Journal of Visualized Experiments

An assessment method and toolkit to evaluate keyboard design on smartphones

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE61796R2
Full Title:	An assessment method and toolkit to evaluate keyboard design on smartphones
Corresponding Author:	Jibo He Tsinghua University Beijing, China CHINA
Corresponding Author's Institution:	Tsinghua University
Corresponding Author E-Mail:	hejibo@live.com;jibo.he@wichita.edu;hejibo@gmail.com
Order of Authors:	Yincheng Wang
	Ke Wang
	Yuqi Huang
	Di Wu
	Jian Wu
	Jibo He
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Beijing,Beijing, China
Please confirm that you have read and agree to the terms and conditions of the author license agreement that applies below:	I agree to the Author License Agreement
Please specify the section of the submitted manuscript.	Behavior
Please provide any comments to the journal here.	

TITLE:

An Assessment Method and Toolkit to Evaluate Keyboard Design on Smartphones

AUTHORS AND AFFILIATIONS:

Yincheng Wang[1], Ke Wang[1], Yuqi Huang[1], Di Wu[2], Jian Wu[3], Jibo He[4,1*]

[1]Psychology Department, School of Social Science, Tsinghua University, Beijing, China
[2]Department of Computer Science, Beijing Normal University, Beijing, China
[3]Haier Innovation Design Center, Haier Company, Qingdao, China
[4]Key Laboratory of Emotion and Mental Health in Chongqing, User Experience and Human-computer Interaction Technology Institute, Chongqing University of Arts and Sciences, Chongqing, China

Email Addresses of Co-authors:

Yincheng Wang	(wang-yc18@mails.tsinghua.edu.cn)
Ke Wang	(ke-wang17@mails.tsinghua.edu.cn)
Yuqi Huang	(huangyq17@mails.tsinghua.edu.cn)
Di Wu	(little_woody@163.com)
Jian Wu	(wuj@haier.com)
Jibo He	(hejibo666@mail.tsinghua.edu.cn)

Corresponding Author:

Jibo He	(hejibo666@mail.tsinghua.edu.cn)

KEYWORDS:

ergonomics, text input, keyboard design evaluation, smartphone, typing task

SUMMARY:

The presented protocol integrates various evaluation methods and demonstrates a method to evaluate the keyboard design on smartphones. Pairs matched by English characters are proposed as the input material, and the transition time between two keys is used as the dependent variable.

ABSTRACT:

Keyboard input has played an essential role in human-computer interaction with a vast user base, and the keyboard design has always been one of the fundamental objects of studies on smart devices. With the development of screen technology, more precise data and indicators could be collected by smartphones to in-depth evaluate the keyboard design. The enlargement of the phone screen has led to unsatisfactory input experience and finger pain, especially for one-handed input. The input efficiency and comfort have attracted the attention of researchers and designers, and the curved keyboard with size-adjustable buttons, which roughly accorded with the physiological structure of thumbs, was proposed to optimize the one-handed usage on large-screen smartphones. However, its real effects remained ambiguous. Therefore, this protocol demonstrated a general and summarized method to evaluate the effect of curved QWERTY keyboard design on a 5-inch smartphone through a self-developed software with detailed

variables, including objective behavioral data, subjective feedback, and the coordinate data of each touchpoint. There is sufficient existing literature on evaluating virtual keyboards; however, only a few of them systematically summarized and took reflection on the evaluation methods and processes. Therefore, this protocol fills in the gap and presents a process and method of the systematic evaluation of keyboard design with code for analysis and visualization that are available, that needs no additional or expensive equipment, and is easy to conduct and operate. In addition, the protocol also helps to get potential reasons for the disadvantages of the design and enlightens the optimization of designs. In conclusion, this protocol with the open-source resources could not only be an in-class demonstrative experiment to inspire the novice to start their studies but also contributes to improving the user experience and the revenue of input method editor companies.

INTRODUCTION:

Keyboard input is the mainstream method of the human-smartphone interaction[1,2], and with the penetration of smartphones, keyboard input gets billions of users. In 2019, the global smartphone penetration rate had reached 41.5%[3], while the United States, with the highest penetration, had come up to 79.1%[4]. Up to the first quarter of 2020, the Sogou mobile keyboard had about 480 million daily active users[5]. Up to May 6, 2020, the Google Gboard had been downloaded more than 1 billion times[6].

Unsatisfactory keyboard input experience increases with the enlargement of the phone screen. Although the enlarged screen aimed to improve the viewing experience, it has changed the gravity, size, and weight of smartphones, causing users to change holding posture repeatedly to reach remote areas (e.g., button A and Q for right-handed users), thus leading to input inefficiency. The stretch of muscle may cause users to suffer from musculoskeletal disorders, hand pains, and different types of disease (e.g., carpal tunnel syndrome, thumb osteoarthritis, and thumb tenosynovitis[7–10]). Users who prefer one-handed usage are under worse conditions[11,12].

Therefore, the evaluation and optimization of keyboard design have become hot topics of psychological, technical, and ergonomic research. Variable keyboard designs and concepts have constantly been proposed by input method editor (IME) companies and researchers to optimize input experience and efficiency, including layout-changed keyboards, character-reordered keyboards: Microsoft WordFlow Keyboard[13], Functional Button Area in Glory of Kings[14], IJQWERTY[15], and Quasi-QWERTY[16].

Existing evaluation methods of keyboard design vary from researcher to researcher except for several highly accepted indicators, and more accurate indicators are proposed. However, with a variety of indicators, there is not a summarized and systematic protocol provided to demonstrate the process of evaluating and analyzing the keyboard design. Fitts' Law[17] and its extended version FFitts' Law[18], which described human-computer interaction, were widely adopted to evaluate keyboard performance[19–22]. Moreover, the functional area of the thumb was proposed to improve keyboard design, and it described a curved motion area for the thumb to comfortably complete the input task[23]. Based on these theories, indicators including word per minute, word

error rate, and subjective feedback (perceived usability, perceived performance, perceived speed, subjective workload, perceived exertion and pain, and intent to use, etc.), which were highly adopted, were partially used in previous studies[24–29] except for modeling and simulation methods. In addition, the fitted ellipse of touchpoints on each button and its offset[30,31] were used in recent years to investigate the accurate performance of inputting events. Also, the galvanic skin response, heart rate, electromyographic activity, hand gesture, and body movement[32–35] were adopted to directly or indirectly evaluate muscle fatigue, comfort, and satisfaction of the users. However, these various methods lack reflection on the appropriateness of the indicators used, and a novice researcher may be confused to select the appropriate indicators for his or her research.

The research about keyboard design is also easy to be conducted, operated, and analyzed. With the boom of screen technology, more behavioral data could be easily collected to evaluate the keyboard design in-depth (e.g., the transition time between two keys and the coordinate data of each touchpoint). Based on the mentioned data, researchers could precisely explore the details of keyboard design and analyze its disadvantages and advantages. When compared with other human-computer interaction research, the research of keyboard design on portable smartphones also has high application value for its vast user base with no expensive equipment, complicated materials, or huge laboratory space needed. The questionnaires, scales, and Python script about the research are open-source and easy to access.

The purpose of this research is to summarize the previous methods to demonstrate a systematic, precise, and general protocol to evaluate and analyze the keyboard design on smartphones. The exemplar experiment and results aim to show whether the curved QWERTY keyboard with size-adjustable buttons could optimize the input experience of one-handed input on a 5-inch smartphone when compared with traditional QWERTY keyboard and share the visualization method and Python script of data analysis.

PROTOCOL:

The study was conducted in accordance with the ethical principle and was approved by the Ethics Committee of Tsinghua University. **Figure 1** shows the process of evaluating the keyboard design of smartphones.

[Place **Figure 1** here]

1. Preparation

1.1 Experiment design

1.1.1 Define the research issue and propose the hypothesis.

1.1.2 Design the experiment according to the hypothesis and define the independent variables (e.g., keyboard layout, typing posture). Use the within-subject design in order to reduce

confounding factors and variance caused by the difference among participants.

1.2 Dependent variables

1.2.1 Use physical data, including the hand length, the length of input finger, and the circumference of input finger, which were measured by a tape measure, as shown in **Figure 2**.

[Place **Figure 2** here]

1.2.2 Use physiological data, including galvanic skin response (measured by the portable wireless physiological detector), heart rate (measured by the portable wireless physiological detector), electromyographic activity (measured by surface electromyography), etc.

1.2.3 Use input performance: word per minute, word error rate, and transition time between two keys.

1.2.3.1 Word per minute refers to the input speed of participants (i.e., the number of correct-inputted words per minute).

1.2.3.2 Word error rate refers to the input accuracy of participants (i.e., the number of incorrect-inputted words divided by the total number of words under one condition). Corrected error rate, uncorrected error rate, and total error rate have also been used in previous studies[36].

1.2.3.3 Transition time between two keys refers to the reaction time of participants between two touchpoints of a correct-inputted word[22] (i.e., the start time of the second touchpoint minus the departure time of the first character).

1.2.4 Use body-movement data such as hand gesture and body (finger) movement. They could be collected by the motion capture system[35].

1.2.5 Use subjective data such as perceived usability, intent to use, perceived accuracy and speed, perceived exertion and pain, and subjective workload, etc. Subjective data can be obtained through existing scales and questionnaires, which are highly reliable as well as valid to better evaluate the subjective feedback of participants about the keyboard design.

1.2.5.1 Use NASA-TLX, a 21-point scale that is used to measure subjective workload through mental, physical, time, performance, effort, and frustration dimensions. A high score indicates a high subjective workload[26].

1.2.5.2 Use the System Usability Scale, a 5-point questionnaire with 10 items, and the responses of one participant will be calculated as a single score from 0 to 100. A high score indicates a high perceived usability[24].

1.2.5.3 Use the Borg CR10 Scale, which is ranged from 0 to 10 to measure perceived pain and

exertion. A high score indicates a high-level perceived pain and exertion[25].

1.2.5.4 Use the Intent to Use Scale: a 10-point questionnaire that is used to measure the likelihood that participants would use the technology or products. A high score indicates a high-level likelihood[28].

1.2.5.5 Perceived speed and perceived accuracy are all measured by 50-point scales, and a high score indicates a good perceived performance[28].

1.2.6 Collect the coordinate data of each touchpoint and change it into the fitted ellipse (95% CI) of touchpoints on each button[30,31]. Adopt the area of each fitted ellipse and the offset from the center of the fitted ellipse to the target center of each button as dependent variables.

NOTE: The coordinate data can be precisely collected by the self-developed application on the smartphone. If it is hard to obtain the coordinate data, objective and subjective data are sufficient to roughly evaluate the keyboard design.

1.3 Materials

1.3.1 Choose the experiment smartphone. Take weight, resolution, and screen size into consideration.

1.3.2 Design and develop the experiment software on smartphones (optional step).

NOTE: Transition time between two keys can be recorded automatically by this software or motion capture sensors (i.e., the accelerometer sensor). It may be difficult to collect it manually (e.g., a clock or stopwatch).

1.3.3 Select the input task from the following suggestions based on the hypothesis and revise it to match the research purpose.

1.3.3.1 For the character pair input task, randomly pair 26 English letters into 676 pairs and averagely divide them into several groups based on the experiment design.

1.3.3.2 For the phrase (sentence) input task, use phrases that are moderate in length, easy to remember, and representative of the target language. If the target language is English, extract 15–20 (or based on research purpose) phrases or words from a 500 phrases set[37].

1.4 Participant recruitment

1.4.1 Use the G*Power software to calculate the sample size.

1.4.2 Post questionnaires to recruit potential participants.

1.4.3 Filter potential participants with wanted characteristics, e.g., age, health, vision, handedness, and input experience. Ensure that the input experience of participants is balanced.

2. Procedure

2.1 Read out the informed consent form of the experiment to participants, including the experiment procedure, task, and whether they will encounter any mental or physical injuries. If participants agree to participate, they need to sign the informed consent form. If not, they can immediately withdraw. According to the informed consent form, participants can also withdraw at any stage of the experiment.

2.2 Collect physical as well as demographic data. Use a tape measure to measure the hand of every single participant (**Figure 2**) in order to eliminate the effect of the hand size difference and also provide repeatable data for future research. Collect demographic data such as age, gender, precise input experience, and occupation.

2.3 Disinfect all devices and clean the body parts of the participant that will touch the devices.

2.3.1 Ask participants to wash their hands and clean the screen of smartphones so that sensors of smartphones can be more sensitive.

2.3.2 Ask participants to wear portable wireless physiological detectors or a motion capture system. Ask participants to wear the portable wireless physiological detection wristband on the non-dominant hand to record galvanic skin response and heart rate with the noise interference avoided.

2.3.2.1. Place passive markers of the motion capture system on the fingernails, the proximal phalanx of the finger, cervical vertebrae (C3–C5), and arm, to collect the precise body and finger movement. Stick wireless electrodes to the skin of two arms and two forearms to detect the electromyographic activity (optional step).

2.3.3 Calibrate all the devices used in the experiment.

2.4 Practice part

2.4.1 Let participants complete the training task. The training task is used to improve participants' familiarity with input tasks and keyboards to reduce the effect of practice or the unfamiliarity on the experiment result. It is composed of 50 pairs or 20 words randomly selected from the 676 English pairs set or 500 phrases set. Only when their input accuracy reaches 80% or more in 150 seconds can they enter the formal trials. The exemplar research adopted inputting 50 pairs as the training task.

2.5 Main task

2.5.1 Let participants complete formal trials under all experimental conditions. They need to ensure their accuracy as quickly as possible during the time of the input task. Formal trials are real input tasks that will be evaluated and analyzed in the research. Each pair, word, or sentence represents a trial, and different experimental designs produce different experimental conditions.

2.5.2 Have participants complete the input task in random order or a balanced order. Methods of the division of input materials are as follows. First, 676 pairs can be randomly divided into each experimental condition (i.e., participants have entered all pairs when they complete all experimental conditions). Second, under each experimental condition, 676 pairs can be divided into several blocks randomly, and participants need to complete these blocks randomly. Third, for inputting words, participants need to complete around 20 trials under each condition. Fourth, for inputting sentences, participants need to complete about 10–15 trials under each condition. Researchers should ensure no significant difference between the number of characters and the number of words entered by the participant under each condition. The exemplar research adopted the first method and had four experimental conditions.

2.5.3 After each condition, ask participants to complete all the questionnaires (scales assessing their subjective experience) at random and give them 1 min to rest.

2.6 At the end of the experiment, let each participant finish the comprehensive questionnaire (Q & A) to obtain subjective feedback.

2.7 Express appreciation to participants with monetary or material rewards.

3. Data analysis

3.1 Hypothesis testing by appropriate parametric or non-parametric tests

3.1.1 Analyze the physical, physiological, body-movement data to test whether the difference between participants would significantly influence the results and inexpressive input experience of users (optional step).

3.1.2 Analyze the input performance of participants to test the input efficiency on the keyboard.

3.1.3 Analyze subjective data to test the perceived usability and subjective feedback of the keyboard.

3.1.4 Figure out whether the practice effect and fatigue effect significantly influence the result. For each condition, trials are divided into two parts according to the timestamp (i.e., the first half part and the second half part). Specifically, under each condition, examine the difference of input performance between the first half part and the second half part to test whether the practice effect or fatigue effect exist.

3.1.5 Analyze the area of the fitted ellipse of touchpoints on each button as well as the offset

from its center to the target center of each button (optional step).

3.1.5.1 Collect all the touchpoints of each button with the software, and they roughly accord with the bivariate Gaussian distribution. The 95% confidence interval of each button in both x- and y-directions is derived through the coordinate data of each touchpoint in pixel, and the 95% confidence ellipses over a 1:1 outline of the button for each keyboard is fitted through Python scripts on pixel coordinate (see **Coding File 2**).

3.1.5.2 Use fitted ellipses (95% CI) and their areas to demonstrate the dispersion of touchpoints on each button. In each button, the offset of fitted ellipse calculated by Python scripts is defined as the center point of the fitted ellipse to the target point of the button, and it could be represented from x- and y-directions (i.e., in X-axis and Y-axis, see **Coding File 3**).

3.2 **Modeling and simulation**

3.2.1 Use the data-driven model as a function of keyboard location and orientation to predict the finger movement by Python scripts. All movements of fingers are divided into eight directions[38] (the top to the bottom, the bottom to the top, the left to the right, the right to the left, the left-top to the right-bottom, the right-bottom to the left-top, the left-bottom to the right-top, the right-top to the left-bottom). For each direction, the average transition time between two keys is calculated to represent the effectiveness of finger movement, which is used to evaluate the keyboard design (optional step).

3.2.2 Use linear regression analysis to build an enhanced Fitts' Law (or its extended version, FFitts' Law) model to predict the transition time between two keys using an integrated cognitive architecture[39] by Python scripts. The enhanced Fitts' Law model could provide a better prediction and evaluation on keyboard design based on its analyses on the location and effective width of keys, as well as the distance of two keys (optional step).

REPRESENTATIVE RESULTS:

The representative study is mainly following the mentioned protocol. The study adopts a 2 (Keyboard layout: Curved QWERTY vs. Traditional QWERTY) × 2 (Button size: large, 6.3 mm × 9 mm vs. small, 4.9 mm × 7 mm) within-subject design to evaluate whether the curved QWERTY could improve the input efficiency and comfort when compared with the traditional QWERTY in different sizes of buttons by the character pair input task through our self-developed software (**Figure 3**). This study has not adopted the expensive physiological detector equipment or motion capture system, and the data analysis did not contain the modeling or simulation.

[Place **Figure 3** here]

A total of 24 right-handed healthy students from Tsinghua University were involved in this study (12 females, M = 22.46 years, SD = 3.04 years). For them, the length of right hand (M = 17.98 cm SD = 1.20 cm), the length of right thumb (M = 6.00 cm, SD = 0.68 cm), and the circumference of right thumb (M = 5.14 cm, SD = 0.52 cm) were measured. The sample size was calculated by

G*Power 3.1.9.2 (effect size f = 0.25, α = 0.05, power = 0.80, correlation among repeated measures = 0.5). The experiment smartphone is a 5-inch smartphone (weight 138 g, screen size 5.0 inch, ppi 294, px 1280 × 720, phone size 143.5 × 69.9 × 7.6 mm).

Input performance (transition time between two keys, word error rate), subjective feedback, and fitted ellipse of each button were collected and analyzed by repeated measures ANOVA. Transition time between two keys instead of word per minute is used in this study because the input material is the character pairs, and the transition time between two keys could evaluate the transition touch event more precisely. The representative results are as follows (**Table 1**).

[Place **Table 1** here]

In the input performance, the interaction between keyboard layout and button size is only significant in the transition time between two keys (**Figure 4**), and it shows that in the curved QWERTY, the transition time between two keys of small button size was significantly longer than that of large button size ($p < 0.001$). The main effect of keyboard layout is significant in both word error rate (**Figure 5**) and transition time between two keys, and it indicates that these of the traditional QWERTY are significantly lower than those of the curved QWERTY. The main effect of button size is significant in both word error rate and transition time between two keys, and it indicates that these of the large button size are significantly lower than those of the small button size. No other significant result is found.

[Place **Figure 4** here]

[Place **Figure 5** here]

In the subjective feedback (**Figures 6–7**), all the interactions between the keyboard layout and button size are not significant. The main effect of keyboard layout is significant in intent to use and subjective workload (mental, performance, effort, and frustration), and it shows that participants perceive less subjective workload (the above four facets) and have more likelihood to use the curved QWERTY when compared with the traditional QWERTY. The main effect of button size is significant in perceived usability and all facets of subjective workload, and it indicates that participants perceive less subjective workload and higher usability in the large button size when compared with small button size. No other significant result is found.

[Place **Figure 6** here]

[Place **Figure 7** here]

In the area of the fitted ellipse (**Figure 8**), the interaction between keyboard layout and button size is significant, and it shows that for both small and large button size, the area of the traditional QWERTY is larger than that of the curved QWERTY ($p < 0.001$), while for both keyboard layouts, the area of the small button is smaller than that of the large button ($p < 0.001$). The main effect of button size and keyboard layout is significant, and it indicates that those areas of the

traditional QWERTY and the large button are larger than those of the curved QWERTY and the small button, respectively. No other significant result is found.

[Place **Figure 8** here]

In the offset of the fitted ellipse (**Figures 9–10**), the interaction between keyboard layout and button size is only significant in the offset in the y-direction, and it shows that in the curved QWERTY, the offset in the y-direction of the small button is significantly shorter than that of the big button ($p < 0.001$), while in both sizes of the button, the offset in the y-direction of the curved QWERTY is significantly shorter than that of the traditional QWERTY. The main effect of keyboard layout is significant in both x- and y-directions, and it indicates that the offset in the y-direction of the curved QWERTY is significantly shorter than that of the traditional QWERTY. No other significant result is found.

[Place **Figure 9** here]

[Place **Figure 10** here]

The practice effect is tested using the *t*-test to compare the input performance (word error rate and transition time between two keys) between the first half and the second half of the character pairs. As for error rate, there is no significant difference between the two groups of character pairs in the curved QWERTY with small button size, $t_{(46)} = 2.03$, p = 0.05, the curved QWERTY with big button size, $t_{(46)} = -0.47$, p = 0.64, the traditional QWERTY with big button size, $t_{(46)} = 0.31$, p = 0.76, and the traditional QWERTY with small button size, $t_{(46)} = 0.05$, $p = 0.97$. As for transition time between two keys, there is no significant difference between the two groups of character pairs in the curved QWERTY with big button size, $t_{(46)} = 0.33$, $p = 0.74$, the curved QWERTY with small button size, $t_{(46)} = 0.22$, $p = 0.83$, the traditional QWERTY with big button size $t_{(46)} = 0.66$, $p = 0.51$, and the traditional QWERTY with small button size, $t_{(46)} = 0.09$, $p = 0.93$. The results indicate that there is no practice effect or fatigue effect during the main process of the input task, and participants have reached and kept the highest effort for each keyboard. The absolute value of the highest effort for different keyboards may be different because the highest effort only indicates that they have been familiar with the keyboard by 100 percent.

This representative study indicates that on the 5-inch smartphone, the curved QWERTY is worse than the traditional QWERTY, and the big button size is better than the small button size. In this representative study, the best keyboard is the traditional QWERTY keyboard with large button size, while the worst keyboard is the curved QWERTY keyboard with small button size. All the results have not been affected by the practice effect and fatigue effect. The word error rate and the transition time between two keys indicate that the curved QWERTY design increases the reaction time of participants between two characters and may augment the recognition workload to characters because of the position of keys and mental rotation, thus leading to unsatisfactory input performance, and the results are the same as the size-reduced button size (QWERTY keyboard with small button size) on a 5-inch smartphone. Although most indicators and dimensions of the subjective feedback are not significant, the subjective workload shows the

higher perceived workload of the QWERTY keyboard with the size-reduced button and the curved QWERTY keyboard. However, from the analysis of fitted ellipses, the results, and **Figures 8–10** show that the curved QWERTY has less offset and its touchpoints are less dispersive, and its offset is mainly toward the upper-left corner for right-handed usage. The results indicate that the curved QWERTY design could be optimized by adjusting the curvature of the keyboard, adding the function of automatic correction, and moderating the size of the buttons. In addition, from the **Figures 8–10**, a curved T9 keyboard, which takes the place of "R, T, Y, U, I, O, D, F, G, H, J, K, X, C, V, B, N, and M" of the curved QWERTY keyboard, may be a potential optimized keyboard, i.e., each key of the curved T9 keyboard takes the place of two letter keys of the curved QWERTY.

Therefore, this representative study only roughly demonstrates the protocol of the evaluation of keyboard design with open-source Python scripts, and the analysis and optimization method could be discussed in-depth based on the research purpose of researchers in the future studies.

FIGURES AND TABLES:

Figure 1: General process of conducting a keyboard experiment and evaluating the keyboard design.

Figure 2: The measurement of the hand.

Figure 3: The interface of the traditional QWERTY keyboard and the curved QWERTY keyboard software. (**A**) Traditional QWERTY keyboard with large button size (letter key size: 6.3 mm × 9 mm). (**B**) Curved QWERTY keyboard with large button size (letter key size: 6.3 mm × 9 mm). (**C**) Traditional QWERTY keyboard with small button size (letter key size: 4.9 mm × 7 mm). (D) Curved QWERTY keyboard with small button size (letter key size: 4.9 mm × 7 mm). The aspect ratio of each letter key is 7:10, and the width of each functional key (Delete, Space, Enter) is twice as that of the letter key. Delete and Space are unworked. Participants click the Enter key to shift to the next trial.

Table1: Statistical analysis of the input performance, subjective feedback, and fitted ellipse of each button. Item with * means $p < 0.05$, item with ** means $p < 0.01$, and item with *** means $p < 0.001$.

Figure 4: The 3D bar graph is the visualization of transition time between two keys (the left is the first character while the right is the second character) in four keyboards. The height of each bar represents the value of transition time. The gradient colors (blue, green, yellow, and red) are used to show the situation of numerical distribution (see **Supplementary Coding File 1**).

Figure 5: The word error rate of each keyboard. The error bars represent 95% CI.

Figure 6: The perceived exertion and pain, intent to use (left Y-axis), perceived accuracy, perceived, and perceived usability (right Y-axis) of each keyboard. The high score of perceived

exertion and pain indicates the unsatisfactory experience, while the other indicators show the opposite. The error bars represent 95% CI.

Figure 7: The six dimensions of subjective workload. The error bars represent 95% CI.

Figure 8: The fitted ellipses (95% CI) of four keyboards. They are drawn by fitting the pixel positions of the touchpoints in four keyboards. The coordinate of the center of the ellipse is the average value of all touchpoints on each button (see **Supplementary Coding File 2**).

Figure 9: The offset of fitted ellipses in the x-direction. The length of the arrow, which is enlarged 1.2 times in proportion in the figure because of the visualization, represents the value of the offset. And different colors visualize the value of standard deviation (±) from the average offset of each button to the offset in the x-direction. The value less than -1σ is green, and the value more than +1σ is red, while the value between -1σ and +1σ is orange (see **Supplementary Coding File 3**).

Figure 10: The offset of fitted ellipses in the y-direction. The length of the arrow, which is enlarged 1.2 times in proportion in the figure because of the visualization, represents the value of the offset. And different colors visualize the value of standard deviation (±) from the average offset of each button to the offset in the y-direction. The value less than -1σ is green, and the value more than +1σ is red, while the value between -1σ and +1σ is orange (see **Coding File 3**, and the script of the y-direction is familiar to that of the x-direction).

Table 1: Summarized data analysis of curved QWERTY keyboard evaluation.

Supplementary Coding File 1: 3D plots of the transition time between two keys.

Supplementary Coding File 2: The fitted ellipse and its area.

Supplementary Coding File 3: The offset of the fitted ellipse.

DISCUSSION:

In this study, based on the development of screen technology, we presented a summarized and general protocol of keyboard design evaluation to assess the keyboard design systematically and precisely. Existing indicators and methods from previous studies, pairs matched by English characters, and transition time between two keys are integrated and modified to generate an effective protocol.

Several critical points need to be noticed in this protocol. The selection of variables and indicators is essential because they decide the perspective of analysis, and it could be used to build the evaluation model in the later stage of the keyboard design evaluation experiment. Except for the objective variables, the subjective variables should also be carefully considered in the experimental design from multiple dimensions, since the subjective data plays a vital role in helping us improve user experience. Coordinate data can be optionally collected and calculated

in the protocol through the self-developed application and Python scripts, e.g., fitted ellipse (95% CI) of touchpoints on each button and the offset from the center of the fitted ellipse to the target center of each button. The analysis and visualization of the fitted ellipse may enlighten the optimization method of the keyboard design. In addition, although physiological measurement and movement measurement, which depend on the wearable equipment, are also optional, they could indeed help to explore the inexpressible experience of keyboard users in-depth.

One crucial step in the procedure of keyboard study is asking participants to wash their hands and clearing the screen before the experiment (the same as the wearable detectors), since hand grease and sweat may affect the sensitivity of the screen sensory, thus influencing the results. The physical data (hand length, thumb circumference) of the participants also needs to be measured or reported because the physical differences between participants may affect the experiment results and the reproducibility as well.

The protocol also cannot escape from the following limitations. All the input materials proposed in this study may mainly concentrate on the language of English without the consideration of other languages. In addition, self-developing a keyboard software to collect the experiment data may be suggested in this protocol, instead of using the traditional manual collection and measurement method. Because a self-developed software could collect and calculate more precise and attributional indicators and help to provide a clear optimization suggestion about the keyboard design rather than only to conclude the effect of the current keyboard design under experimental conditions. Besides, other expensive devices or equipment adopted by previous studies have not been included in the representative results, such as the portable wireless physiological detector or motion capture system, and researchers should choose their specific experimental devices based on their research problem and hypothesis. Finally, followers of the New Statistics or Bayesian enthusiasts could try to adopt more statistical methods to analyze and evaluate the keyboard design.

For future applications and directions, this protocol can be adopted in the keyboard design evaluation process on other smart devices. In addition to smartphones, more and more intelligent devices have gained popularity, for instance, wearable smartwatches and bracelets (iWatch), tablet PC (iPad), and virtual reality devices (VR glasses). This protocol can be used to evaluate various keyboard designs on these devices and helps optimizations (indicators and processes may be slightly adjusted). In this sense, this study opens up new opportunities to re-examine the benefits and importance of keyboard design evaluation study in the touch screen of smart devices. Therefore, it provides an inexpensive and easy-to-conduct research method with the open-source resources in the field of human-computer interaction, computer science, and psychology, thus making contributions to helping the novice researchers and students to start their studies or being an in-class demonstrative experiment.

ACKNOWLEDGMENTS:
This research is supported by the Tsinghua University Initiative Scientific Research Program (Ergonomic design of curved keyboard on smart devices). The authors appreciate Tianyu Liu for his kind suggestions and coding assistance on figures.

DISCLOSURES:

The authors declared no financial disclosure or conflicts of interest.

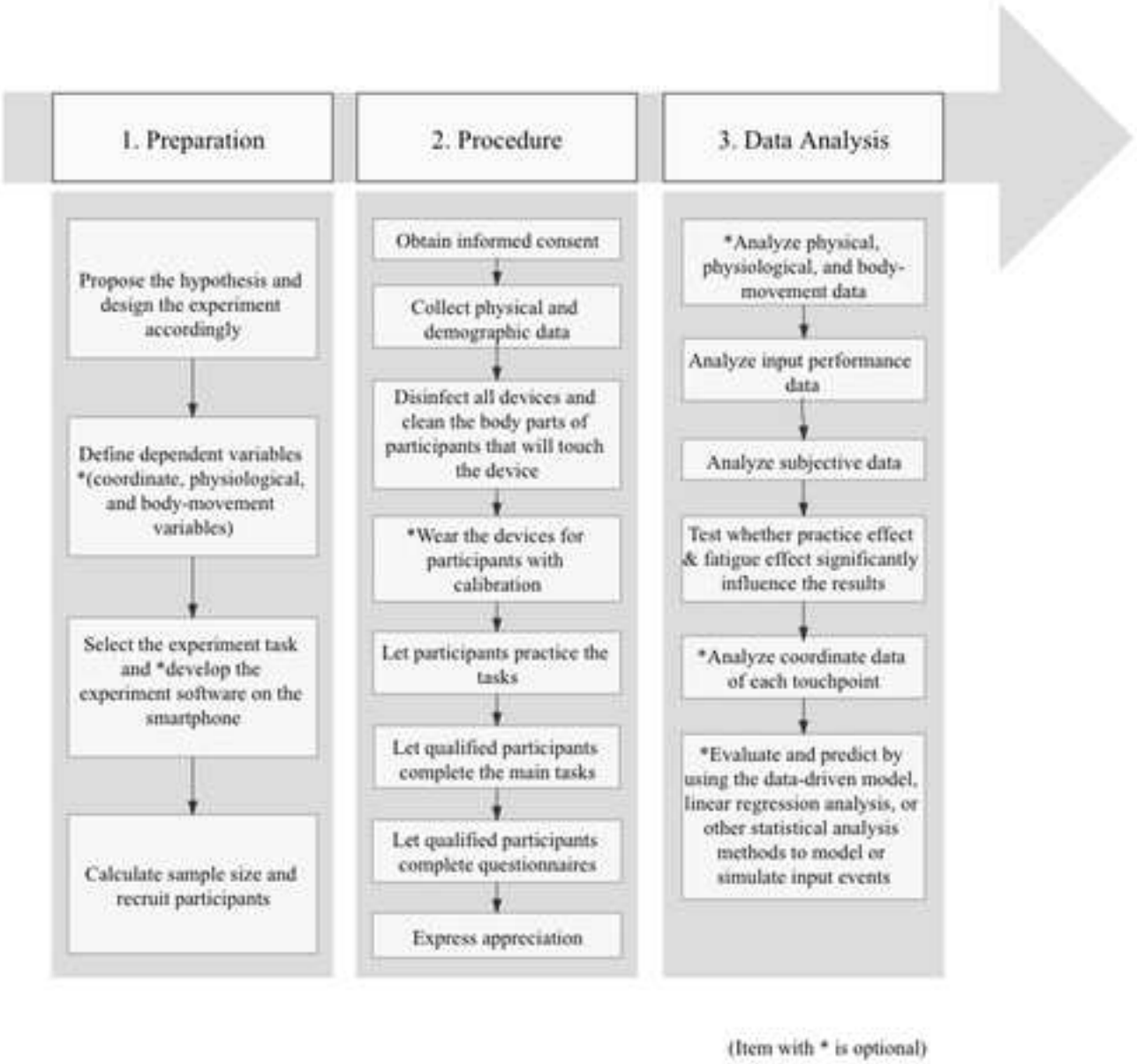
REFERENCES:

1. Lee, S., Zhai, S. The performance of touch screen soft buttons. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. USA (2009).
2. Smith, B. A., Bi, X., Zhai, S. Optimizing touchscreen keyboards for gesture typing. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. USA (2015).
3. Statista. *Global smartphone penetration rate as share of population from 2016 to 2020* [Fact sheet]. https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/ (2020).
4. Newzoo. *Top Countries by Smartphone Users* [Fact sheet]. https://newzoo.com/insights/rankings/top-countries-by-smartphone-penetration-and-users/ (2019).
5. Sogou. *Sogou Announces Fourth Quarter and Full Year 2019 Results* [Press release]. http://ir.sogou.com/2020-03-09-Sogou-Announces-Fourth-Quarter- and-Full-Y ear-2019-Results (2020).
6. Google Play. *Gboard - the Google Keyboard* [Press release]. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin &hl=en (2020).
7. Eitivipart, A. C., Viriyarojanakul, S., Redhead, L. Musculoskeletal disorder and pain associated with smartphone use: A systematic review of biomechanical evidence. *Hong Kong Physiotherapy Journal*. **38** (2), 77–90 (2018).
8. Chang, J., Choi, B., Tjolleng, A., Jung, K. Effects of button position on a soft keyboard: Muscle activity, touch time, and discomfort in two-thumb text entry. *Applied Ergonomics*. **60**, 282–292 (2017).
9. Gehrmann, S. V. et al. Motion deficit of the thumb in CMC joint arthritis. *Journal of Hand Surgery*. **35** (9), 1449–1453 (2010).
10. Kim, G., Ahn, C. S., Jeon, H. W., Lee, C. R. Effects of the Use of Smartphones on Pain and Muscle Fatigue in the Upper Extremity. *Journal of Physical Therapy Science*. **24** (12), 1255–1258 (2012).
11. Girouard, A. et al. One-handed bend interactions with deformable smartphones. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. USA (2015).
12. Lee, M., Hong, Y., Lee, S., Won, J., Yang, J., Park, S. The effects of smartphone use on upper extremity muscle activity and pain threshold. *Journal of Physical Therapy Science*. **27** (6), 1743–1745 (2015).
13. Microsoft Garage. *Word Flow keyboard* [Press release]. https://www.microsoft.com/en-us/garage/profiles/word-flow-keyboard/ (2020).
14. Tencent Games. *The glory of kings* [Press release]. https://pvp.qq.com/ (2020).

15. Bi, X., Zhai, S. Ijqwerty: what difference does one key change make? Gesture typing keyboard optimization bounded by one key position change from qwerty. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. USA (2016).
16. Bi, X., Smith, B.A., Zhai, S. Quasi-qwerty soft keyboard optimization. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. USA (2010).
17. Fitts, P. The information capacity of the human motor system is controlled by the amplitude of movement. *Journal of Experimental Psychology*. **47**, 381–391 (1954).
18. Bi, X., Li, Y., Zhai, S. FFitts law: modeling finger touch with fitts' law. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. USA (2013).
19. Dunlop, M., Levine, J. Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. USA (2012).
20. Li, Y., Chen, L., Goonetilleke, R. S. A heuristic-based approach to optimize keyboard design for single-finger keying applications. *International Journal of Industrial Ergonomics*. **36** (8), 695–704 (2006).
21. Benligiray, B., Topal, C., Akinlar, C. SliceType: fast gaze typing with a merging keyboard. *Journal on Multimodal User Interfaces*. **13** (4), 321–334 (2019).
22. Wang, Y., Ai, H., Liang, Q., Chang, W., He, J. *How to optimize the input efficiency of keyboard buttons in large smartphone? A comparison of curved keyboard and keyboard area size* [Conference presentation]. International Conference on Human-Computer Interaction. Berlin, Germany (2019)
23. Bergstrom-Lehtovirta, J., Oulasvirta, A. Modeling the functional area of the thumb on mobile touchscreen surfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Canada (2014).
24. Brooke, J. SUS: A retrospective. *Journal of Usability Studies*. **8** (2), 29–40 (2013).
25. Borg, G. Principles in scaling pain and the Borg CR Scales. *Psychologica*. **37**, 35–47 (2004).
26. Hart, S. G., Staveland, L. E. Development of NASA-TLX (task load index): results of empirical and theoretical research. In *Human mental workload.* Edited by Hancock, P. A., Meshkati, N. 139–183, Oxford (1988).
27. Trudeau, M. B., Asakawa, D. S., Jindrich, D. L., Dennerlein, J. T. Two-handed grip on a mobile phone affords greater thumb motor performance, decreased variability, and a more extended thumb posture than a one-handed grip. *Applied Ergonomics*. **52**, 24–28 (2016).
28. Turner, C. J., Chaparro, B. S., He, J. Text input on a smartwatch qwerty keyboard: tap vs. trace. *International Journal of Human Computer Interaction*. **33** (1–3), 143–150 (2017).
29. Zhai, S., Kristensson, P. O. The word-gesture keyboard: reimagining keyboard interaction. *Communications of the ACM*. **55** (9), 91–101 (2012).
30. Azenkot, S., Zhai, S. Touch behavior with different postures on soft smartphone keyboards. *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. New York, USA (2012).
31. Yi, X., Yu, C., Shi, W., Shi, Y. Is it too small?: Investigating the performances and preferences of users when typing on tiny qwerty keyboards. *International Journal of Human Computer Studies*. **106**, 44–62 (2017).
32. Li, Y., You, F., Ji, M., You, X. Smartphone text input: effects of experience and phrase complexity on user performance, physiological reaction, and perceived usability. *Applied*

Ergonomics. **80**, 200–208 (2019).
33. Gerard, M. J., Jones, S. K., Smith, L. A., Thomas, R. E., Wang, T. An ergonomic evaluation of the Kinesis ergonomic computer keyboard. *Ergonomics*. **37**(10), 1661–1668 (1994).
34. Van Galen, G. P., Liesker, H., Haan, A. Effects of a vertical keyboard design on typing performance, user comfort and muscle tension. *Applied Ergonomics*. **38**(1), 99–107 (2007).
35. Baker, N. A., Cham, R., Cidboy, E. H., Cook, J., Redfern, M. S. Kinematics of the fingers and hands during computer keyboard use. *Clinical Biomechanics*. **22**(1), 34–43 (2007).
36. Soukoref, R. W., MacKenzie, I. S. Metrics for text input research: an evaluation of MSD and KSPC, and a new unified error metric. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 113–120 (2003).
37. Mackenzie, I. S., Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*. 754–755 (2003).
38. Trudeau, M. B., Sunderland, E. M., Jindrich, D. L., Dennerlein, J. T., Federici, S. A data-driven design evaluation tool for handheld device soft keyboards. *Plos One*. **9** (9), e107070 (2014).
39. Cao, S., Ho, A., He, J. Modeling and predicting mobile phone touchscreen transcription typing using an integrated cognitive architecture. *International Journal of Human-Computer Interaction*. **34** (4–6), 544–556 (2018).

Figure 1



(Item with * is optional)

Figure 2

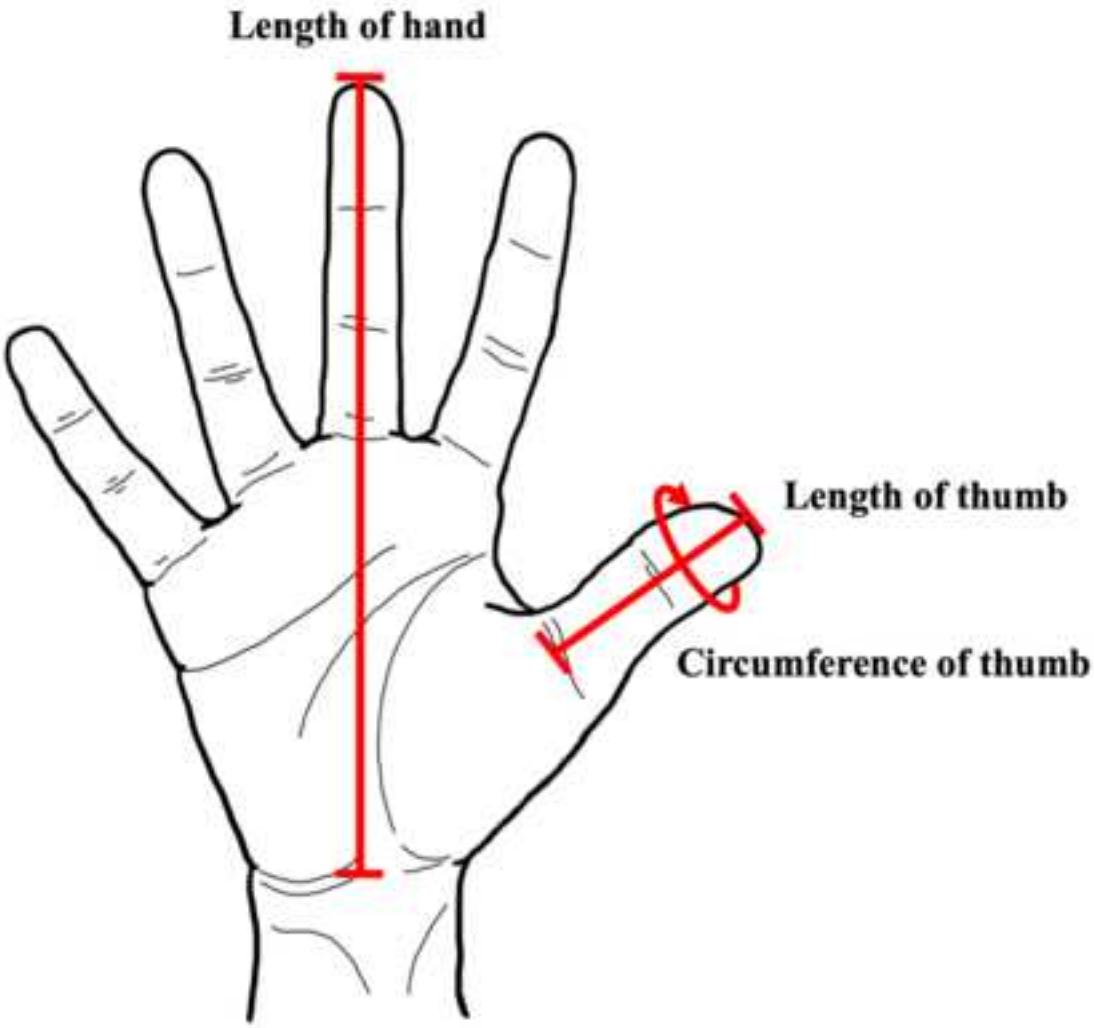


Figure 3

Click here to access/download;Figure;Figure3.psd

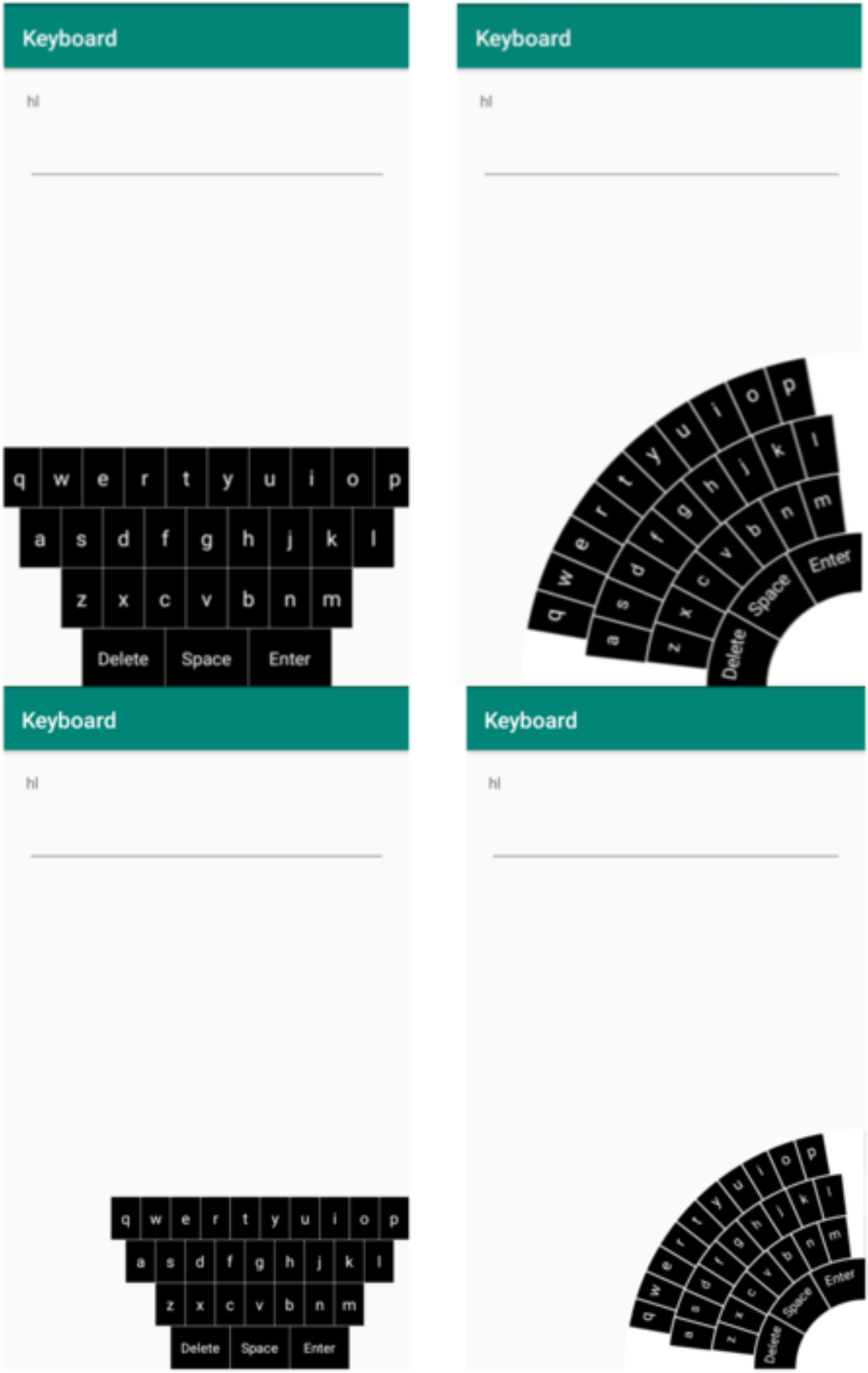


Figure 4

Click here to access/download;Figure;Figure4.psd

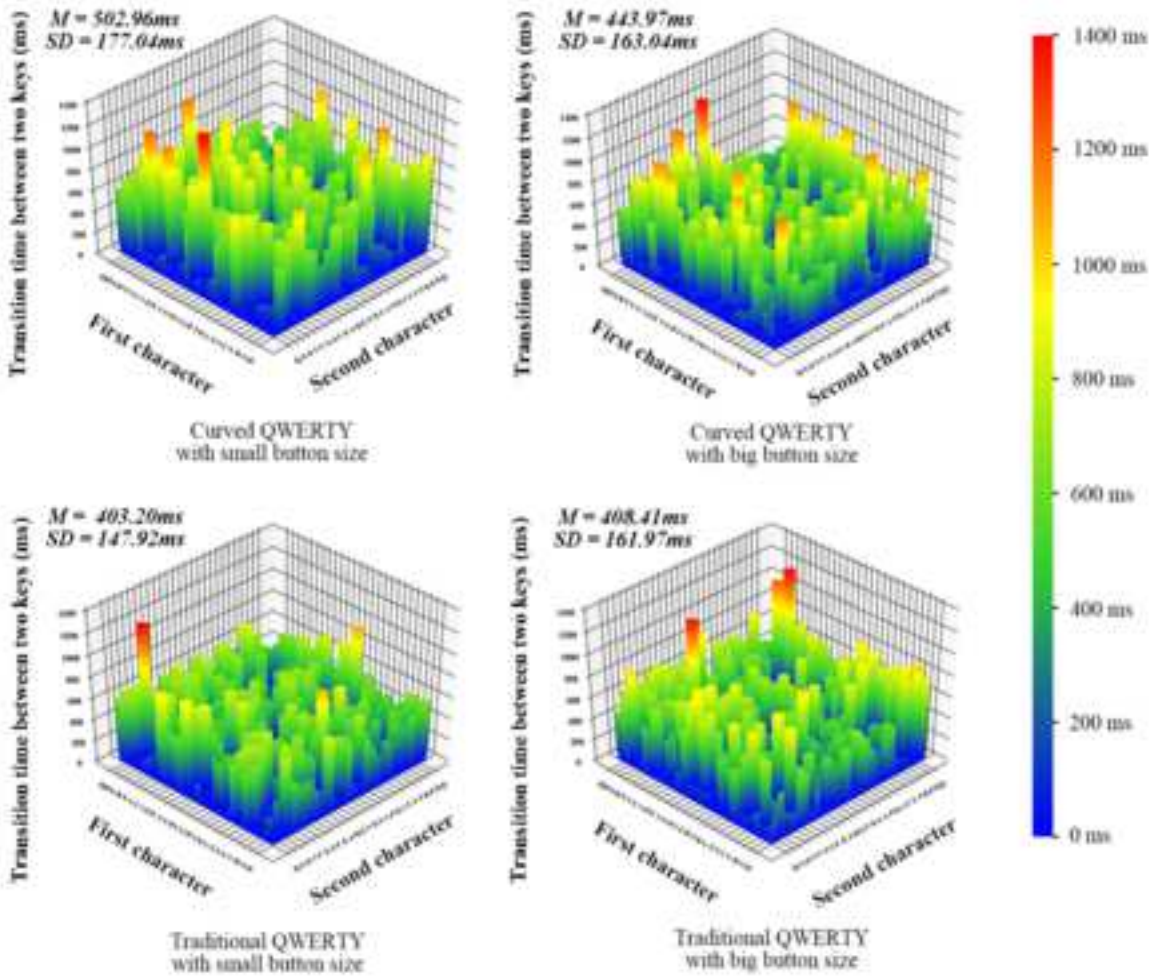


Figure 5

Click here to access/download;Figure;Figure5.psd

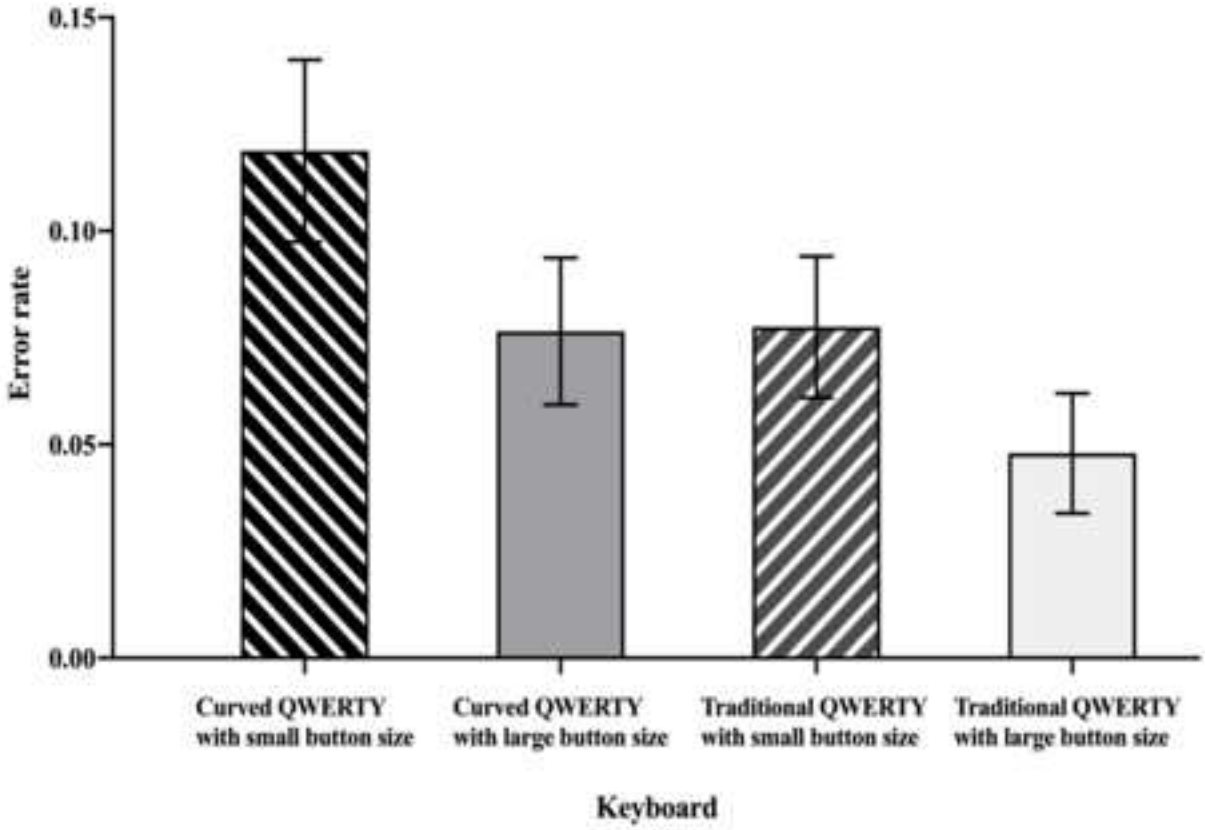


Figure 6

Click here to access/download;Figure;Figure6.psd

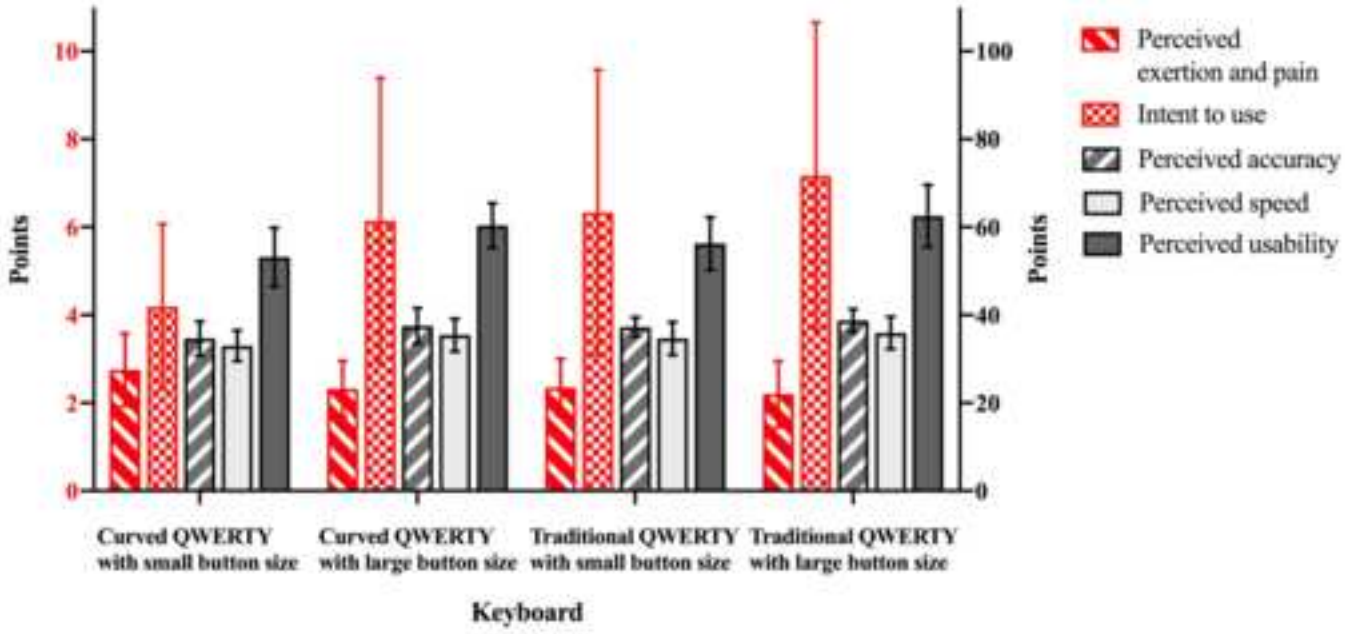


Figure 7

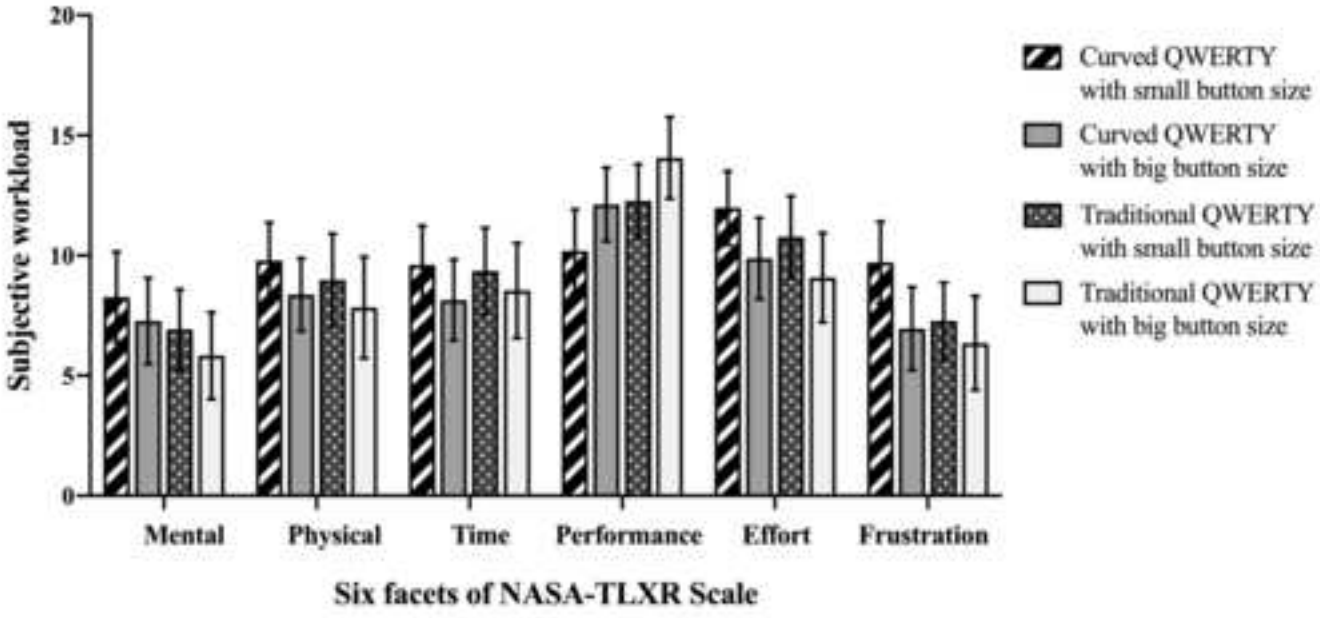


Figure 8

Click here to access/download;Figure;Figure8.psd

Traditional QWERTY keyboard
with large button size

Traditional QWERTY keyboard
with small button size

Curved QWERTY keyboard
with large button size

Curved QWERTY keyboard
with small button size

Figure 9

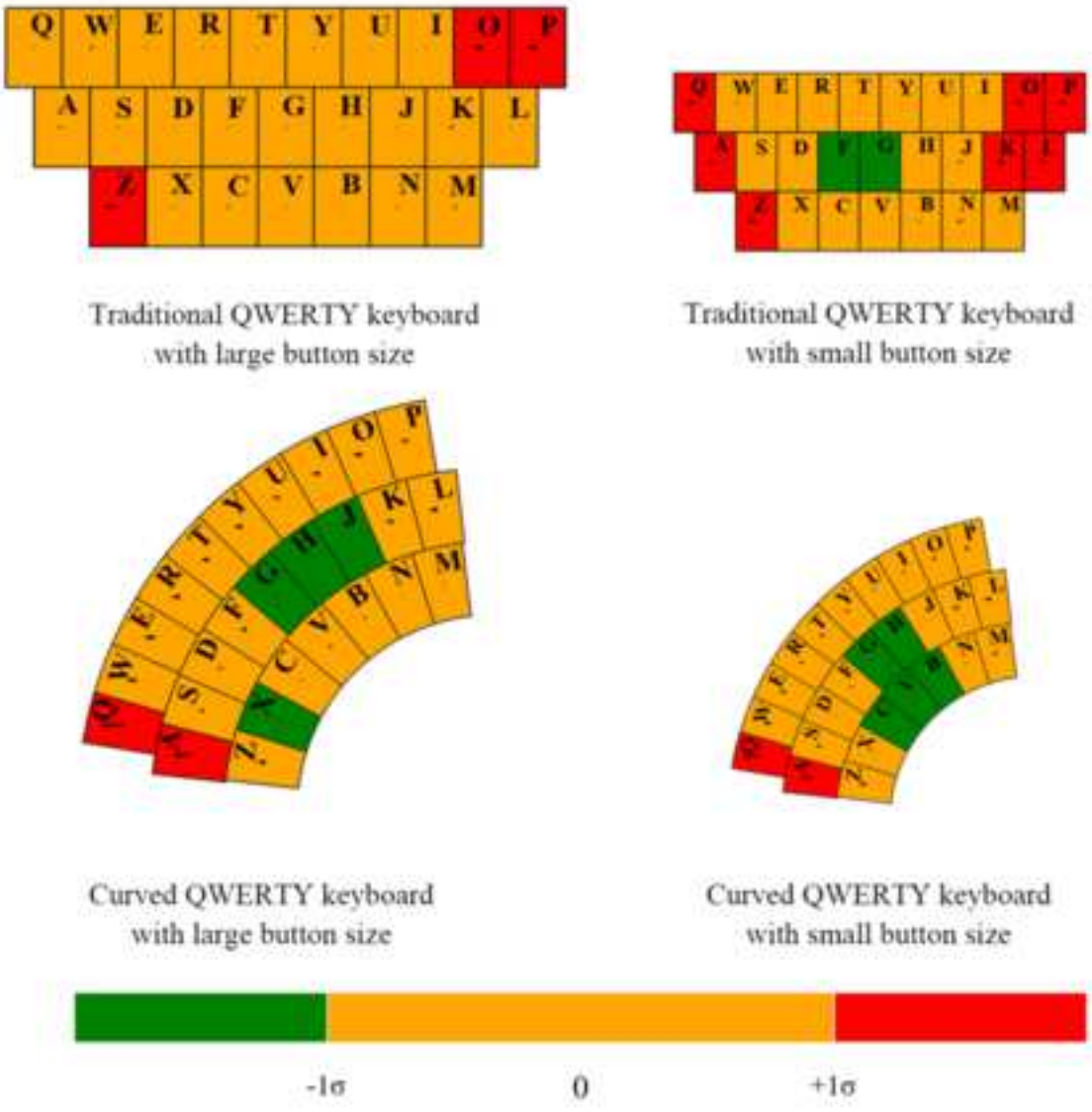


Figure 10

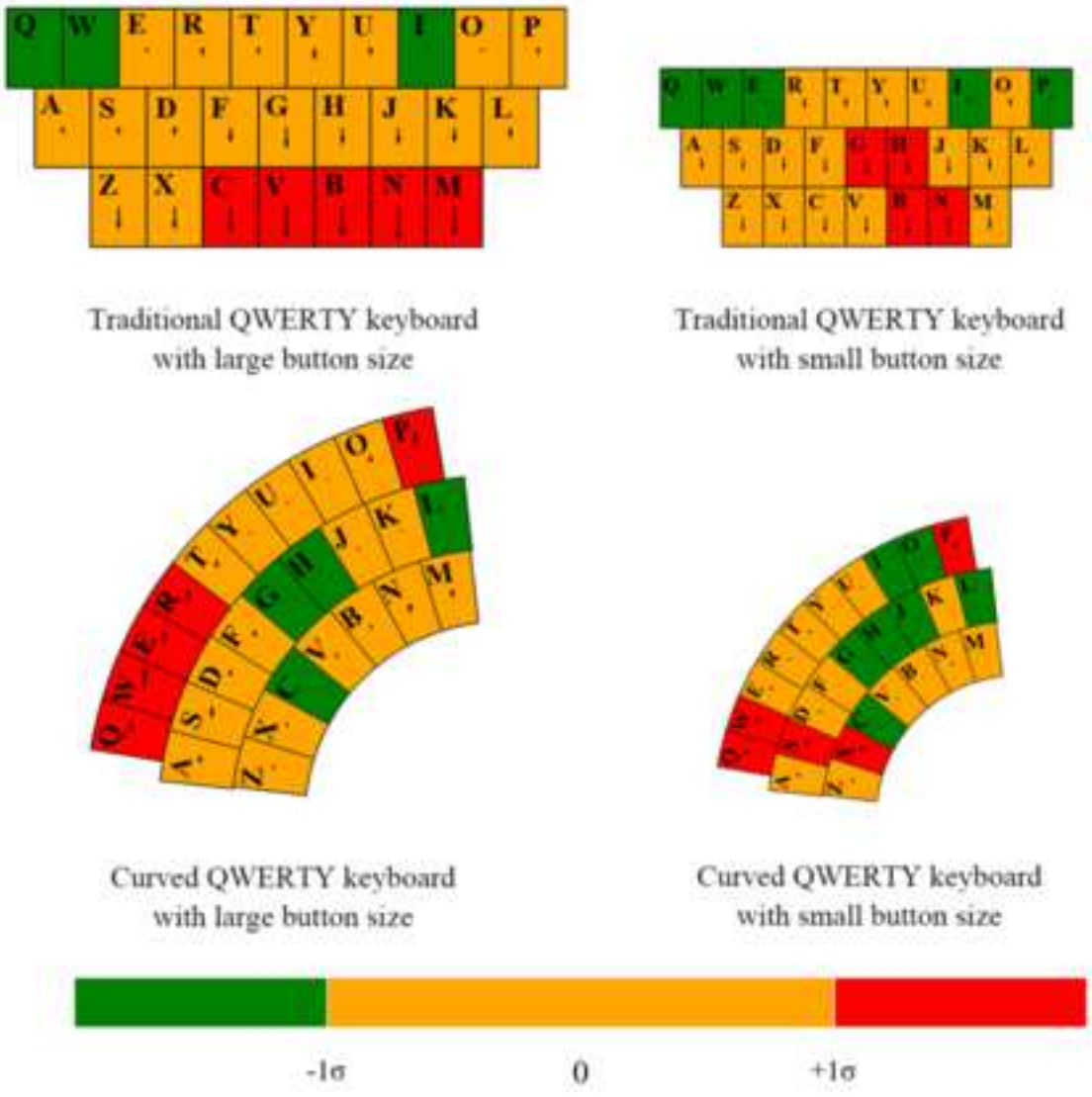


Table 1

		Keyboard layout			Button size			Keyboard layout × Button size		
		F	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Word error rate		48.90	<.001***	0.68	30.57	<.001***	0.57	2.63	0.12	0.10
Transition time between two keys		10.19	.004**	0.31	43.57	<.001***	0.66	12.75	.002**	0.36
Perceived exertion and pain		2.33	0.14	0.09	1.36	0.26	0.06	0.28	0.60	0.01
Intent to use		7.41	.012*	0.24	3.62	0.07	0.14	0.63	0.44	0.03
Perceived accuracy		1.32	0.26	0.54	2.94	0.10	0.11	0.69	0.42	0.03
Perceived speed		0.56	0.47	0.02	0.98	0.33	0.04	0.25	0.62	0.01
Perceived usability		0.63	0.44	0.03	5.48	.028*	0.19	0.03	0.87	0.001
Subjective workload	Mental	19.30	<.001***	0.46	8.88	.007**	0.28	0.01	0.91	0.001
	Physical	2.41	0.13	0.10	5.55	.027*	0.19	0.07	0.78	0.003
	Time	0.02	0.9	0.001	10.26	.004**	0.31	0.37	0.55	0.02
	Performance	11.51	.003**	0.33	12.25	.002**	0.35	0.02	0.90	0.001
	Effort	4.66	.042*	0.17	16.33	.001**	0.42	0.13	0.72	0.006
	Frustration	9.32	.006**	0.29	8.87	.007**	0.28	2.11	0.16	0.08
Area of fitted ellipse		90.00	<.001***	0.78	1368.78	<.001***	0.98	31.99	<.001***	0.56
Offset of fitted ellipse	X-direction	10.94	.003**	0.30	1.4	0.25	0.05	6.08	0.21	0.19
	Y-direction	23.49	<.001***	0.48	0.48	0.50	0.02	13.74	.001**	0.36

Table of Materials

Name of Material/Equipment	Company	Catalog number	Comments
Changxiang 6S smartphone	Huawei		Smartphone used in the examplar study
Curved QWERTY keyboard software	Tsinghua University		Developed by authors
SPSS software	IBM		Data analysis software
G*Power software	Heinrich-Heine-Universität Düsseldorf		Sample size calculation
E4 portable wireless wristband	Empatica		Recording Galvanic skin response and heart rate
Arqus	Qualysis		Motion capture camera platform
Passive marker	Qualysis		Appropriate sizes: 2.5 mm, 4 mm, and 6.5 mm
Trigno sEMG	Delsys		Recording electromyographic activity
Visual Studio Code	Microsoft		Python editor

Reviewer comments

Reviewer #1:

Manuscript Summary:

The manuscript explains a descriptive analysis of the impact of touchscreen device on the human performance and musculoskeletal structure of the hand. It presents significant findings on the type of layout to be used by illustrating the differences among 4 types of keyboard. The protocol established line ups with the intended aims and objectives of the study.

Major Concerns:

There are no major concerns

Minor Concerns:

 # Comments 1:

Gap in the literature should be stated and the significance of this study should be emphasized. Figure 1 should be explained within the manuscript at lines 102-103. Figure 2 cannot be on P.4 line 116 and P.6 line 180. One of them should be preferred.

*** Response 1:**

Thanks for your kind suggestions. Figure 1 and Figure 2 have been corrected as follows:

1.

PROTOCOL:

The study was conducted in accordance with the ethical principle and was approved by the Ethics Committee of Tsinghua University. Figure 1 shows the process of evaluating the keyboard design of smartphones.

[Place **Figure 1** here]

2.

1.2 Dependent variables

1.2.1 Physical data were composed of hand length, the length of input finger, and the circumference of input finger which was measured by a tape measure as shown in Figure 2.

[Place **Figure 2** here]

1.2.2 Physiological data consist of galvanic skin response (measured by the portable wireless physiological detector), heart rate (measured by the portable wireless physiological detector), electromyographic activity (measured by Electromyography), etc.

Gap in the literature has been added and the significance of this study has been re-written as follows:

1. In the section of Summary

   The presented protocol integrates various evaluation methods and demonstrates a summarized method and Python scripts for evaluating the keyboard design on smartphones. Pairs matched by English characters is proposed as the input material and the transition time between two keys is used as the dependent variable. The representative results show whether the curved QWERTY keyboard design with size-adjustable buttons could optimize the input efficiency and experience on smartphones for one-handed usage when compared with the traditional QWERTY keyboard.

2. In the section of Abstract

   ……There is sufficient existing literature on evaluating virtual keyboards, however, little of them systematically summarized and took reflection on the evaluation methods and processes. Therefore, this protocol fills in the gap and presented a process and method of the systematic evaluation of keyboard design with code for analysis and visualization that are available, need no additional or expensive equipment required, and are easy to conduct and operate. In addition, the protocol also helps to get potential reasons for the disadvantages of the design and enlighten the optimization of designs. In conclusion, this protocol with the open-source resources not only could be an in-class demonstrative experiment to inspire the novice to start their studies but also contributes to improving the user experience and the revenue of input method editor companies.

And, based on the above, some words and sentences about the significance were re-written throughout the manuscript.

Reviewer #2:

Manuscript Summary:

The authors are presenting a "protocol" for assessing keyboard layouts on virtual keyboards (touch keyboards) with a specific focus on curved keyboards.

Major Concerns:

 # Comment 1:

The text input literature is large and also the part addressing virtual keyboards. There are some established conventions for how to conduct experimental assessments of text entry experiments. As such I do not see that this protocol adds to anything new that has not already been covered by others.

*** Response 1:**

Thanks for your comments. To be sure, there are currently many studies that use text input to evaluate virtual keyboard design. During the evaluation process, the subjects mainly input sentences from the phrase set proposed by Mackenzie and Soukoreff (2003). Traditionally, researchers collected and analyzed their Word Per Minute, Word Error Rate (objective variables), and some different subjective feedbacks. The main contributions and innovations of this paper are as follows:

- There is much existing literature on virtual keyboards but little of them summarizes the evaluation methods and processes. This article will systematically sort out the evaluation methods and materials commonly used by the previous studies. Also, this protocol could **help novice researcher to select the appropriate method and indicators for their own experiments rather than being confused with such various methods and indicators**.

- WPM commonly used in existing research is a descriptive indicator. According to our accepted paper "Usability Evaluation of Smartphone Keyboard Design from an Approach of Structural Equation Model" (coming online) by HCII 2020, WPM is not a appropriate evaluation indicator. During the collection process, it is affected by many confounding factors, such as cognitive factors. Therefore, we proposed Pairs (matching 26 letters in pairs; a total of 676 pairs) as the new text input material. All words and abbreviations (including LX, QL, which are not words, but abbreviations for companies or names) are composed of multiple pairs. Then, we added the transition time between two keys as the new objective variable to **help researchers better explore the reaction time and input behavior of the participants during the transition of fingers**. Its results can also better help designers optimize the virtual keyboard.

Based on the above, we found that the description of our contribution was unclear and **we have re-written as follows:**

- **In the section of Summary**

  The presented protocol integrates various evaluation methods and demonstrates a summarized method and Python scripts for evaluating the keyboard design on smartphones. Pairs matched by English characters is proposed as the input material and the transition time between two keys is used as the dependent variable. The representative results show whether the curved QWERTY keyboard design with size-adjustable buttons could optimize the input efficiency and experience on smartphones for one-handed usage when compared with the traditional QWERTY keyboard.

- **In the section of Abstract**

  ……There is sufficient existing literature on evaluating virtual keyboards, however, little of them systematically summarized and took reflection on the evaluation methods and processes. Therefore, this protocol fills in the gap and presented a process and method of the systematic evaluation of keyboard design with code for analysis and visualization that are available, need no additional or expensive equipment required, and are easy to conduct and operate. In addition, the protocol also helps to get potential reasons for the disadvantages of the design and enlighten the optimization of designs. In conclusion, this protocol with the open-source resources not only could be an in-class demonstrative experiment to inspire the novice to start their studies but also contributes to improving the user experience and the revenue of input method editor companies.

And, based on the above, some words and sentences about the significance were re-written throughout the manuscript.

 # Comment 2:

The protocol seems very specific to the curved keyboards which is a bit odd, and not so useful for other types of experiments. The authors have also failed to relate to the other work on "curved" or "tilted" keyboard design. There are quite a lot of such works (especially in the ergonomics literature).

*** Response 2:**

Thanks for your comment. This protocol aims at all experiments on smartphones to evaluate the keyboard design. Curved keyboard study was conducted based on the protocol, and we have nearly collected all variables (without the optional) mentioned by the protocol to fully present the protocol while existing studies might just select several indicators which were much appropriate for their research goals. This manuscript did not try to in-depth explore the usability of curved design, therefore, after the general protocol, we have not discussed the related works on “curved” and “tilted” keyboard design in the section of representative results, and we just simply showed the results and draw a conclusion. In addition, the exemplar python scripts are also available for other keyboard evaluation experiments on smartphones by just changing some parameters and functions.

However, our description about the protocol may confused readers that this protocol is specific to curved keyboard evaluation, therefore, **the content of the protocol has been refined and added more evaluation methods as follows:**

- **In the section of Protocol**

1.2 Dependent variables

……

1.2.2 Physiological data consist of galvanic skin response (measured by the portable wireless physiological detector), heart rate (measured by the portable wireless physiological detector), electromyographic activity (measured by the surface Electromyography), etc.

……

1.2.4 Body-movement data contain hand gesture and body (finger) movement. They could be collected by the motion capture system [35].

2. Procedure

……

2.3 Disinfect all devices and clean the body parts of the participant that will touch the devices.

2.3.1 Ask participants to wash their hands and clean the screen of smartphones so that sensors of smartphones can be more sensitive.

2.3.2 Wear portable wireless physiological detectors or motion capture system for participants. Wear the portable wireless physiological detection wristband on the non-dominant hand of participants to record galvanic skin response and heart rate with the noise interference avoided. Place passive markers of the motion capture system on the fingernails, the proximal phalanx of the finger, cervical vertebrae (C3-C5), and arm, to collect the precise body and finger movement. Stick wireless electrodes to the skin of two arms and two forearms to detect the electromyographic activity. (Optional step)

2.3.3 Calibrate all the devices used in the experiment.

- **In the section of Representative Results**

The representative study is mainly following the mentioned protocol. The study adopts a 2 (Keyboard layout: Curved QWERTY vs. Traditional QWERTY) × 2 (Button size: large, 6.3 mm × 9 mm vs. small, 4.9 mm × 7 mm) within-subject design to evaluate whether the curved QWERTY could improve the input efficiency and comfort when compared with

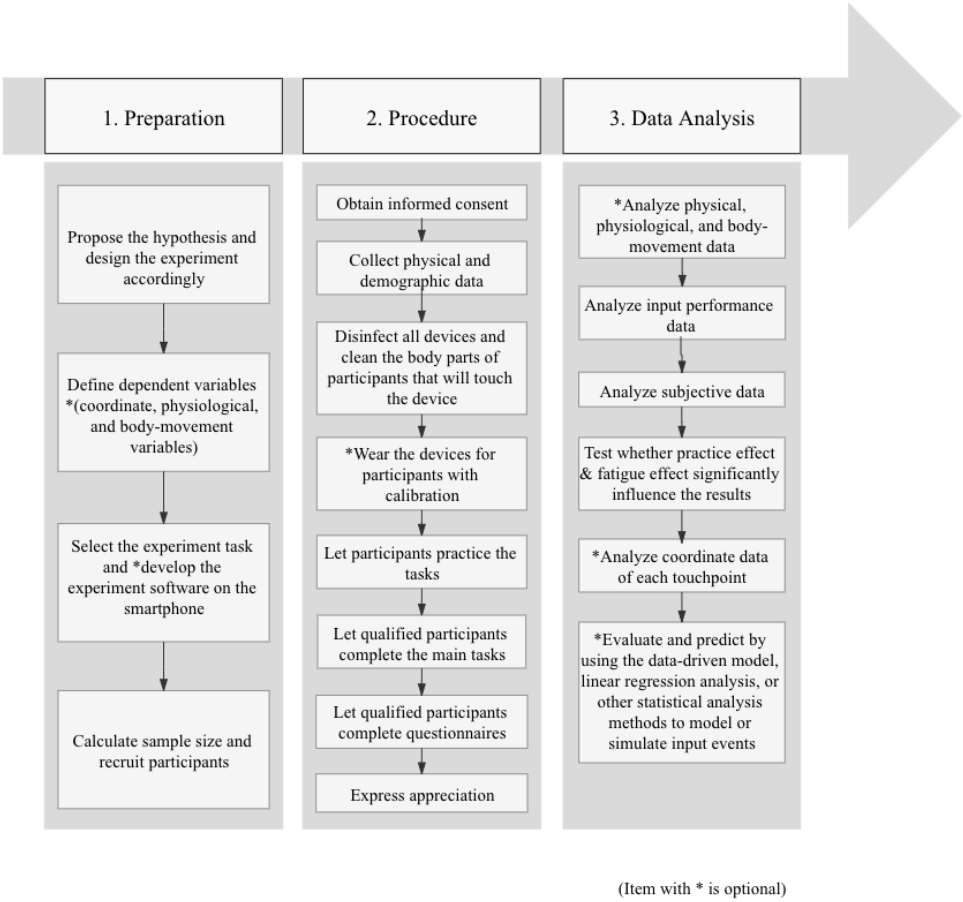
the traditional QWERTY in different sizes of buttons by the character pair input task through our self-developed software (Figure 3). This study has not adopted the expensive physiological detector equipment or motion capture system, and the data analysis did not contain the modeling or simulation.

- In the section of Discussion

Limitation (lines 615-618):

……Besides, other expensive devices or equipment adopted by previous studies have not been included in the representative results, such as the portable wireless physiological detector or motion capture system and researchers should choose their specific experimental devices based on their research problem and hypothesis……

Based on the above, some words and sentences about the significance were re-written throughout the manuscript. Also, Figure 1 (the evaluation process) was changed as follows:



(Item with * is optional)

 # Comment 3:

The statistical analysis part of the protocol lacks reflection on the appropriateness of the statistical tests used. The authors simply rely on parametric tests, while in some situations the assumptions for these tests are not satisfied and non-parametric substitutes are needed. Others would argue that the hypothesis testing paradigm is old fashioned (the followers of the New Statistics, or Bayesian enthusiasts).

*** Response 3:**

Thanks for your kind suggestions. To be sure, the statistical analysis part of the protocol should pay much attention to both parametric tests and non-parametric tests. Although this protocol concentrates on evaluation, and hypothesis testing is commonly used, modeling and other statistical analysis methods could also enlighten the evaluation by their indexes or indicators. In the section of data analysis, we have added some new methods, such as modeling and simulation (do not appear in representative results). In the limitation, we hope that researchers could explore more evaluation analysis method. Thanks a lot!

- **In the section of Data Analysis**

3. Data analysis

3.1 Hypothesis testing by appropriate parametric or non-parametric tests

3.1.1 Analyze the physical, physiological, body-movement data to test whether the difference between participants would significantly influence the results and inexpressive input experience of users. (Optional step)

3.1.2 Analyze the input performance of participants to test the input efficiency on the keyboard.

3.1.3 Analyze subjective data to test the perceived usability and subjective feedback of the keyboard.

3.1.4 Figure out whether the practice effect and fatigue effect significantly influences the result. For each condition, trials are divided into two parts according to the timestamp, i.e., the first half part and the second half part. Specifically, under each condition, examine the difference of input performance between the first half part and the second half part to test whether it exists the practice effect or fatigue effect.

3.1.5 Analyze the area of the fitted ellipse of touchpoints on each button as well as the offset from its center to the target center of each button. (Optional step)

3.1.5.1 All the touchpoints of each button were collected by our self-developed software, and they roughly accord with the bivariate Gaussian distribution. The 95% confidence interval of each button in both x- and y-directions is derived through the coordinate data of each touchpoint in pixel, and the 95% confidence ellipses over a 1:1 outline of the button for each keyboard is fitted through Python scripts on pixel coordinate (see coding file 2).

3.1.5.2 Fitted ellipses (95% CI) and their areas demonstrate the dispersion of touchpoints on

each button. In each button, the offset of fitted ellipse calculated by Python scripts is defined as the center point of the fitted ellipse to the target point of the button, and it could be represented from x- and y-directions, i.e., in X-axis and Y-axis (see coding file 3).

3.2 **Modeling and simulation**

3.2.1 Use the data-driven model as a function of keyboard location and orientation to predict the finger movement by Python scripts. All movements of fingers are divided into 8 directions [38] (the top to the bottom, the bottom to the top, the left to the right, the right to the left, the left-top to the right-bottom, the right-bottom to the left-top, the left-bottom to the right-top, the right-top to the left-bottom). For each direction, the average transition time between two keys is calculated to represent the effectiveness of finger movement which is used to evaluate the keyboard design. (Optional step)

3.2.2 Use linear regression analysis to build an enhanced Fitts' Law (or its extended version, FFitts Law) model to predict the transition time between two keys by using an integrated cognitive architecture [39] by Python scripts. The enhanced Fitts' Law model could provide a better prediction and evaluation on keyboard design based on its analyses on the location and effective width of keys, as well as the distance of two keys. (Optional step)

- In the section of Discussion

  Limitation (lines 618-620)

  ……Finally, followers of the New Statistics or Bayesian enthusiasts could try to adopt more statistical methods to analyze and evaluate the keyboard design.