently restricted so that individual text styles do not vary much (Experiment 2 and 3). Second, participants' compositions can be reliably judged by either the experimenter conducting the evaluation or by crowdsourced workers (Experiment 4). This last step is crucial because it is vital that the error rate can be sufficiently controlled in order to ensure it is possible to compare two different text entry methods' entry rates. Thus, it is possible to use the composition task to compare two text entry methods against each other. Alternatively, a single text entry method can be examined in a large-scale composition task experiment conducted via, for example, crowdsourcing to get an idea of the realistic entry rate distribution in the general population.

However, due to the inescapable tradeoff between internal and external validity, the composition task method we have presented here is not suitable for fine-grained low-level comparisons between similar text entry methods. For instance, if one wishes to test subtle differences in two touchscreen keyboard error correction algorithms, it is likely that the standard transcription task would provide more reliable results.

In the end, a field as diverse as HCI requires multiple approaches to tackle a vast array of research questions. This is not a new argument (see, e.g., Dix [2010]), but it is important to highlight. A text entry method—as any other user interface—has many design dimensions and a single experimental methodology, such as the transcription task cannot, due to the need to control for confounding factors and to ensure internal validity, possibly examine all of them. For example, Kristensson [2007] proposes and analyzes 22 design dimensions for mobile text entry and many of these dimensions regard real-world constraints. The composition task enables us to better understand a largely neglected design dimension: users' text entry rates in actual practice. We hope our four experiments have convinced researchers that despite several initial reservations raised in the beginning of this article, it is indeed possible to use the composition task in text entry evaluations.

APPENDIX

A. STIMULI TEXT IN EXPERIMENT 1

In Experiment 1, two of the composition conditions asked participants to respond to a given message or situation. Here we list the set of possible stimuli for these two conditions. These stimuli may be useful for other researchers conducting similar short composition style text entry experiments. They are also available online.[5]

A.1. Reply Condition

(1) The bus is late again, and I'm getting very irritated!
(2) Hey i got 2/3 but not d minus sign.
(3) Wanna go to the park?
(4) I hope that you are okay.
(5) What do u want me to cook tonight? Chicken ok?
(6) almost there about 15 mins more. sorry.
(7) where are you now?I miss you a lot darling!

[5]http://keithv.com/software/composition

(8) i cant come til after 10
(9) play while play work while work
(10) ill be home around 23:30 because i just missed the bus.
(11) Can't reach you, please call me back.
(12) Did you see those Uggs she had on with a tight skirt and fishnet stockings?
(13) partu at my house, this saturday, hope to see u there. bye
(14) be there in 5
(15) what time u get off from work
(16) hai how r u what u think of hanging out today?. huh
(17) Great talkers are little doers- those people who talk a lot and always teaching others usually do not do much work
(18) how long do i have?
(19) Hey beautiful, will u b able to join me for dinner?
(20) my wallet is lost
(21) I want to come personally to meet you at your home
(22) can you stop at the store for me? im runnin late
(23) pete just told me he'll b there
(24) Nah, I'm at the cafe.
(25) i love you
(26) howve you been? its been awhile
(27) What time?
(28) Hey whats up?? its been ages
(29) have u seen my phone?
(30) I'm running late.
(31) DUDE!! I beat the game today
(32) hi, whatcha doing?
(33) Please send the the date, time and venue of the party.
(34) hey get me my laptop tomorrow man... i am gonna need it. And my sister told me to do some paper work for her..please bring it tomorrow..after that you can use it for your project
(35) When are you going to come over here
(36) R U here yet?
(37) I'd rather have Chinese food tonight.
(38) Hey when you gonna be home?
(39) Hi r u awake? If yes can u tell me what happened today in d class?
(40) Can you take me to the dentist at 4? Mom just told me she wont be at home on time :-(
(41) how is your job going on
(42) im gonna skip today
(43) Is everything allright with u? Havent heard from u in a while. Call me back.
(44) What movie did you end up seeing last night?
(45) This show is hilarious.
(46) have u seen my phone
(47) hi how r u... how is your job going on...
(48) why don't u reveal ur identity
(49) What kind of pizza do you want?
(50) a terrible song is on right now
(51) Happy birthday bro!! Hope you have an awesome day :) xxx
(52) please call me after you reach home
(53) I just finished tht quiz. very hard
(54) cant go 2nite gtg bathe my grandma :P
(55) Leave the shopping list on the table.

(56) Missed my flight, so will be late. Don't worry.
(57) Are you thinking of me?
(58) i heard that u're coming to NYC, call if u want to hang out tonight, kissess paul.
(59) Let's go to lunch together. 12 ok?
(60) pick up some milk please
(61) where is mom
(62) nothing much what about you?
(63) can you come for cup of coffe
(64) Hi how are you. Do you get my gift.
(65) call this number 4 work
(66) wut movie do u want 2 see?
(67) My sister is in town, so I'm having dinner with her 2nite.
(68) Many Happy Returns of the Day. Happy Birthday Sarah!
(69) I think I like her.
(70) okay wanna meet at 8?
(71) Heard about floods in your area. Hope you are safe, am worried.
(72) is tony coming? need 2 know asap
(73) These papers you gave me are full of useless stuff. Try to be more imaginative next time.
(74) not heard fm u
(75) How about pasta for dinner?
(76) Where's a good place to hang out this weekend?
(77) OMG is he okay? let me know what happens!
(78) want to meet for drinks 2morrow nite?
(79) Meet me in the park in 20 minutes! I'll be right next to the hotel. Don't be late!
(80) Why haven't you called? I want to talk.
(81) Whatev loser!
(82) when will you come back?
(83) man, stop with the drugs, those things are killing you...
(84) Kids want bananas. Can u buy some and bring here?
(85) Hey, where r u?
(86) sorry can't im busy
(87) Kindly send me the file.
(88) The electricity keeps going out because of the wind.
(89) Celebration tonight at my place! Be there at 8.
(90) U r cute!
(91) call me as soon as you see this
(92) All right, I'm totally lost and some weird dude is looking at me. Map quest me!
(93) Can we carpool tomorrow?
(94) Haha Dude switch to channel 5, you'll love it ;) x
(95) have u tried mturk.com easy money. love it! lol
(96) hey i have missed you alot whats up
(97) When do you leave for Florida?
(98) Your email was weird.
(99) s i do agree compatibility is psble with adjustments
(100) wanna go 2 six flags?
(101) call me later after work
(102) Please call after 5 PM.
(103) hey dude brb ill call you when im home
(104) I love that new song she just put out but they don't play it enough.
(105) what are you thinking about
(106) What time does the movie start?

(107) 8pm at the Old Pony for a couple?
(108) what made him say that
(109) wanna go to the store
(110) can you meet me in 10
(111) CONGRATULATIONS! AND CELEBRATIONS! I HEARD OF YOUR GETTING JOB! WHEN R U GIVING PARTY?
(112) what made her say that thing to you
(113) I was awake all night reading that book. Now I need a strong coffee.
(114) I nearly cried when he slipped in the last corner of the race.
(115) where are going i came to your room you wasn't there....where are you know? message me or call me dear
(116) Off to bed. good night.
(117) come play wow with us, we are on a new private serverm i think you'll enjoy...
(118) What do you want to do tonight?
(119) Where do you want to meet?
(120) no, u call me
(121) Please chek your email, so that there is a message for u
(122) Wish you and ur family a happy and prosperous New Year.
(123) worries don't reduce yesterday's sorrows.but it reduces today's strength.so don't worry.Be happy.
(124) there's a website that pays you for writing reviews and other things, u need to see that. it's awesome!
(125) tonight will be our fisrt date.... im gonna get crazy *-*
(126) It is heavy rain here.That's why cool now.What about there?
(127) where are the boys?
(128) good morning! wassup? slept good? :)
(129) HEY ? R U
(130) need to vst studio b4 picking you up. how about 330?
(131) Congrats on ur promotion. Waiting to get a treat from you.
(132) tried... going 2 bed. love u
(133) Don't forget to print out your grades for your father.
(134) WAT R U DNG
(135) Hey, i heard you broke your arm, you alright?
(136) sorry i didnt hear you call! everything ok?
(137) dinner?
(138) where are you?
(139) What time will you be off work? Miss you.
(140) Where did you leave the Lucky Charms?
(141) are you commited?
(142) Hi Frnds, gud mrng & ms u lot
(143) Good luck today!
(144) thnks...can we get 2 know each other by email
(145) Meet you in 10 minutes, don't be late.
(146) Please call me. We need to talk.
(147) here heavy rain
(148) Always keep a check on your words before it comes out of the mouth as words once spoken cannot be taken back.
(149) Hi! How are you? Long time. Let's catch up sometime?
(150) Hey wats up r u gonna b @ tha party 2nite i hope so tha last 1 was gr8 lol, call me l8r and let me kno wat you plan on doin l8r.
(151) Can I come to collect my payment today?
(152) How is weather now there?It is very hot here!

(153) Just thought I would let you know that I love you and I miss you and I can't wait to see you later :)
(154) watz good
(155) What were you thinking?
(156) son,make me proud
(157) Hai sweet heart I'm late tonight please have your dinner and sleep. Take care bye ...bye....
(158) Trip is going really well - meetings have been long but productive.
(159) looking forward to see u
(160) when r u coming?
(161) who is this
(162) hey how are you?
(163) nice pic. groose
(164) Don't put all your money on Boston in the finals!
(165) what made you think that
(166) Pick up bread and milk on the way home please
(167) i just bought me a new phone testing it out by texting u
(168) friendships are opened doors,each with a different view.but none could be more beautifull view than the door that leads to you.
(169) please pick up some dinner on your way home.
(170) How was Mexico?
(171) R U there yet?
(172) will b late... lost my wallet!
(173) Are we ever going to meet? This is the third time youve cancelled :(Im starting to think u dont like me.
(174) Lunch tomorrow?
(175) Crap. I got a flat tire. I'm waiting for AAA.
(176) Where r u rght now
(177) Hey 2mrw we have which lecture?
(178) Late like running behind or late like having a baby?
(179) what a day it was Sachin made a miracle in the field first he batted throughout the innings and next he created a world record by scoring first double ton Yummmy
(180) Are you going to wear something flirty tonight?
(181) i am between love and hate
(182) Sorry that u are sick - Hope u feel better soon!
(183) What time will you be back?
(184) What r u doing
(185) Hey do you wanna go grab a cup of coffee? I have some amazing gossip I need to tell somebody x
(186) let go to this club near my house
(187) the teacher will come at 8 a clock
(188) Do you want a ride or can you walk?
(189) are u coming to school today? i thought will not becoz u was ill yesterday
(190) Did the band play my favorite song?
(191) what are you doing
(192) congratulation you past the exams
(193) I can't find it. SRY
(194) waz up?
(195) The phone is not working, did u pay the bill?
(196) Did I just see you drive by?
(197) do not 4get me on t occasion 2moro
(198) just heard u got back. how was it?

(199) I won playing some poker today
(200) Nice people are like wind.You can never see them.But You will always feel their prescence.

A.2. Situation Condition

(1) Your housemate has been sick for the last week. You are currently shopping downtown. See if he requires anything.
(2) You want to have lunch with Laura and would also like to see her trekking photos from Nepal. Arrange to do both.
(3) Your electricity bill must be paid every month. It is your housemate's responsibility this month. Inform him of the situation.
(4) Your friend is picking you up at the airport but you are still waiting for your bags. Inform your friend of the situation.
(5) You co-worker possess the latest electronic versions of your vendor's price lists. You require these lists. Make a request to your co-worker.
(6) Your friend Carol is at the local sandwich shop. Request a ham and cheese baguette.
(7) You are playing golf when there is a loud boom and a flash of light in the sky. Make a suggestion to your group.
(8) You are driving to visit Katie. Last time it took 45 minutes to get there without any traffic. Inform Katie of the situation.
(9) Your friend is in hospital with a dislocated shoulder. Send her your sympathies.
(10) You will be out of the office for the next week. Your co-worker Jackie will handle any crucial issues. Inform people of the situation.
(11) Your household is trying to reduce its electric bill. Compact fluorescent light bulbs utilize significantly less energy. Make a suggestion.
(12) Your friend is in trouble with the law. You have had good experience with the law firm of Thornberg and Sutley. Make a suggestion.
(13) You were supposed to attend the three o'clock movie with Megan. You failed to get to your bus before it departed. Inform Megan of the situation.
(14) You are planning to take a trip by cycle to the apple orchard south of town. Paula just bought a new cycle. Make a suggestion to Paula.
(15) Your best customer is coming to visit and needs to be picked up at the airport. Make a request to the office administrator to handle the situation.
(16) Your friend wants to do something tonight. You had planned to watch Casablanca on television tonight. Make a suggestion.
(17) Your co-worker Carl will be 29 years old today. Send him a greeting.
(18) You are expecting a call from a friend but your mobile phone is almost out of electricity. You will be with Michael who also has a phone. Inform your friend of the situation.
(19) The prototype of the new hydraulic press will not be ready for testing for another two weeks. Inform your co-worker of the situation.
(20) You are meeting with a customer on the twelfth flour of your hotel in the Golden Daffodil room. Give your customer directions.
(21) Your phone number is 321 1942. You want to communicate with somebody in accounts payable. Leave a message.
(22) You are expecting an important package to be delivered to your home. You are currently not at home but your housemate is. Make a request to your housemate.
(23) You normally feed and walk your dog after work. You have to work late but your housemate is home. Make a request to your housemate.
(24) You are taking your friend Vanessa to your parents for dinner. She is allergic to all types of shellfish. Inform your parents of the situation.

(25) You have booked a holiday next Thursday and Friday. Your boss Sara is asking for people's upcoming availability. Inform Sara of the situation.
(26) You need to provide food for a meeting for ten people, three of whom are vegetarians. Tell the caterer what you need.
(27) You have an all day dentist's appointment on February 28th. Inform your colleagues.
(28) You are meeting your friend whose initials are RAV for lunch at Lui's. Tell your other friend your plans.
(29) Your company has bought new adjustable height desks. Your knees are hitting the bottom of your new desk. Make a request to the maintenance department.
(30) Let your housemate know that the concert was sold out and you couldn't get any more tickets.
(31) Your classmate wants to go out drinking tonight. You have a big Spanish exam on Monday and need to study. Inform your classmate of the situation.
(32) You have arranged to donate blood today between one and two pm. Inform your co-workers when you will not be present.
(33) You are building a shed behind your house and a battery powered drill is needed. Your neighbour Wilson has such a drill. Make a request to Wilson.
(34) You have just left on holiday. You are worried you did not turn the oven off. Make a request to your housemate.
(35) You are to meet your classmate at 4:30 in the university library. Confirm the time and location.
(36) You have just completed reading the book Aesop's Fables. Your friend has asked if you've been doing any reading lately. Inform her of the situation.
(37) Despite very cold weather, you and Amy are planning a snowshoe trip this Saturday. You understand Gooseberry Falls is attractive when frozen. Make a suggestion.
(38) Julie wants to go to the beach tomorrow. You prefer to go to the beach only when it's sunny. Inform Julie of the situation.
(39) Your friend was victorious in his football match today. Wish him good fortune at the semi-finals tomorrow.
(40) There is a new coffee shop on Westfield road. You want to drink coffee with Samantha. Make a suggestion.
(41) There is a meteor shower tonight and you have a telescope. Juan asked if you were going to watch the shower. Make a suggestion to Juan.
(42) Your car won't start. Your co-worker drives in close proximity to your house on their way to work. Arrange your transport to work.
(43) You tripped and hurt your ankle yesterday. The doctor today indicated it was a bad sprain. Inform Dianne of the situation.
(44) After weeks of coughing, your doctor has diagnosed you with pneumonia and put you on antibiotics. Lara has inquired about your health. Respond to Lara.
(45) You require plain yogurt and bay leaves to prepare dinner tonight. Tell Pat to acquire these ingredients.
(46) Joseph borrowed some music CDs over a year ago. Make a demand to Joseph.
(47) Aubrey handles procurement of supplies in your office. Your printer no longer contains magenta or cyan ink. Make a request to Aubrey.
(48) Tim got lost on his way to your house. He is currently five streets south of your street, Chestnut Avenue. Tell him what to do.
(49) Your enjoy collecting antique coffee bean grinders. You currently have 148. Describe your collection.
(50) Tonight there will be a big fireworks display downtown. Last year you could see them from the patio of your building. Make a suggestion to your friends.

(51) The last problem in this week's assignment is similar to the example in chapter three. Provide advice to your classmate.
(52) A client has inquired about the details of their last contract. You can't access your computer because the electricity has failed. Inform your customer of the situation.
(53) In the third quarter, your company's sales decreased to 2.7 million. Inform your management of the situation.
(54) Your dog Drea got lacerations on her hind legs crawling through a barbed wire fence. Ask your veterinarian friend for advice.
(55) Warren has lost his billfold. You observed something that might have been a billfold next to the tea kettle. Inform Warren of the situation.
(56) Ask Peter if he can acquire brown bread and a couple litres of skimmed milk on his way home.
(57) Agree to meet for dinner at the Black Ox after you finish work at five.
(58) Your laptop computer has stopped working again. Last time Bruce came to your residence to troubleshoot it. Make a request to Bruce.
(59) You cannot locate your keys and can't gain entry to your building. Richard is upstairs and possibly sleeping. Make a request of Richard.
(60) You live in a brick house at 2727 Whispering Willows Road. Kyle is coming over to visit. Describe your residence.
(61) You are applying for an account and are asked to provide a piece of personal information. Your mother's maiden name is Waycaster. Make a suggestion.
(62) You are meeting four friends tonight at 8:30 for dinner. Your friend is handling the dinner reservation. Tell her what to do.
(63) You are estimating the cost of some repair work for your best customer James Abdalla. Your boss is inquiring what you are occupied with. Inform him of the situation.
(64) Ethan has an important job interview tomorrow. Provide reassurance to Ethan.
(65) Your friend wants to come over to visit. You are leaving now to swim at the gym. Inform your friend of the situation.
(66) You are traveling by train to your cousin Linda's and are two stops away. Inform Linda of your situation.
(67) Betty operates a farm and also gives horseback riding lessons. You want to learn to ride. Make a request to Betty.
(68) Your classmate is on his way to buy a used copy of the class textbook. You would also like a used copy. Inform your classmate of the situation.
(69) Yesterday, you took a pottery making class. After working on your coil pot all day, it exploded in the kiln. Tell Brian what you did yesterday.
(70) You can no longer recollect your computer password. Victor administers your company's computers. Ask Victor for help.
(71) You would like Tyler to give a half-hour presentation at the annual company meeting on his Zero Injuries safety project. Make a request to Tyler.
(72) Check with your friend Carolyn on the current weather conditions downtown.
(73) Your sister is shopping for your father's birthday. You saw a beautiful bamboo fly fishing rod on sale at a local shop. Make a suggestion.

B. DEMOGRAPHIC DATA FROM EXPERIMENT 1

Table VIII. Self-Reported Gender of Participants in Experiment 1

Sex	Number	Percentage
female	96	53.0
male	85	47.0

Table IX. Self-Reported Country of Participants in Experiment 1

Country	Number	Percentage
United States	115	63.5
India	33	18.2
United Kingdom	7	3.9
Serbia	4	2.2
Philippines	2	1.1
Switzerland	2	1.1
Pakistan	2	1.1
Canada	2	1.1
Bangladesh	2	1.1
Indonesia	2	1.1
Dominican Republic	1	0.6
Australia	1	0.6
Trinidad and Tobago	1	0.6
New Zealand	1	0.6
Mexico	1	0.6
Thailand	1	0.6
Barbados	1	0.6
Singapore	1	0.6
Romania	1	0.6
Macedonia	1	0.6

Table X. Self-Reported Age of Participants in Experiment 1

Age range	Number	Percentage
0–19	17	9.4
20–24	56	30.9
25–29	36	19.9
30–34	34	18.8
35–39	12	6.6
40–44	13	7.2
45–49	4	2.2
50–54	5	2.8
55–59	2	1.1
60–64	2	1.1

Table XI. Self-Reported English Proficiency of Participants in Experiment 1

English proficiency	Number	Percentage
native	128	70.7
advanced	36	19.9
moderate	17	9.4
beginner	0	0.0

Table XII. Self-Reported Amount That Participants "Looked at Keyboard While Typing" in Experiment 1

Looked at keyboard	Number	Percentage
sometimes	106	58.6
never	52	28.7
always	23	12.7

Table XIII. Self-Reported Typing Speed of Participants in Experiment 1

Typing speed	Number	Percentage
moderate	118	65.2
fast	57	31.5
slow	6	3.3

Table XIV. Self-Reported Computer Type of Participants in Experiment 1

Computer type	Number	Percentage
laptop	100	55.2
desktop	81	44.8
mobile device	0	0.0
tablet	0	0.0
other	0	0.0

C. DEMOGRAPHIC DATA FROM EXPERIMENT 2

Table XV. Self-Reported Gender of Participants in Experiment 2

Sex	Number	Percentage
female	51	54.3
male	43	45.7

Table XVI. Self-Reported Country of Participants in Experiment 2

Country	Number	Percentage
United States	46	48.9
India	27	28.7
Canada	5	5.3
Australia	3	3.2
United Kingdom	3	3.2
Philippines	1	1.1
Germany	1	1.1
Sweden	1	1.1
Trinidad and Tobago	1	1.1
Serbia	1	1.1
Mexico	1	1.1
Finland	1	1.1
Portugal	1	1.1
The Former Yugoslav Republic of Macedonia	1	1.1
Romania	1	1.1

Table XVII. Self-Reported Age of Participants in Experiment 2

Age range	Number	Percentage
0–19	7	7.4
20–24	18	19.1
25–29	21	22.3
30–34	20	21.3
35–39	12	12.8
40–44	5	5.3
45–49	3	3.2
50–54	5	5.3
55–59	1	1.1
60–64	2	2.1

Table XVIII. Self-Reported English Ability of Participants in Experiment 2

English ability	Number	Percentage
native	57	60.6
advanced	22	23.4
moderate	15	16.0

Table XIX. Self-Reported Amount That Participants "Looked at Keyboard While Typing" in Experiment 2

Looked at keyboard	Number	Percentage
sometimes	63	67.0
never	21	22.3
always	10	10.6

Table XX. Self-Reported Typing Speed of Participants in Experiment 2

Typing speed	Number	Percentage
moderate	66	70.2
fast	26	27.7
slow	2	2.1

Table XXI. Self-Reported Computer Type of Participants in Experiment 2

Computer type	Number	Percentage
laptop	50	53.2
desktop	44	46.8

D. DEMOGRAPHIC DATA FROM EXPERIMENT 3

Table XXII. Self-Reported Gender of Participants in Experiment 3

Sex	Number	Percentage
female	36	72.0
male	14	28.0

Table XXIII. Self-Reported Country of Participants in Experiment 3

Country	Number	Percentage
United States	50	100.0

Table XXIV. Self-Reported Age of Participants in Experiment 3

Age range	Number	Percentage
0–19	6	12.0
20–24	14	28.0
25–29	14	28.0
30–34	3	6.0
35–39	7	14.0
40–44	3	6.0
50–54	2	4.0
55–59	1	2.0

Table XXV. Self-Reported English Ability of Participants in Experiment 3

English ability	Number	Percentage
native	49	98.0
advanced	1	2.0

Table XXVI. Self-Reported Amount That Participants "Looked at Keyboard While Typing" in Experiment 3

Looked at keyboard	Number	Percentage
sometimes	28	56.0
never	20	40.0
always	2	4.0

Table XXVII. Self-Reported Typing Speed of Participants in Experiment 3

Typing speed	Number	Percentage
moderate	31	62.0
fast	17	34.0
slow	2	4.0

Table XXVIII. Self-Reported Computer Type of Participants in Experiment 3

Computer type	Number	Percentage
laptop	37	74.0
desktop	12	24.0
tablet	1	2.0

REFERENCES

Tamara Broderick and David J. C. MacKay. 2009. Fast and Flexible Selection with a Single Switch. *PLoS ONE* 4, 10 (10 2009), e7481.

Steven J. Castellucci and Scott I. MacKenzie. 2008. Graffiti vs. unistrokes: an empirical comparison. In *CHI'08: Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM Press, 305–308.

Edward Clarkson, James Clawson, Kent Lyons, and Thad Starner. 2005. An empirical study of typing rates on mini-QWERTY keyboards. In *CHI'05: Extended abstracts on Human Factors in Computing Systems*. ACM Press, 1288–1291.

Alan Dix. 2010. Human-computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with Computers* 22 (January 2010), 13–27. Issue 1.

Mark D. Dunlop and Michelle Montgomery Masters. 2009. Pickup Usability Dominates: A Brief History of Mobile Text Entry Research and Adoption. *International Journal of Mobile Human Computer Interaction* 1, 1 (2009), 42–59.

Sandra G. Hart and Lowell E. Stavenland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, P. A. Hancock and N. Meshkati (Eds.). Elsevier, Chapter 7, 139–183.

Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI'10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 203–212.

Poika Isokoski and Timo Linden. 2004. Effect of foreign language on text transcription performance: Finns writing English. In *NordiCHI'04: Proceedings of the Third Nordic Conference on Human-Computer Interaction*. ACM Press, 109–112.

Poika Isokoski and Roope Raisamo. 2000. Device independent text input: a rationale and an example. In *AVI'00: Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM Press, 76–83.

Akiyo Kano, Janet C. Read, and Alan Dix. 2006. Children's phrase set for text input method evaluations. In *NordiCHI'06: Proceedings of the Fourth Nordic Conference on Human-Computer Interaction*. ACM Press, 449–452.

Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In *CHI'99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 568–575.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI'09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 453–456.

Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *European Conference on Machine Learning*. 217–226.

Per Ola Kristensson. 2007. *Discrete and Continuous Shape Writing for Text Entry and Control*. Ph.D. Dissertation. Linköping University.

Per Ola Kristensson. 2009. Five Challenges for Intelligent Text Entry Methods. *AI Magazine* 30, 4 (2009), 85–94.

Per Ola Kristensson and Leif C. Denby. 2009. Text entry performance of state of the art unconstrained handwriting recognition: a longitudinal user study. In *CHI'09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 567–570.

Per Ola Kristensson and Keith Vertanen. 2012a. Performance Comparisons of Phrase Sets and Presentation Styles for Text Entry Evaluations. In *IUI'12: Proceedings of the International Conference on Intelligent User Interfaces*. ACM Press, 29–32.

Per Ola Kristensson and Keith Vertanen. 2012b. The Potential of Dwell-Free Eye-Typing for Fast Assistive Gaze Communication. In *ETRA'12: Proceedings of the ACM Symposium on Eye-Tracking Research and Applications*. 241–244.

Kent Lyons, Thad Starner, and Brian Gane. 2006. Experimental evaluations of the Twiddler one-handed chording mobile keyboard. *Human-Computer Interaction* 21 (November 2006), 343–392. Issue 4.

I. Scott MacKenzie and R. William Soukoreff. 2002. Text Entry for Mobile Computing: Models and Methods, Theory and Practice. *Human-Computer Interaction* 17 (2002), 147–198.

Ian Scott MacKenzie and William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI'03: Extended Abstracts on Human Factors in Computing Systems*. ACM Press, 754–755.

I. Scott MacKenzie and Shawn X. Zhang. 1999. The design and evaluation of a high-performance soft keyboard. In *CHI'99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, USA, 25–31.

Tim Paek and Bo-June Hsu. 2011. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In *CHI'11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 2477–2480.

R. W. Soukoreff and I. S. MacKenzie. 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *CHI'03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 113–120.

Keith Vertanen and Per Ola Kristensson. 2009. Parakeet: A Continuous Speech Recognition System for Mobile Touch-Screen Devices. In *IUI'09: Proceedings of the* 14^{th} *International Conference on Intelligent User Interfaces*. ACM Press, 237–246.

Keith Vertanen and Per Ola Kristensson. 2011a. The imagination of crowds: conversational AAC language modeling using crowdsourcing and large data sources. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*. ACL, 700–711.

Keith Vertanen and Per Ola Kristensson. 2011b. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *MobileHCI'11: Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM Press, 295–298.

David J. Ward and David J. C. MacKay. 2002. Fast Hands-free writing by Gaze Direction. *Nature* 418, 6900 (2002), 838.

Jacob O. Wobbrock, Duen Horng Chau, and Brad A. Myers. 2007. An Alternative to Push, Press, and Tap-tap-tap: Gesturing on an Isometric Joystick for Mobile Phone Text Entry. In *CHI'07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 667–676.

Shumin Zhai, Per Ola Kristensson, and Barton A. Smith. 2005. In search of effective text input interfaces for off the desktop computing. *Interacting with Computers* 17, 3 (2005), 229–250.

Received June 2013; revised November 2013; accepted December 2013