

Drunk User Interfaces: Determining Blood Alcohol Level through Everyday Smartphone Tasks

Alex Mariakakis¹, Sayna Parsi^{2,3}, Shwetak N. Patel¹, Jacob O. Wobbrock³

¹Computer Science & Engineering, ²Human Centered Design & Engineering, ³The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
{atml5, shwetak}@cs.washington.edu, {parsi, wobbrock}@uw.edu

ABSTRACT

Breathalyzers, the standard quantitative method for assessing inebriation, are primarily owned by law enforcement and used only after a potentially inebriated individual is caught driving. However, not everyone has access to such specialized hardware. We present *drunk user interfaces*: smartphone user interfaces that measure how alcohol affects a person’s motor coordination and cognition using performance metrics and sensor data. We examine five drunk user interfaces and combine them to form the “*DUI app*”. *DUI* uses machine learning models trained on human performance metrics and sensor data to estimate a person’s blood alcohol level (BAL). We evaluated *DUI* on 14 individuals in a week-long longitudinal study wherein each participant used *DUI* at various BALs. We found that with a global model that accounts for user-specific learning, *DUI* can estimate a person’s BAL with an absolute mean error of $0.005\% \pm 0.007\%$ and a Pearson’s correlation coefficient of 0.96 with breathalyzer measurements.

Author Keywords

Situational impairments; alcohol; mobile; smartphones; health; drunkenness; inebriation; safety; driving.

ACM Classification Keywords

K.4.1. Computers and Society: Public Policy Issues – *human safety*; J.3. Computer applications: Life and Medical Sciences – *health*.

INTRODUCTION

In 2014, 27 people died every day as a result of drunk driving in the United States [35]. Portable breathalyzers were invented in 1931 [19] to allow law enforcement to prosecute cases of drunk driving; however, breathalyzers

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

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5620-6/18/04...\$15.00

<https://doi.org/10.1145/3173574.3173808>

are typically used after a drunk driver has been caught, rarely to prevent people from driving in the first place. Jewett *et al.* [14,23] estimate that the average drunk driver has driven drunk over 80 times before their first arrest. There remains a need of being able to catch cases of drunk driving without the presence of law enforcement or relying on people to determine their own limits for personal safety.

One can view inebriation as a temporary “situational impairment” that affects a person as they interact with the world around them [36,40,41,46,51]. From this ability-based perspective, we propose *drunk user interfaces (DUIs)*: smartphone-based tasks that challenge and assess a person’s motor coordination and cognition. When a person manipulates a drunk user interface, the smartphone can measure how well that person performs the required task using human performance metrics and features derived from embedded sensors (*e.g.*, the touchscreen, accelerometer). For example, a person’s ability to type a sentence on a smartphone can be measured by both counting typing errors and by measuring how the user strikes keys using accelerometer and touchscreen data.

In this paper, we describe and evaluate five different drunk user interfaces. We combine different drunk user interfaces into a single smartphone app that creates a detailed snapshot of a person’s abilities. We call this app the *Drunk User Interfaces* app, or *DUI* (pronounced “doo-eee”).

What would motivate a person to use *DUI* in the first place? We envision a number of possible use cases:

1. Services like OnStar from General Motors can allow individuals to unlock their vehicles with their smartphones¹. A car insurance company could offer a discount to customers who agree to use *DUI* whenever they try to unlock their car after 10 PM or leave an establishment that serves alcohol. If they fail *DUI*, their car will not start.
2. Bartenders are obliged to refuse service to customers who seem overly intoxicated. Either a bartender or a customer may wish to check their

¹ <https://www.onstar.com/us/en/services/services.html>

blood alcohol level (BAL) to ensure safe drinking behavior.

3. Many teenagers fear “drunk texting” – when a person sends a text message that they normally would not because alcohol has impaired their judgment. *DUI* failures could lock a person out of his or her messaging app until the next day.
4. Individuals might benefit from increased self-awareness or education about how they respond to alcohol and how quickly their motor coordination and cognition degrade.

DUI measures the side effects that alcohol has on a person’s own abilities, not the alcohol concentration in a person’s blood directly. Furthermore, some of our proposed use cases only require a binary decision between sobriety and inebriation, not a precise estimate of BAL. Nevertheless, we strive to achieve the most difficult goal possible: estimating a person’s BAL. We do this through a data-driven approach. We collected data from 14 participants in a 5-day longitudinal study where participants used *DUI* at various BALs. This study design provides several benefits over previous alcohol studies in the HCI community [2,22,26], the main benefits being that it allows us to account for learning effects and control for fatigue, which can result in behavior that appears similar to inebriation. Using a combination of five different drunk user interfaces, *DUI* is able to estimate BAL with a mean absolute error of $0.005\% \pm 0.007\%$ when the app accounts for the user’s learning curve².

The task interfaces comprising the design of *DUI* are not necessarily novel; most of the tasks are borrowed from literature in the HCI and medical communities [18,21,27]. Rather, their combination in *DUI* and their ability to produce data that informs an accurate BAL estimate and inebriation decision are the key breakthroughs in this paper.

The two primary contributions of this work are: (1) the *DUI* app, comprising (a) tasks that challenge a person’s psychomotor control in a mobile setting, and (b) the use of machine learning to translate a person’s performance into a BAL estimate; and (2) a 14-person longitudinal study of *DUI* demonstrating its ability to track different BALs for the same user against a breathalyzer baseline.

RELATED WORK

DUI draws inspiration from work at the intersection of situational impairments and mobile devices. We briefly highlight some of this work, followed by a summary of research and products aimed at measuring BAL.

Situational Impairments

We view inebriation as a situational impairment [36,40,41,46,51], *i.e.*, a factor that affects a person’s ability

to interact with others and the world around them. Situational impairments can be imposed by the user’s external environment (*e.g.*, cold weather [12]), by internal changes (*e.g.*, medicine-induced motor-impairment [45]), or by a combination thereof (*e.g.*, divided attention [34]). Smartphones bring situational impairments to the forefront because they are used in a variety of different mobile scenarios [24]; at the same time, smartphones are instrumented with sensors that can interpret and understand these scenarios, providing the opportunity for ameliorating the effects of situational impairments within them [50].

These works and others view situational impairments as problems that can be addressed by sensing the user’s current state and adapting the interface accordingly. In this work, however, we stop short of adapting the interface and instead use the sensed indicators of the user’s state to train a machine learning model that outputs a description of the user’s state – specifically a BAL measurement.

Hardware for Measuring Alcohol Consumption

Breathalyzers are the *de facto* method of measuring BAL outside of a medical setting [4]. Most people are familiar with the handheld breathalyzers carried by law enforcement, but companies have produced different form factors for personal use. For example, Tokyoflash³ produces an LCD watch with a built-in breathalyzer for \$139.00 USD. At one point, Breathometer⁴ produced a breathalyzer that could interface with a smartphone via the audio jack or Bluetooth, for \$49.99 and \$99.99 USD, respectively; the FTC later initiated an investigation and found their accuracy claims to be false [9].

There are other methods for measuring BAL that are meant to be easier than a blood draw. TruTouch⁵ is a device that measures BAL non-invasively using a method called photoplethysmography (PPG). Alcohol slightly changes the blood’s color, which can be quantified by shining different wavelengths of light onto the fingertip and measuring the intensity that is reflected back. SCRAM has a device that measures BAL through the wearer’s perspiration every 30 minutes⁶. The device is intended for high-risk drunk driving offenders who are court-ordered to monitor their drinking behavior. Finally, Jung *et al.* [25] developed a smartphone attachment that performs color analysis on pads that react to saliva.

Each of these systems is able to measure BAL at some biological level; however, these systems either require the purchase of extra hardware or certain specifications from a person’s smartphone. We view these as limitations towards

²In the United States, BAL is typically reported as the fraction of a person’s blood that contains alcohol by volume. The units are interchangeable with g/dl ($0.10\% = 0.10$ g/dl).

³<https://www.tokyoflash.com/en/watches/kisai-intoxicated-silicone>

⁴<https://www.breathometer.com/>

⁵<http://ttinc.com/>

⁶<https://www.scramsystems.com/products/scram-continuous-alcohol-monitoring/>

ubiquitous BAL sensing, which is why we propose drunk user interfaces that can work on an unmodified smartphone.

Mobile Software for Measuring Alcohol Consumption

Because smartphones are ubiquitous, researchers have explored ways that mobile devices can be used to curb alcohol abuse without supplemental hardware. One area where smartphones have been used is education. Hundreds of publicly available apps, such as BAC Calculator⁷ and IntelliDrink PRO⁸, allow users to log their drinking behavior. Using demographic information (*e.g.*, height, weight) and data on the drinks themselves (*e.g.*, proof, frequency, quantity), these apps estimate the users' BAL; however, a study by Weaver *et al.* [47] found that the estimates reported by 98 such apps were inaccurate compared to a breathalyzer. Of course, these apps also rely on self-report, which is prone to error.

Shifting to more automatic means of sensing inebriation, Hossain *et al.* [22] mined geotagged tweets to determine whether or not people were drunk. They assumed that tweets with words like “hangover” and “drunk” came from drunk individuals. They then propagated that inference to tweets that were posted by the same person near that time. One of the most common tasks explored by the HCI and ubicomp communities for predicting inebriation is gait analysis. The vision of these projects is an app that continuously processes the smartphone's accelerometer data for features such as step amplitude and cadence variation [2,26]. BreathalEyes [5] reports a BAL estimate by detecting nystagmus, or involuntary eye movement, during horizontal gaze shifts. To the best of our knowledge, there is no publicly available study that describes BreathalEyes' accuracy. Our work is most similar to that of Bae *et al.* [3], who detected heavy drinking episodes in a study involving the collection of mobile sensor data and experience sampling methods for ground truth. Their sensor data included location, network usage, and motion data. Unlike our work, Bae *et al.* did not use human performance data. They also made a categorical assessment (sober, tipsy, or drunk), not a continuous-scale BAL estimate as we do.

THE DESIGN OF *DUI*

The *DUI* app comprises five different drunk user interfaces: (1) typing, (2) swiping, (3) balancing+heart rate, (4) simple reaction, and (5) choice reaction. For each task, we cite a subset of clinical experiments that informed them, how they were adapted for use on a mobile device, and some of the features calculated on human performance and sensor data. Unfortunately, limitations of space preclude a complete listing of every feature used for each task. A more detailed listing can be found on the project's webpage⁹. We then

describe how those features are processed and analyzed to produce a final BAL estimate.

(1) Typing Task

DUI's typing task is intended to measure the user's fine motor coordination abilities and cognition as they text. Anecdotal evidence suggests that texting is more difficult while a person is inebriated; to the best of our knowledge, though, there has been no work that has quantitatively analyzed the effect of alcohol on smartphone touchscreen typing. However, research in medicine and psychology has examined similar tasks that require small, controlled movements, such as the Purdue Pegboard Test [6].

For *DUI*'s typing task, the user is presented with a random phrase from the MacKenzie-Soukoreff phrase set [33] and asked to type the phrase “as quickly and accurately” as possible, relying on their own internal speed-accuracy tradeoff. Auto-correct is disabled, and no cursor is provided for the user to jump back to make corrections; if the user makes a mistake, they must decide for themselves whether or not to remedy the mistake with a backspace or to leave it. We imposed these restrictions in keeping with standard text entry evaluation methodology [52].

There are two levels of features that emerge from this test. At a high level, *DUI* utilizes the error rate analysis proposed by Soukoreff and MacKenzie for text entry analysis [42]. In such an analysis, each character is classified into one of four categories: “correct” (C), “fix” (F), “incorrect fixed” (IF), and “incorrect not fixed” (INF). *DUI* calculates different text entry metrics involving these character categories that not only measure how often the user made mistakes, but also how often they decided to correct those mistakes. Other quantities that can be calculated include “utilized bandwidth” (*i.e.*, the fraction of correct keystrokes made) and “participant conscientiousness” (*i.e.*, the fraction of mistakes corrected):

$$\text{utilized bandwidth} = \frac{C}{C + F + IF + INF}$$

$$\text{conscientiousness} = \frac{IF}{IF + INF}$$

At a lower level, *DUI* examines the mechanics of the user's typing through the touchscreen, accelerometer, and gyroscope, similar to how Goel *et al.* [16] used those sensors to compensate for typing errors that were made while walking. *DUI*'s typing task uses a custom keyboard, similar in appearance to the smartphone's default keyboard, which records the precise position and radius of each touch. From this data, *DUI* calculates features like the Euclidean distance between the center of the selected key and the user's touch position. Motion sensor features include the peak acceleration before a touch and variation in phone orientation during the task. One interesting hypothesis within this task is that people could have different reactions to mistakes that could be detected through sensor data. If a person is drunk, they could overreact to the mistake and

⁷<https://play.google.com/store/apps/details?id=com.simonm.bloodalcoholcontentcalculator>

⁸<https://itunes.apple.com/us/app/intellidrink-pro-blood-alcohol-content-bac-calculator/id440759306>

⁹ https://atm15.github.io/extra/DUI_feature_list.csv

jostle their hand in a more pronounced manner than if they were sober; on the other hand, they may overlook the mistake and not react at all.

(2) Swiping Task

Whereas the *DUI* typing task measures fine motor control in the form of repeated target selection, the swiping task measures fine motor control through gesturing. The swiping task can be considered a progressive goal-crossing task where the user is asked to pass through different targets [1]. For feature extraction, we also treat the swiping task as a steering task with implicit tunnels. To our knowledge, the effect of inebriation on swiping gestures has yet to be explored, but there have been related studies involving tracing. Hindmarch *et al.* [21], for example, saw that participants' ability to track a moving target with a joystick worsened after consuming alcohol.

The swiping task shows a screen that mimics the 3×3 lock screen of many Android devices (Figure 1). The user traces a random 4-digit passcode on the screen. The passcode is generated in such a way so that the user must change the direction of his or her finger after each digit. Each circular cell in the grid has a moderate diameter, but a digit is only triggered if the user's finger passes over the small gray center.

Although the user believes that they are simply entering a passcode for accuracy, the features *DUI* calculates for the swiping task come from comparing the trajectory of the user's finger (solid trace, Figure 1) to the ideal 3-segment shape that connects the 4 digits (dashed lines, Figure 1). The user does not see the ideal trajectory, only their own trace. Of course, the user is not expected to move their finger from point-to-point in the most efficient manner possible, but the hypothesis is that the user's finger would move more efficiently while sober than while drunk. One metric we use to compare the gesture shapes is the proportional shape matching metric described by Kristensson and Zhai [30], which compares the form of two shapes regardless of when their points are sampled. We also examine each gesture segment individually by slicing the data between the time when the user's finger enters and exits the gray center of the digit cell. For each segment, we calculate the path-based accuracy features proposed by MacKenzie *et al.* [32] for evaluating how a trajectory between two points deviates from the shortest path between them. For example, "movement variability" measures the standard deviation of the distance between the ideal path and the user's trajectory. Finally, we also calculate time-based measurements, such as maximum finger velocity, acceleration, and jerk, for each segment; these features were found to be informative by Flash and Hogan for characterizing human motion [15].

(3) Balancing+Heart Rate Task

DUI's balancing+heart rate task serves two purposes. The original intent of the task was to measure just the user's heart rate; a person's average heart rate slows down after

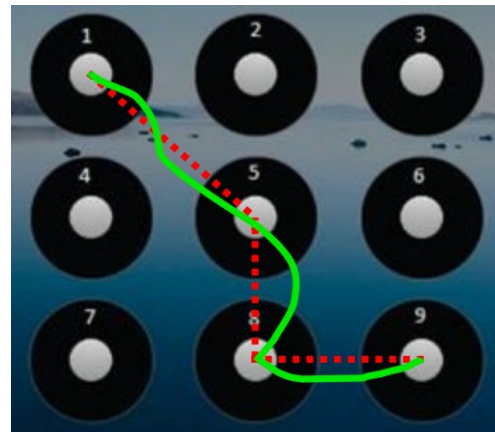


Figure 1. The *DUI* swiping task resembles an Android 3×3 lock screen. The straight dashed red lines show the ideal gesture (hidden from the user) for the code 1-5-8-9, while the curvy solid green path shows the user what they have drawn.

alcohol consumption because of alcohol's depressive effects [39]. Han *et al.* [18] recently demonstrated a method of measuring heart rate using a technique called photoplethysmography (PPG) through the smartphone camera. In short, PPG measures the transparency of the finger as blood rushes in and out while circulating. For the PPG measurement to be clear, the user must hold his finger completely still on the camera. We realized that this also offers the chance for a test that challenges the user's coordination while their heart rate is being measured. For example, Tianwu *et al.* [43] cite diminished vestibular control with alcohol consumption.

In our *DUI* task, the user is instructed to hold the smartphone parallel to the floor. The user is then told to place their index finger over the flash and the camera simultaneously so that their heart rate can be measured for 10 seconds. The user sees two widgets on the bottom of the screen. One widget shows a preview of what the camera sees so that the user can adjust his or her fingertip if it is not in the correct position. The other widget shows a constantly updated "flatness score" that the user is supposed to keep as high as possible; unbeknownst to the user, the score is a function of the accelerometer reading along the *z*-axis (*i.e.*, through the screen).

The features for the balancing+heart rate task relate to both the user's heart rate and their ability to keep the smartphone flat. The user's average heart rate is measured using Han *et al.*'s PPG algorithm [18] from the camera video. If the calculation fails or the algorithm misses a couple of beats, *DUI* uses that as an indication that the user was unable to comply with the instructions, which could indicate inebriation. The user's ability to maintain balance with his or her hand is measured using the standard deviation of the acceleration in the *z*-direction.

(4) Simple Reaction Task

DUI's simple reaction task is intended to capture the user's alertness and, to a lesser extent, motor speed. Multiple

studies [21,37] have linked alcohol consumption to impaired reaction times. *DUI*'s task for measuring reaction is a variation of PVT-Touch [27], a smartphone-based version of the clinically validated Psychomotor Vigilance Task (PVT) by Dinges and Powell [11] to measure alertness. *DUI* utilizes two of the four touchscreen input techniques that were investigated in Kay *et al.*'s work on PVT-Touch: “touch down” and “finger lift”. These gestures were selected because Kay *et al.* found that the “touch down” gesture was most comparable to the traditional PVT and the “finger lift” gesture was the most precise.

For *DUI*'s simple reaction task, the user is asked to perform a “touch down” gesture and then a “finger lift” gesture in response to a randomly-timed stimulus. That stimulus is a single square shown in the middle of the screen. When the screen changes from red to green, the user must perform a “touch down” gesture; when the square changes from green to red, the user must perform a “finger lift” gesture. The events were spaced within a 7-second period such that the “touch down” would occur randomly within the first 3 seconds and the “finger lift” would occur randomly within the last 3 seconds. The user was not instructed to use a particular finger, but we found that most used their thumb.

From a human performance standpoint, *DUI* records the time difference between the square's color change and the expected action, *i.e.*, “touch down” or “finger lift”. From a sensing standpoint, *DUI* records data from the touchscreen, accelerometer, and gyroscope. It also records touch pressure through the touchscreen and the motion of the smartphone as the user performs the task.

(5) Choice Reaction Task

Like the simple reaction task, the choice reaction task is intended to assess alertness and motor speed; we treat the two tasks independently, as psychology has done. Instead of the single square in the middle of the touchscreen, the choice reaction task for *DUI* shows four squares arranged in a 2×2 grid. Only one of the four squares, selected at random, changes from red-to-green and then green-to-red. In addition to the features described for the simple reaction task, *DUI* also computes the user's accuracy at selecting the correct square.

Excluded Tasks

Many other tasks could be made into drunk user interfaces, each with their own intended purpose, benefits, and drawbacks. We explored a few in concept or in practice. For example, we considered walking [2,26], but felt that requiring the user to move would lead to a poor user experience. We also considered speech analysis [7,28], but the diversity of accents led to difficulties. Finally, we considered short term memory [38], but the typical word recall task simply took too long (over one minute).

Machine Learning

Each task generates a set of human performance metrics that can be used as features for training a regression model

that estimates BAL. Not only are the human performance metrics of an individual trial interesting, but also the variation of those metrics across different trials. For instance, a person may have the same average reaction time when they are sober and when they are drunk, but they may have a larger spread of times while drunk. In our user study, we asked participants to perform each task multiple times. The performance metrics across different trials of the same task are aggregated using means and standard deviations.

Fifty-one features are available for training, but some are more informative for estimating BAL than others. Automatic feature selection is used to select the most explanatory features and eliminate redundant ones. The top 25% of the features that explain the data according to the mutual information scoring function are used in the final models. Mutual information measures the dependency between two random variables [29]. Automatic feature selection works best when all of the features are normally-distributed. We assume that this is the case with most of the features except for those that are time-based (*e.g.*, reaction times). Prior research has noted that such measures tend to be log-normally distributed [8,31], so they are log-transformed after they are aggregated before feature selection and training.

DUI uses random forest regression models [44] for estimating BAL. A single decision tree regressor would force features to be split sequentially in the same tree; random forest regression learns shallower, more isolated trees instead, reducing the possibility of nonsensical interactions between features across tasks. The disadvantage of random forest regression is that it cannot extrapolate beyond the BAL levels that were reached in the study. Models like linear regression can extrapolate, although there is no guarantee that they would do so correctly. We chose random forest regression because it outperformed the other models we tried for the data we had. The feature extraction and machine learning models were built in Python using the scikit-learn package.

USER STUDY FOR *DUI*

We conducted a longitudinal user study of *DUI* with the intent of collecting human performance data at different BALs for the same users over time. Our study design allowed us to control for fatigue while modeling any learning that occurred as users gained familiarity with *DUI*'s tasks.

Participants

Fourteen participants (9 male, 5 female) ranging from 21 to 35 years old ($M = 25.7$, $SD = 4.8$) were recruited for our study. The participants were a mix of Caucasian, Asian, and South Asian races. All participants owned and used a smartphone on a daily basis.

Apparatus

Participants used our custom smartphone app on a third-generation Moto G smartphone that has a 5-inch capacitive

screen with 720×1080 pixels. The app was designed with five different screens, one for each of the drunk user interfaces. Each time a person used *DUI*, they saw the typing, swiping, and balancing+heart rate task in that order five times; after that, they saw the simple reaction task five times in a row, and then the choice reaction task five times in a row. In other words, one “use” of *DUI* entailed five trials of each task in our study. The task order was selected for the participants’ convenience.

Each screen had a consistent presentation, including instructions at the top, a button to start and stop the task, and a red “recording light” icon to indicate when data was being recorded. Navigation between the tasks happened automatically; even if the participant felt that they made a mistake, redoing tasks was disallowed and the app progressed to the next screen.

Procedure

Prospective participants were required to satisfy guidelines set by the National Advisory Council on Alcohol Abuse and Alcoholism (NIAAA)¹⁰ in order to participate in the study. They also had to provide state- or federally-approved identification that verified that they were at least 21 years old, the legal drinking age in the United States. Participants also had to confirm that they did not have a family history of alcoholism and that they were not taking any medication that interacted with alcohol. Finally, in accordance with the NIAAA and at the insistence of our IRB, female participants were required to take a pregnancy test on the first day of the study to confirm that they were not pregnant.

Participants satisfying those criteria were scheduled to participate in our study for five sessions. The sessions were scheduled in 24-hour intervals with a tolerance of one hour; if there was a scheduling conflict, the remaining sessions were pushed by another 24 hours to maintain time-of-day. No two sessions were scheduled more than 48 hours apart. Maintaining this schedule was important as it allowed us to control for time-of-day related fatigue. Had participants done one session in the afternoon and another session late at night, the latter session would have included the confounding effects of fatigue. The default schedule was each weekday (Monday through Friday) at 4 PM. Although 4 PM is earlier than when most people start drinking, starting early helped make each session quicker since it was right before people had eaten dinner.

In the first session, the participant was introduced to *DUI*. A research staff member explained each task, but did not mention the specific metrics that were being recorded. Participants were allowed to operate the smartphone with their own texting style (e.g., one finger, two thumbs, etc.) and swiping posture. The only restriction was that they had to hold their phone in their hands.

Once the participant had used *DUI* in the first session, they were free to leave. The remaining four sessions started with the participant using *DUI* sober as before. Once the participant finished using *DUI* sober, he or she was required to reach a predetermined BAL between 0.02% and 0.08% in increments of 0.02%. The alcohol levels were increased incrementally for all participants as a safety precaution; if someone felt uncomfortable at a lower BAL, they were free to withdraw without consequence.

The decision to use 0.08% as the maximum BAL was for both safety and practicality. The NIAAA recommends that research participants not be given more alcohol than they would normally consume unless absolutely necessary; we decided with our IRB that 0.08%, the legal limit in the United States, would be a reasonable limit. Another relevant guideline was that participants should not leave the study until they were back below 0.04% to ensure they would not drive while impaired. Increasing the upper BAL limit would have led to a longer study and possibly more attrition.

To reach the target BAL, we used the same procedure as Hashtroudi *et al.* [20]. The research staff member estimated the amount of alcohol needed based on the participants’ weight, the prescribed frequency of alcohol administration, and the alcohol’s proof¹¹. Based on that estimate, the participant was given one shot of 80-proof vodka (1.5 fluid ounces, 40% alcohol) every ten minutes. Once the alcohol was consumed, participants waited for 15-20 minutes to allow their bodies to absorb the alcohol. Participants then had their BAL measured with a breathalyzer after rinsing their mouths with water. Participants used *DUI* once their BAL was within 0.003% of the target. If their BAL was too low, they were given at most one shot and then delayed for their body to absorb the alcohol; if their BAL was too high, they waited and periodically used the breathalyzer until they reached the target. If a participant’s BAL did not reach the target within two hours, he or she used *DUI* at their current BAL regardless.

Participants took 261.1 ± 27.4 seconds on average to use *DUI* once in its entirety, including five trials of the five tasks. The only two tasks that did not have a fixed duration by design were the typing (10.8 ± 6.7 s) and swiping (3.4 ± 2.2 s) tasks.

Design & Analysis

Early on, we found that a model without any user-specific adjustments would be infeasible. Each participant had their own baseline abilities, which made it difficult to compare results across users. Two different methods of user-specific calibration were explored to account for this confound.

The first method compares performance to that of the sober first session (Figure 2, left). This would be akin to asking

¹⁰ <http://www.niaaa.nih.gov/Resources/ResearchResources/job22.htm>

¹¹ http://www.clevelandclinic.org/health/interactive/alcohol_calculator.asp

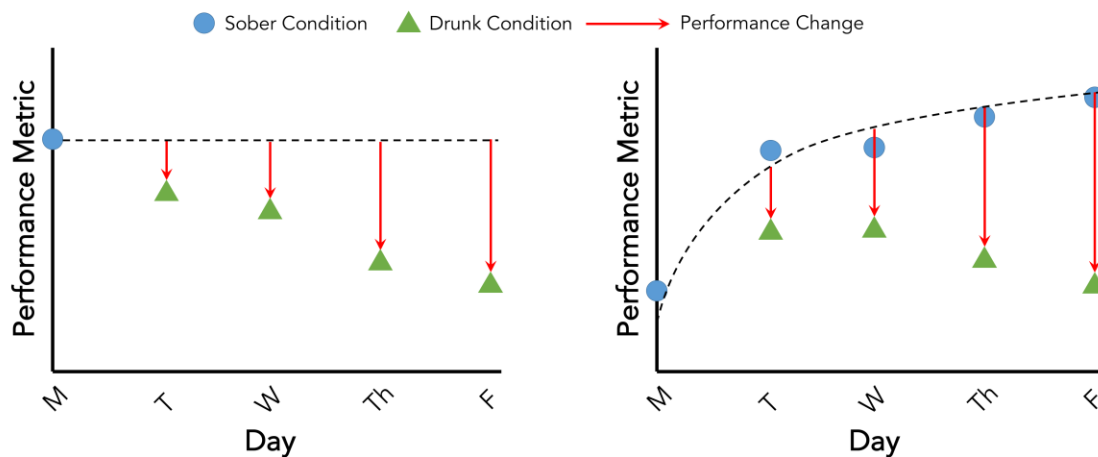


Figure 2. Idealized illustrations of two user-specific calibration methods enabled by the study design: **(left)** single-day baseline, where all performance metrics are compared to the first session, **(right)** learning curve, where all performance metrics are compared to a learning curve fit to the sober data. Our evaluation reveals that calibration with a learning curve leads to better accuracy results.

the user to go through *DUI* once upon installation. Rather than using the raw performance metrics calculated from a particular session, the features given to the model are the difference between a given session's features and those from the initial session. The problem with this approach is that it does not account for learning.

The second method compares performance to a learning curve fit to sober measurements taken at the beginning of each session (Figure 2, right). The measures recorded when the participant was sober were collected and indexed according to the number of times the participant had used *DUI* before (*i.e.*, 0th, 1st, 3rd, 5th, and 7th). The metrics were then individually fit to exponential learning curves of the following form:

$$Y = aX^b$$

where X is the session number index, Y is the performance measure, and a and b are regression coefficients for the user's baseline and learning rate, respectively. Once this learning curve was found, the contribution due to alcohol consumption could be isolated by subtracting the contribution of learning estimated for that given session. Not all metrics were considered learnable. Features that were considered performance-based (*e.g.*, swiping accuracy, reaction time) were fit to learning curves, whereas more biological-based features (*e.g.*, heart rate, typing motion acceleration) were not.

All experiments were conducted using leave-one-out cross-validation across users. In other words, to generate results for *DUI* for each participant, that participant's data was excluded from training *DUI*'s models. Once the models were trained, three different measures were obtained: (1) the absolute mean error of *DUI*'s BAL estimates compared

to the breathalyzer readings, (2) the Pearson correlation coefficient of *DUI*'s BAL estimates compared to the breathalyzer readings, and (3) sensitivity (true positive) and specificity (true negative) rates when classifying individuals as sober or drunk. For our purposes, we set our model's decision boundary for sobriety at 0.04% in accordance with the National Advisory Council on Alcohol Abuse and Alcoholism. In this context, sensitivity was defined as how often *DUI* correctly identified drunk individuals, whereas specificity was defined as how often *DUI* correctly identified sober individuals.

RESULTS

In this section, we present the results of three different experiments. The first experiment determined whether *DUI* would perform better with a single calibration session or multiple. The second experiment investigated whether or not multiple trials for the same drunk user interface had an effect on performance. The third experiment examined how well *DUI* can estimate BAL when multiple drunk user interfaces are combined.

It should be noted that two participants did not complete the entire protocol. One participant withdrew after the third session because she was uncomfortable reaching an elevated BAL, while the other withdrew after the fourth session because of short-notice travel plans. Nevertheless, their data is included in our analysis since they went through enough of the protocol to experience different BALs. This means that $(12 \text{ participants} \times 5 \text{ sessions}) + (1 \text{ participant} \times 4 \text{ sessions}) + (1 \text{ participant} \times 3 \text{ sessions}) = 67$ sessions were included in our analyses.

Methods of User-Specific Calibration

Our first experiment compared the two methods of user-specific calibration: single-day baseline and learning curve

Table I. Absolute mean error and Pearson correlation coefficient using different numbers of trials for the same task.

Task	1 trial		2 trials		3 trials		4 trials		5 trials	
	Error (%)	<i>r</i>	Error (%)	<i>r</i>	Error (%)	<i>r</i>	Error (%)	<i>r</i>	Error (%)	<i>r</i>
T	0.017 ± 0.016	0.65	0.017 ± 0.016	0.64	0.019 ± 0.018	0.55	0.018 ± 0.018	0.58	0.016 ± 0.015	0.68
S	0.017 ± 0.015	0.65	0.016 ± 0.015	0.66	0.016 ± 0.013	0.72	0.016 ± 0.015	0.67	0.015 ± 0.015	0.69
BHR	0.019 ± 0.019	0.54	0.020 ± 0.018	0.53	0.019 ± 0.019	0.55	0.020 ± 0.019	0.53	0.017 ± 0.018	0.57
SR	0.017 ± 0.017	0.63	0.018 ± 0.017	0.60	0.018 ± 0.017	0.60	0.018 ± 0.017	0.59	0.017 ± 0.016	0.65
CR	0.004 ± 0.009	0.95	0.004 ± 0.007	0.96	0.005 ± 0.008	0.94	0.004 ± 0.007	0.96	0.001 ± 0.008	0.97

T = typing, S = swiping, BHR = balancing+heart rate, SR = simple reaction, CR = choice reaction

Table II. Selected features for the single task experiment.

Task	Most important features
Typing	Mean touch radius while lifting Mean distance from key center Mean force during touch Mean touch duration
Swiping	Mean segment speed Min segment speed Min segment jerk Mean throughput Mean touch radius
Balancing+Heart Rate	Mean heart rate
Simple Reaction	Mean finger lift time
Choice Reaction	Mean finger lift time

(Figure 2). All tasks and all trials were used for this analysis to provide the most information possible to the regression models. Using the single-day baseline, *DUI* achieved an absolute mean error of $0.015\% \pm 0.013\%$, much higher than the breathalyzer's claimed accuracy of 0.005% . *DUI* with the single-day baseline also had a Pearson correlation coefficient of 0.73. The model generally underestimated user BAL by an average of 0.002% ; however, the model led to a sensitivity of 81.8% and a specificity of 61.8%. The absolute mean error of the model that factored in learning across multiple sessions was $0.005\% \pm 0.007$, and the Pearson correlation coefficient was 0.96. Unsurprisingly, both the sensitivity (87.9%) and specificity (91.2%) of that model were better than the model calibrated off the single sober session. Although less practical, accounting for user learning led to much stronger results. Participants clearly became familiar with *DUI*, showing that the single calibration session was insufficient. The results for the other experiments calibrate using multiple sessions.

Single Task

The second experiment investigated the efficacy of each task individually for predicting BAL. Table I shows the absolute mean error and Pearson correlation coefficients for the regression models trained using a different number of trials of the same task. As a reminder, study participants performed five trials of each task whenever they used *DUI*. All possible combinations of trials were used for training and testing to ensure that the random selection of an outlier would not skew the results. Doing so yielded more training and testing samples when two or three trials are used since $\binom{5}{2} = \binom{5}{3} = 10$. More training samples can improve accuracy by providing the model with more examples, while more testing samples can worsen accuracy by challenging the model with more outliers. Keeping the same number of samples would have subjected the results to random selection.

Table II lists the features that were selected through automatic feature selection; the descriptions of these features can be found at the project's webpage⁹. Note that the features selected from this experiment were not necessarily the ones that were selected when tasks were combined since feature selection depends on features that complement one another.

Using more trials did not have a significant effect on the results. Although more trials would be expected to better represent the user's performance, using multiple combinations of fewer trials compensated for that effect. The choice reaction task performed well on its own, yielding Pearson correlation coefficients greater than 0.90; the other tasks did not exceed 0.75.

Multiple Tasks

Our final experiment examined how combining features across tasks could improve *DUI*'s results. Table III shows the absolute mean error and Pearson correlation coefficients for all possible combination of tasks. For this analysis, all five trials of each task were included.

Table III. Absolute mean error and Pearson correlation coefficient using different tasks.

Test	Error (%)	<i>r</i>	Test	Error (%)	<i>r</i>	Test	Error (%)	<i>r</i>
T	0.016 ± 0.015	0.68	S+CR	0.005 ± 0.007	0.96	S+ BHR+CR	0.005 ± 0.007	0.96
S	0.015 ± 0.015	0.69	BHR+SR	0.017 ± 0.017	0.58	S+SR+CR	0.005 ± 0.007	0.96
BHR	0.017 ± 0.018	0.57	BHR+CR	0.016 ± 0.014	0.69	BHR+SR+CR	0.015 ± 0.014	0.69
SR	0.017 ± 0.016	0.65	SR+CR	0.016 ± 0.015	0.68	T+S+ BHR+SR	0.015 ± 0.014	0.71
CR	0.001 ± 0.008	0.97	T+S+BHR	0.014 ± 0.013	0.76	T+S+ BHR+CR	0.004 ± 0.007	0.96
T+S	0.014 ± 0.013	0.75	T+S+SR	0.014 ± 0.013	0.76	T+S+ SR+CR	0.005 ± 0.007	0.96
T+BHR	0.014 ± 0.014	0.74	T+S+CR	0.014 ± 0.013	0.75	T+BHR+ SR+CR	0.013 ± 0.012	0.78
T+SR	0.014 ± 0.014	0.73	T+BHR+SR	0.014 ± 0.014	0.74	S+BHR+ SR+CR	0.004 ± 0.007	0.96
T+CR	0.005 ± 0.007	0.96	T+ BHR +CR	0.014 ± 0.014	0.74	T+S+BHR+ SR+CR	0.005 ± 0.007	0.96
S+BHR	0.015 ± 0.014	0.71	T+SR+CR	0.014 ± 0.014	0.73			
S+SR	0.014 ± 0.014	0.74	S+BHR+SR	0.014 ± 0.015	0.72			

T = typing, S = swiping, BHR = balancing+heart rate, SR = simple reaction, CR = choice reaction

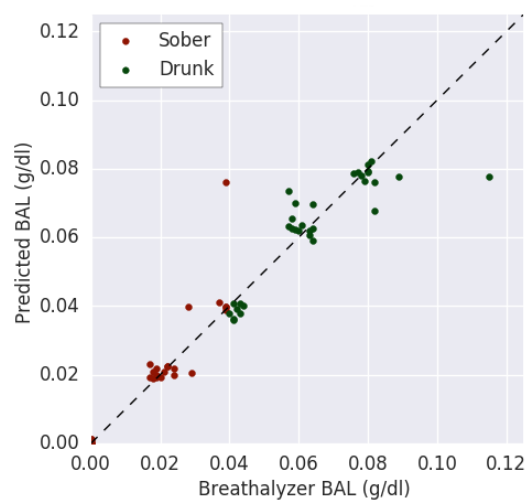


Figure 3. The correlation plot showing *DUI*'s BAL predictions with all 5 trials of all 5 tasks against the breathalyzer output.

Although many combinations of tasks lead to similar performance results as the individual tasks, we found some combinations to be more promising. Using all of the tasks together led to an absolute mean error of $0.005\% \pm 0.007\%$ and a Pearson correlation coefficient of 0.96 (Figure 3). The estimates have a slightly positive bias, overestimating by an average of 0.0005% across all BALs. Not all tasks were needed, though, as we found through the previous experiments that the choice reaction task performed well on its own. Combining that task with the typing or swiping tasks added to *DUI*'s complexity without compromising accuracy, which is important if a user becomes practiced at the choice reaction task on its own.

To frame the results in a different manner, *DUI* can be treated as a classifier that determines whether or not a person should operate a vehicle. The National Advisory Council on Alcohol Abuse and Alcoholism defines a person to be sober at or below 0.04% BAL. With that decision threshold and using all of the tasks, *DUI* leads to a sensitivity of 93.9% and a specificity of 82.3%. Like any other model with a decision boundary, *DUI*'s threshold can be tuned for a specific need; the threshold can be decreased to increase specificity at the cost of lower sensitivity, or the reverse could be done. As a point of comparison for the results, Bae *et al* [3] reported 96.6% accuracy when separating between three categories of inebriation. That being said, their results are not necessarily comparable since they were drawn from self-reports on the number of drinks consumed.

DISCUSSION & LIMITATIONS

Our goal was to develop a smartphone app that assesses a person's psychomotor control and translates that assessment into an estimate of their BAL without the need for additional hardware. To achieve this, we developed and evaluated *DUI*, a combination of five different tasks that challenge a person's motor coordination and cognition. *DUI* uses random forest regression to combine human performance metrics and features derived from smartphone sensor data to estimate a person's BAL. We conducted a week-long study to train and evaluate *DUI*. Through this study, we found that by accounting for user learning through multiple sessions (Figure 2, right), *DUI* is able to estimate BAL with a mean absolute error of $0.005\% \pm 0.007\%$ relative to a breathalyzer. This means that in order for people to use the *DUI*, they would have to perform the tasks periodically while sober. *DUI* will likely not work if a user purposefully botches a baseline. For scenarios when

DUI is used for enforcement (e.g., the car insurance scenario), baselines can be recorded under the supervision of a trusted entity (e.g., an insurance agent). Future studies could be conducted to examine the differences between how drunk people and non-compliant users fail tests, perhaps differently.

Some clinical studies incorporate a placebo into their protocol to ensure that participants are not simply acting drunk after being given alcohol [48]. Our protocol did not have a placebo condition because we did not have enough participants to utilize some in a purely placebo condition. There are two undesirable scenarios from a human performance perspective. The first is when a person is given alcohol but tries to act sober when using *DUI*. Participants were instructed to complete each task to the best of their abilities to avoid these issues. On the other hand, a participant could be given a placebo but act intoxicated. Although this scenario is not included in our dataset, the cost of a false positive from *DUI* is much less than that of a false negative (i.e., mild inconvenience vs. serious danger).

A limitation of *DUI* is that it does not measure BAL directly, but rather the behavioral manifestations of inebriation. Many researchers have noted similar symptoms between inebriation and sleepiness, including impairments to hand-eye coordination and short-term memory [10,49]. In fact, the source of inspiration for *DUI*'s simple reaction and choice reaction tasks, PVT-Touch [27], was designed for sleep loss assessment. We designed our study such that each session was held at roughly the same time of day to reduce fatigue variance across sessions. With even more participant cooperation, our study could be extended to having multiple sessions each day (i.e., every morning, afternoon, and evening for a week). Nevertheless, there are many tasks that both sleep-deprived individuals and inebriated individuals should not perform, so while distinguishing between the two is technically interesting, we believe that it is not crucial for many use cases.

Many would-be participants were eager to join our study, but one or more of the study's high demands (e.g., repeated alcohol consumption, scheduling impositions, and requirement of pregnancy testing) led them not to participate. Unsurprisingly, younger people were far more willing to overlook such burdens for the sake of research. One way to shorten the protocol to a single day would be to replace the breathalyzer ground truth with a clinical one wherein participants' BALs are manipulated through intravenous infusion [17]. Changing BAL intravenously has the additional benefit of higher accuracy and control, although breathalyzers have been used in past clinical alcohol studies [26,49]. Eventually, we decided against this method for fear that participants would not operate the smartphone naturally while in a clinical setting.

FUTURE WORK

We recognize that there is more work to be done to demonstrate the generalizability of *DUI* and its constituent

drunk user interfaces. By conducting our study in a quiet office space with seating, participants were not exposed to additional situational impairments that would occur outdoors or in a noisy bar. Doing the study itself in-the-wild would have been difficult for repeated recruitment and control over alcohol consumption; nevertheless, more work needs to be done to investigate how *DUI* would perform in such scenarios.

DUI, in its current instantiation, is an app that must be explicitly operated by the user to produce an estimate. There were many reasons for this, including control over the stimuli that the participant saw and being able to record fine-grained sensor data from a custom keyboard. Nevertheless, we believe that all of the tasks that we have selected map well to actions that users normally perform on their smartphones. Many people use a swipe password to unlock their smartphone, which was the inspiration for the *DUI* swiping task. Texting is also a common action performed for a variety of tasks, and prior work has shown how text entry accuracy measures can be obtained from text entry "in the wild" [13]. We believe it would be possible to integrate *DUI* more fully into a user's everyday smartphone use. Another way that *DUI* could be consolidated is by combining tasks in a game. For example, a game like Fruit Ninja¹² entails swiping in response to random stimuli, combining both the swiping and reaction tasks.

CONCLUSION

Incidents involving inebriation often occur because they happen before an intervention can take place, highlighting the need for a blood alcohol level (BAL) system more ubiquitous than a breathalyzer. We have introduced drunk user interfaces (DUIs), smartphone tasks that use sensing and human performance metrics to estimate a person's BAL. The combination of different DUIs led us to create the *DUI* app, which utilizes machine learning on human performance and sensor data features to attain a more complete snapshot of the user's current state. To evaluate *DUI*, we conducted a rigorous longitudinal study in which participants used the app at different ground-truth BALs. We trained models that accounted for learning and found that *DUI* was able to estimate BAL with an absolute mean error of $0.005\% \pm 0.007\%$. It is our hope that other researchers will recreate our study or improve upon it with their own tasks so that ubiquitous inebriation assessment might become a reality.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grant IIS-1217627, a Google Faculty Award, and the Mani Charitable Foundation. We also thank Abdullah Ali for his help with the study, as well as the participants for their time.

REFERENCES

1. Johnny Accot and Shumin Zhai. 1997. Beyond Fitts'

¹² <https://fruitninja.com/>

- law: Models for trajectory-based HCI tasks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '97)*, 295–302. <https://doi.org/10.1145/258549.258760>
2. Zachary Arnold, Danielle LaRose, and Emmanuel Agu. 2015. Smartphone Inference of Alcohol Consumption Levels from Gait. In *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, 417–426. <https://doi.org/10.1109/ICHI.2015.59>
 3. Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. 2017. Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2: 1–36. <https://doi.org/10.1145/3090051>
 4. Robert F Borkenstein and H W Smith. 1961. The breathalyzer and its application. *Medicine, Science and the Law* 1: 13. Retrieved March 23, 2017 from <http://msl.sagepub.com/content/2/1/13.short>
 5. Clay Bradley and Gage A Brasher. 2013. Nystagmus evaluation system. Retrieved July 13, 2017 from <https://www.google.com/patents/US9042615>
 6. Robert L Breckenridge and R Scott Berger. 1990. Locus of Control and Perceived Alcohol Ingestion in Performance of a Fine Motor Skill. *Psychological Reports* 66, 1: 179–185. <https://doi.org/10.2466/pr0.1990.66.1.179>
 7. Malcolm Brenner and JR Cash. 1991. Speech analysis as an index of alcohol intoxication—the Exxon Valdez accident. *Aviation, space, and environmental medicine* 62, 9 Pt 1: 893–8. Retrieved August 19, 2017 from <http://www.ncbi.nlm.nih.gov/pubmed/1930083>
 8. Stuart K Card, Thomas P Moran, and Allen Newell. 1983. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates. <https://doi.org/10.1007/s13398-014-0173-7.2>
 9. Federal Trade Commission. 2017. FTC v. Breathometer, Inc. 1–20. Retrieved from www.ftc.gov/system/files/documents/cases/170123breathometer_dkt_4-1_-_stipulated_order.pdf
 10. Drew Dawson and Kathryn Reid. 1997. Fatigue, alcohol and performance impairment. *Nature* 388, 6639: 235–235. <https://doi.org/10.1038/40775>
 11. David F Dinges and John W Powell. 1985. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers* 17, 6: 652–655. <https://doi.org/10.3758/BF03200977>
 12. Edward Elton, D Dumolo, and Colette A Nicolle. 2010. Have I just pressed something? The effects of everyday cold temperatures on dexterity. IN: *Designing Inclusive Interactions: Inclusive Interactions between. In 5th Cambridge Workshop on Universal Access and Assistive Technology*, 22–25. Retrieved July 14, 2017 from http://www.academia.edu/download/42191250/PUB_20571_20The_20effects_20of_20everyday_20cold_20temperatures_20on_20dexterity.pdf
 13. Abigail Evans and Jacob Wobbrock. 2012. Taming Wild Behavior: The Input Observer for Obtaining Text Entry and Mouse Pointing Measures from Everyday Computer Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1947–1956. <https://doi.org/10.1145/2207676.2208338>
 14. Federal Bureau of Investigation. 2014. *Crime in the United States - 2014*. Retrieved July 4, 2017 from <https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-29>
 15. T Flash and N Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of neuroscience* 5, 7: 1688–1703. <https://doi.org/4020415>
 16. Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2687–2696. <https://doi.org/10.1145/2207676.2208662>
 17. S A Grant, K Millar, and G N Kenny. 2000. Blood alcohol concentration and psychomotor effects. *British journal of anaesthesia* 85, 3: 401–406. <https://doi.org/10.1093/bja/85.3.396>
 18. Teng Han, Xiang Xiao, Lanfei Shi, John Canny, and Jingtao Wang. 2015. Balancing Accuracy and Fun: Designing Camera Based Mobile Games for Implicit Heart Rate Monitoring. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 847–856. <https://doi.org/10.1145/2702123.2702502>
 19. Rolla N Harger. 1948. Chemical Tests for Intoxication as an Aid to Traffic and Industrial Safety. *Industrial Medicine and Surgery* 17, 2: 52–56. Retrieved July 4, 2017 from <https://www.cabdirect.org/cabdirect/abstract/19482701549>
 20. Shahin Hashtroudi, Elizabeth S Parker, Lynn E DeLisi, Richard J Wyatt, and Sharon A Mutter. 1984. Intact retention in acute alcohol amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10, 1: 156–163. Retrieved July 6, 2017 from

- <http://psycnet.apa.org/journals/xlm/10/1/156/>
21. I Hindmarch, J S Kerr, and N Sherwood. 1991. The effects of alcohol and other drugs on psychomotor performance and cognitive function. *Alcohol and alcoholism* 26, 1: 71–79. Retrieved July 14, 2017 from <http://www.ncbi.nlm.nih.gov/pubmed/1854375>
 22. Nabil Hossain, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry Kautz. 2016. Inferring Fine-grained Details on User Activities and Home Location from Social Media: Detecting Drinking-While-Tweeting Patterns in Communities. Retrieved July 11, 2017 from <http://arxiv.org/abs/1603.03181>
 23. Amy Jewett, Ruth A Shults, Tanima Banerjee, and Gwen Bergen. 2015. Alcohol-impaired driving among adults—United States, 2012. *Morbidity and Mortality Weekly Report (MMWR)* 64, 30: 814–817. Retrieved July 4, 2017 from <https://www.cdc.gov/Mmwr/preview/mmwrhtml/mm6430a2.htm>
 24. Peter Johnson. 1998. Usability and Mobility; Interactions on the move. In *First workshop on human computer interaction with mobile devices*.
 25. Youngkee Jung, Jinhee Kim, Olumide Awofeso, Huisung Kim, Fred Regnier, and Euiwon Bae. 2015. Smartphone-based colorimetric analysis for detection of saliva alcohol concentration. *Applied Optics* 54, 31: 9183–9. <https://doi.org/10.1364/AO.54.009183>
 26. Hsin-Liu (Cindy) Kao, Bo-Jhang Ho, Allan C Lin, and Hao-Hua Chu. 2012. Phone-based gait analysis to detect alcohol usage. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, 661–662. Retrieved July 11, 2017 from <http://dl.acm.org/citation.cfm?id=2370354>
 27. Matthew Kay, Kyle Rector, Sunny Consolvo, Ben Greenstein, Jacob Wobbrock, Nathaniel Watson, and Julie Kientz. 2013. PVT-Touch: Adapting a Reaction Time Test for Touchscreen Devices. In *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques*, 248–251. <https://doi.org/10.4108/icst.pervasivehealth.2013.252078>
 28. F. Klingholz, R. Penning, and E. Liebhardt. 1988. Recognition of low-level alcohol-intoxication from speech signal. *Journal of the Acoustical Society of America* 84, 3: 929–935. <https://doi.org/10.1121/1.396661>
 29. LF Kozachenko and NN Leonenko. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii* 23, 2: 9–16. Retrieved March 28, 2017 from <http://www.mathnet.ru/eng/ppi797>
 30. Per-Ola Kristensson and Shumin Zhai. 2004. SHARK²: A Large Vocabulary Shorthand Writing System for Pen-Based Computers. *Proceedings of the 17th annual ACM symposium on User interface software and technology - UIST '04* 6, 2: 43–52. <https://doi.org/10.1145/1029632.1029640>
 31. RJ Lawrence. 1988. *The log-normal as event-time distribution*. Dekker, New York, New York, USA.
 32. I Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. 2001. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01*, 9–16. <https://doi.org/10.1145/365024.365028>
 33. I Scott MacKenzie and R William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI '03 extended abstracts on Human factors in computing systems - CHI '03*, 754–755. <https://doi.org/10.1145/765891.765971>
 34. Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N Patel, and Jacob O Wobbrock. 2015. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 2953–2962. <https://doi.org/10.1145/2702123.2702539>
 35. National Highway Traffic Safety Association. 2015. *Traffic Safety Facts. 2014 Data*. <https://doi.org/http://dx.doi.org/10.1016/j.annemergmed.2013.12.004>
 36. Allen F Newell. 1995. *Extra-ordinary human-computer interaction*. Cambridge University Press. Retrieved September 12, 2017 from <http://dl.acm.org/citation.cfm?id=215600>
 37. Mary E Nicholson, MinQi Wang, Collins O Airhihenbuwa, Beverly S Mahoney, and Dolores W Maney. 1992. Predicting alcohol impairment: Perceived intoxication versus BAC. *Alcoholism: Clinical and Experimental Research* 16, 4: 747–750. Retrieved July 6, 2017 from <http://onlinelibrary.wiley.com/doi/10.1111/j.1530-0277.1992.tb00671.x/full>
 38. Suchismita Ray and Marsha E Bates. 2006. Acute alcohol effects on repetition priming and word recognition memory with equivalent memory cues. *Brain and Cognition* 60, 2: 118–127. <https://doi.org/10.1016/j.bandc.2005.07.009>
 39. J M Ryan and L G Howes. 2002. Relations between alcohol consumption, heart rate, and heart rate variability in men. *Heart (British Cardiac Society)* 88, 6: 641–2. Retrieved July 18, 2017 from <http://www.ncbi.nlm.nih.gov/pubmed/12433903>
 40. Andrew Sears and Julie A. Jacko. 2008. *Physical*

- disabilities and computing technologies: An analysis of impairments*. Lawrence Erlbaum Associates, Hillsdale, New Jersey. <https://doi.org/10.1201/9781410615862>
41. Andrew Sears, Min Lin, Julie Jacko, and Yan Xiao. 2003. When computers fade: Pervasive computing and situationally-induced impairments and disabilities. *HCI International 2*, 3: 1298–1302.
 42. R William Soukoreff and I Scott MacKenzie. 2003. Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the conference on Human factors in computing systems - CHI '03*, 113–120. <https://doi.org/10.1145/642611.642632>
 43. H Tianwu, Yoshinori Watanabe, M Asai, K Shimizu, S Takada, and K Mizukoshi. 1995. Effects of alcohol ingestion on vestibular function in postural control. *Acta Oto-Laryngologica* 115, S519: 127–131. <https://doi.org/10.3109/00016489509121886>
 44. Tin Kam Ho. 1995. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition 1*: 278–282. <https://doi.org/10.1109/ICDAR.1995.598994>
 45. Shari Trewin. 2002. An invisible keyguard. In *Proceedings of the fifth international ACM conference on Assistive technologies*, 143–149. <https://doi.org/10.1145/638249.638275>
 46. Gregg C Vanderheiden. 1997. Anywhere, anytime (+anyone) access to the next-generation WWW. *Computer Networks and ISDN Systems* 29, 8: 1439–1446. [https://doi.org/10.1016/S0169-7552\(97\)00067-6](https://doi.org/10.1016/S0169-7552(97)00067-6)
 47. Emma R Weaver, Danielle R Horyniak, Rebecca Jenkinson, Paul Dietze, and Megan Sc Lim. 2013. “Let’s get Wasted!” and Other Apps: Characteristics, Acceptability, and Use of Alcohol-Related Smartphone Applications. *JMIR mHealth and uHealth* 1, 1: e9. <https://doi.org/10.2196/mhealth.2709>
 48. JM Weiler, JR Bloomfield, GG Woodworth, AR Grant, TA Layton, TL Brown, DR McKenzie, TW Baker, and GS Watson. 2000. Effect of fexofenadine, diphenhydramine, and alcohol on driving performance. *Annals of Internal Medicine* 132, 5: 354–363. Retrieved August 21, 2017 from <http://annals.org/aim/article/713291/effects-fexo-fenadine-diphenhydramine-alcohol-driving-performance-randomized-placebo-controlled-trial>
 49. A M Williamson. 2000. Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication. *Occupational and Environmental Medicine* 57, 10: 649–655. <https://doi.org/10.1136/oem.57.10.649>
 50. Jacob O Wobbrock. 2006. The Future of Mobile Device Research in HCI. In *Proc. CHI '06 Workshops*, 131–134. Retrieved January 6, 2018 from <http://faculty.washington.edu/wobbrock/pubs/chi-06.05.pdf>
 51. Jacob O Wobbrock, Krzysztof Z Gajos, Shaun K Kane, and Gregg C Vanderheiden. Ability-Based Design. *Communications of the ACM*.
 52. Jacob O Wobbrock and Brad A Myers. 2006. Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction* 13, 4: 458–489. <https://doi.org/10.1145/1188816.1188819>