Evaluation of Different Modalities for Self Measuring Impulsivity

Jason Waterman, PhD[1], George Whiteside[1], Hongyi Wen[2], Dylan Horowitz[1], JP Pollak, PhD[2], Deborah Estrin, PhD[2]
[1]Vassar College, Poughkeepsie, NY, USA; [2]Cornell Tech, New York, NY, USA

Abstract

Mobile devices are becoming an increasingly integral part of modern life. As the popularity of smartphones, smartwatches, and voice assistants continues to rise, more people are using them to track health and medical statistics. We enable easy and consistent data collection by extending ResearchStack, a framework for developing mHealth applications, to support Android Wear devices. We also evaluate user acceptance of different modalities by collecting feedback from 30 users using a smartphone, a smartwatch, and a voice assistant version of the Balloon Analogue Risk Task (BART), a computerized measure of risk taking behavior. The smartphone was the preferred platform, with users stating ease of use and familiarity with the platform as reasons for their preference. However, almost 17 percent of users listed the smartwatch platform as their favorite, stating that it was also easy to use and quicker than the other versions. The voice assistant application was rated the lowest, but it can be a viable option for others who are not able to use a smartphone, expanding the participation base for data collection. Users also provided feedback on how to improve these new modalities, which will be incorporated into the next version of these applications.

Introduction

Poor self-control and impulsivity are underlying symptoms of numerous mental health problems such as obesity, substance abuse, ADHD, gambling, binge eating, bipolar disorder, borderline personality disorder, and suicidal behaviors[1]. This makes it one of the most important personal and public health intervention targets. The classic marshmallow experiment[2] determined that the inability to delay gratification in childhood was predictive of lower SAT scores and higher BMI in adulthood. However, assessments of impulsivity and poor self-regulation are rarely included in routine medical care because of time and financial constraints.

The smartphone has changed our ability to assess and intervene with individuals remotely, providing an avenue for ambulatory diagnostic testing and just-in-time adaptive interventions that can be accessed by billions of people. Newer methods of assessment using smartphone mHealth platforms, including ResearchKit and ResearchStack[3], provide the opportunity for powerful assessments of impulsivity beyond simple self-report.

Before we can conduct large-scale population based studies in impulsivity, we need to validate mobile impulsivity assessments in real-world settings. In a broader collaboration with Northwell Health, we aim to do just this with collaboration from researchers, clinicians, and patients. Our goal is to develop and validate a remote assessment task of impulsivity called the Digital Marshmallow Test (DMT) using both Apple and Android mobile applications for widespread dissemination to researchers, clinicians and the general public to assess impulsivity.

While our primary target is smartphones, the use of wearable devices and smartwatches is growing[4]. In addition, the popularity of voice assistants such as Apple's Siri, Amazon's Echo devices, and Google's voice assistant means there are new modalities for collecting data. Having multiple ways of interacting with the user to collect data can help to improve compliance with surveys and tests. This can be increasingly important if data needs to collected frequently over a long period of time, where survey fatigue can be an issue. A framework to guide development on these new platforms and an evaluation of acceptance of these new modalities is needed and is the focus of this paper.

While there has been research on using smartwatches for both passive data collection[5] and active data collection[6], our work is one of the first examples of a non-smartphone mHealth application framework and evaluation of user acceptance of these applications on different modalities. Our work has two main contributions. First, we extend ResearchStack to support Android Wear devices. ResearchStack is a community based, open-source platform for building mHealth research study applications on Android. Our extensions to ResearchStack are freely available and allow anyone to run ResearchStack applications on Android smartwatches with minimal modifications.

Second, we evaluate how receptive users are to taking tests with two new modalities, smartwatches and voice assistants.

To this end, we developed three versions of the Balloon Analogue Risk Task (BART)[7], a computerized measure of risk taking behavior. One version runs on Android smartphones, another version runs on Android Wear smartwatches, and the third version runs on top of Google Assistant, a voice activated assistant similar to Apple's Siri. Google Assistant runs on Android and IOS smartphones, as well as stand alone hardware devices.

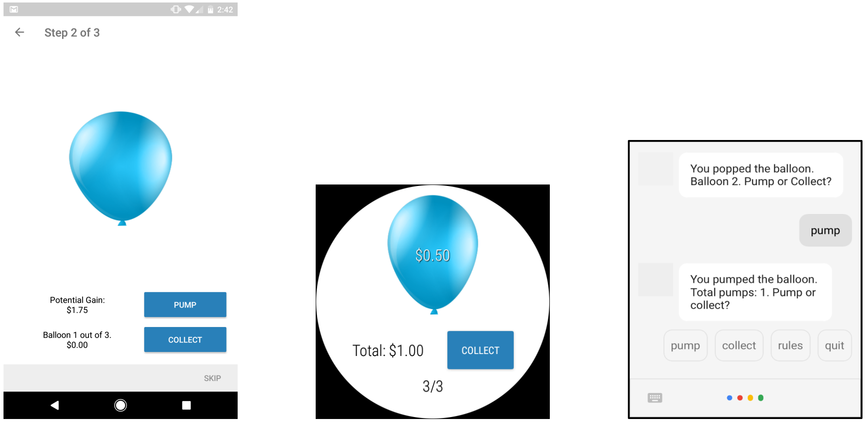
We collected data from 30 users who used all three versions of the test. Users were the most comfortable with and favored the smartphone version of the test. However, 17% (5 out of 30) users preferred taking the test on the smartwatch, citing ease of use and speed of taking the test as their reasons. We also received valuable feedback on how to improve the interfaces of the smartwatch and voice assistant applications.

Methods

The BART test models real-world risk behavior by balancing the potential for reward versus loss. The task is presented as a number of trials. Each trial gives the user a chance to earn money by inflating a virtual balloon. A turn in a trial presents the user with two choices. They can pump the balloon to inflate it in an attempt to earn money or they can stop inflating the balloon and collect the money they have earned in the trial so far. With each inflation attempt, there is a chance that the balloon will pop, with the chance of popping increasing with each turn, up to some maximum. If the balloon pops, the user loses any money they had accumulated for that trial.

We used ResearchStack to develop the smartphone version of the test. ResearchStack is similar to Apple's ResearchKit, with an overarching goal of making it easier to port ResearchKit applications to Android. Applications developed with ResearchStack are built using JSON and HTML files, making it possible to build basic applications without having extensive knowledge of the Android platform. It also supports secure connectivity to multiple data collection backends.

To support the smartwatch version of the BART test, we extended ResearchStack to support Android Wear devices. This allowed us to reuse much of the existing smartphone code and allows us to port any other ResearchStack application to the smartwatch with minimal effort. It also means that features supported in the mobile version of the application, such as secure backend data storage, are also supported on the smartwatch as well. Most of the effort in porting a ResearchStack application to an Android Wear device is adjusting to the much smaller screen on the watch interface. We tested the application on two Android Wear devices, the LG Watch R, and the LG Watch Sport, though the application should run without modification on any Android Wear device.



You popped the balloon. Balloon 2. Pump or Collect?

pump

You pumped the balloon. Total pumps: 1. Pump or collect?

(a) Android Smartphone **(b)** Android Wear Watch **(c)** Google Assistant

Figure 1: Screenshots of BART Test

The Google Assistant version of the BART test was developed using Actions on Google and the Actions SDK, which allows the user to take the test through a conversational interface. The application is written in Javascript using the Firebase platform, a mobile and web application development framework. It is currently being hosted using Google Cloud services, but it also possible for the application to be hosted on a local server. The application can be run on any device that supports Google Assistant, including Android and IOS smartphones, and the Google Home device.

We tested the application on both Android and IOS devices. The primary mode of interacting with this interface is via voice, though if you are interacting with the voice assistant on a smartphone, you can say your response or press one of four action buttons: pump, collect, rules, or quit (which also serves as a reminder as to what responses are valid). On the smartphone, the results of each round are both spoken out loud and shown in a text bubble, similar to a text chat application. If you were performing the test on a device such as Google Home, the interaction would be done entirely through voice. User evaluations were done using the smartphone interface to Google Assistant on both Android and IOS devices. Screenshots on all three platforms are shown in Figure 1.

Our goal was to keep the tests as similar as possible across the three different platforms, subject to the constraints of each platform. The smartwatch is a much smaller form factor with limited screen "real estate". This led to some differences between the smartphone and smartwatch applications, some of which can be seen in the screenshots. The first difference is that to save space on the smartwatch screen, we removed the "pump" button. Instead, the users tap on the balloon to inflate it. Also, the amount of money collected in a round (called "Potential Gain" in the smartphone version) was moved to inside the balloon itself, with no label describing it. In the smartphone application, when the user presses the pump button, there is an animation inflating the balloon. There is also an animation when the balloon pops. We debated adding the inflation and popping animations to the smartwatch version, but decided against it given the constraints of the screen size.

There is a much bigger difference between the smartphone/smartwatch version of the application and the voice assistant version. By design, interaction with the application is done primarily through voice, however as mentioned above, you can press one of the action buttons in the chat window on the smartphone instead of speaking. Some users did interact with the system in this manner, however we did not record how many responses were done via voice versus pressing the action buttons.

To evaluate these prototypes, we had users try all three versions of these applications. For all users, we verbally explained how the test works and had them first use the smartphone version of the application, where they conducted three rounds of the test, inflating three balloons. We started with the smartphone because all participants were familiar with the form factor, so it was easiest to explain how the test works on a smartphone. We expect that a common use case will be applications running on the phone supported by a smartwatch or voice assistant, so we are not concerned with order effects with respect to the smartphone in this preliminary work.

For their second test, we randomly picked either the smartwatch or the voice assistant, where they also completed three rounds of the test. The user then finished with the version of the application they had not yet tried, again inflating three balloons. After they completed all three versions of the tests, we asked them a series of short answer questions and a number of statements which they could agree or disagree with using a one to five Likert scale. We asked the same set of statements for each version of the application: smartphone, watch, and voice assistant. The survey questions we asked are shown in Table 1 and the statements we asked them to rate are shown in Table 3.

Q1. What interface do you think would be the best for using every day? List them in order of your preference, with one (1) being your most preferred, and three (3) being your least preferred.
Q2. What did you like about the test you chose as number one (the best)?
Q3. What didn't you like about the test you chose as number three (the worst)?
Q4. Thinking about how you might use the watch version of the test on a daily basis, how would you improve the user interface for the test?
Q5. Thinking about how you might use the voice assistant application on a daily basis, how would you improve the user interface for the test?

Table 1: Short Answer Questions

Results

We evaluated the three prototypes described above with 30 students and employees at Vassar college. Of the 29 users who reported their gender, 19 (65.52%) were female and 10 (34.48%) were male. All users own a smartphone and are familiar with its operation.

The first question (Q1) asked the users to rank their preference among the three modalities. These results are shown in Table 2. Most users preferred the smartphone, with 83.33% users rating it their top choice and all users listing is as either their first or second choice. The most frequent reason given why users preferred the smartphone application was familiarity with the platform, with one user stating "I'm so used to using my phone already. It was intuitive. I already knew what to do and it was familiar."

Application	Ranked First	Ranked Second	Ranked Third
Smartphone	83.33%	16.67%	0.00%
Smartwatch	16.67%	53.33%	30.00%
Voice Assistant	0.00%	30.00%	70.00%

Table 2: User Rankings

Visual appeal was another reason why users preferred the smartphone. The application gave users an animation when the user pressed "pump" to inflate the balloon. The balloon got bigger on the screen with every press. This gave them a visual indication that they had pressed the pump button, as well as adding a bit of suspense before seeing if the balloon popped. Due to the small screen size, this feature was not added to the smartwatch version. It was also not part of the voice application.

There were some users (16.67%) who preferred the smartwatch over the smartphone. In addition, 70% of the users listed the smartwatch as either their first or second choice. Those who preferred the smartwatch stated ease of use and convenience as their top reasons. One user said "It was really easy for people to have. Everything is right there." while another user said "It was a lot quicker," and "It felt appropriate for the screen size of the watch."

Of the users who ranked the smartwatch last, some had more general concerns about the platform itself, with one user saying "I am concerned about battery life. I can't see myself using it." and another saying "I don't wear a watch." The other issue had to do with size of the device, with several users saying it was small and text on the watch was hard to read. They also missed the animations available on the smartphone application.

The voice assistant version of the application was the least favorite application. The biggest issue people had was the length of time it takes to take the test compared to the other versions. It takes awhile for the application to read back the results of each turn, in addition to time it takes to say your choice and have the application recognize what you said. A poor network connection can also slow each turn down, as processing is done on a remote server and not locally on the device.

Several users mentioned that they are concerned about privacy, with one user saying "If I take this test every day, surrounded by people, then they will know what I am doing. From a privacy standpoint, I would like the test to be more discrete." However, we expect that the most common use case for the voice assistant would be in the home or private office where this is less of a concern. Several people also had issues with the speech recognition, with users saying "The voice recognition wasn't really that on. It misheard me a couple of times. The touch screens were more accurate." and "It was slow and sometimes it wouldn't register my voice correctly." Again, if used in the home or private office, the speech recognition should work much better.

To get a quantitative measure of acceptance of these new modalities, we asked a series of statements, which users could strongly agree (by giving a five) or strongly disagree (by giving a one). The results are shown in Table 3. The first three columns show the average score for the smartphone, smartwatch, and voice assistant respectively. The fourth column shows the difference in ratings between the smartwatch and smartphone, and the fifth column shows the difference in ratings between the voice assistant and smartphone. Overall, the results are consistent with the rankings and short answer responses the users gave. That is, they liked the smartphone the best and the voice assistant the least, with the smartwatch in the middle, but close to the smartphone in overall acceptance.

Looking at the averages of all the statements, users were in between "agree" and "strongly agree" for the smartphone, with an average of 4.47. The average for the smartwatch was 3.90, putting it very close to "agree" with most statements. The users were more neutral with the voice assistant, with the average being 2.83, which is just below "neither agree or disagree".

Statement	Mobile	Watch	Voice	Watch vs. Mobile	Voice vs. Mobile
S1. The application is easy to use.	4.87	4.10	3.67	-0.77	-1.20
S2. I would likely use the application daily during down time.	4.17	3.57	2.20	-0.60	-1.97
S3. It is enjoyable to take the test on this application.	4.17	3.57	2.60	-0.60	-1.57
S4. I would likely take the test using the application soon after receiving a reminder to take the test.	4.27	3.93	2.63	-0.33	-1.63
S5. I learned to use the application quickly.	4.90	4.53	3.83	-0.37	-1.07
S6. I am satisfied taking the test with this application.	4.67	3.93	2.57	-0.73	-2.10
S7. I application helps motivate me to finish the test.	4.27	3.63	2.33	-0.63	-1.93
Grand Total	4.47	3.90	2.83	-0.58	-1.64

Table 3: Survey Results

Comparing the smartwatch to the smartphone, users favored the smartphone by 0.58. Given how familiar people are with smartphones and how new smartwatches are, this is encouraging. By implementing the smartwatch application improvements suggested in the next section, it is quite possible to narrow this gap with an improved version of the smartwatch application.

For the smartwatch, statement S4, which measures how likely a user would take the test upon receiving a notification, had the least difference from the smartphone, with a rating of 0.33 less than the smartphone. This may be an indication that some users appreciate the speed and convenience of taking a quick test on their wrist without having to pull out their phone. More study is needed in this area.

The two statements that had the biggest difference between the smartwatch and smartphone were S1 and S6, which asks about ease of use and satisfaction. In these two statements, users still agreed with these statements (4.10 and 3.93), just not as much as they did with the smartphone. This difference is most likely due to the difficulty of reading instructions on the smartwatch and the lack of animations on the application. We discuss how we plan to improve these areas in the next section.

Differences between the smartphone and voice assistant were more pronounced, with the average difference being 1.64. The statement with the biggest difference was S6, which asks about application satisfaction. Users rated this statement 2.10, which is a strong signal that users are not as happy with this platform compared to the smartphone. This may also suggest that this test is not a good match for this particular platform.

Design Implications

We asked the users specifically how to improve the smartwatch and voice assistant applications. We summarize their suggestions here and discuss how we plan to implement these improvements in the next version of these applications.

For the smartwatch, most users did not like the long block of instruction text that first pops up to explain how to take the test. The text was hard to read and it required users to scroll through the text, which some users did not know how to do because this was the first time they used an Android Wear smartwatch. They suggested we find a way to replace this “wall of text” with something else.

Based on the feedback we collected, it is clear that any smartwatch application that needs lots of text to operate is not a good match for this platform. For the next version of the BART test, we plan on replacing the documentation with a tutorial that will show up the first time the user performs the test. This tutorial will have a minimum amount of text per screen and will walk them through the basics of the game (e.g., how to pump the balloon and collect money).

The second smartwatch improvement that several people mentioned was to add balloon animations to give the user feedback they pressed the button to inflate the balloon. Users missed this feature on the watch, along with the animation of the balloon popping. The future version of this test will add the inflation and popping animations. To get around the lack of a larger screen size, we plan on showing an animation of the balloon inflating, and if successful, have the screen display some indication of success (perhaps by showing a dollar sign or displaying a message such as “success” or “winner”). After the message is shown, the balloon will go back to its previous size. If we did not do this, the balloon would have to start out very small (and be hard to press) if it kept getting bigger with every pump.

We also asked users how we could improve the voice assistant application. The top suggestion is to improve the speed of taking the test. Most users felt the test took too long compared to the smartphone and smartwatch versions of the application. Also, users missed the visualizations that are available in the smartphone version of the application. Users also didn’t like the amount of time it took for the instructions to be read to them. One user wished for better feedback when you said something incorrect, and several wished for the test to be less repetitive. Another user summed up things by saying “It doesn’t feel like a conversation.”

Given the constraints in the platform, we are a bit more limited in the improvements we can make, but for our next version of the application we will try to streamline the flow of each turn in the test. It is possible to send back graphical results (when using voice assistant on a smartphone) as part of the voice response, but animations are not allowed at this time. We can attempt to make it more conversational by accepting more phrases instead of just “pump” or “collect”, but it may simply be that this modality is not the best fit for this type of test.

Conclusion

The user feedback we received shows a good degree of acceptance for the smartwatch platform. We received valuable feedback about how to improve this test in particular along with some general guidelines for designing applications using this new modality. There was less acceptance of the voice assistant. However, it may be appropriate in situations where users are unable to use a smartphone, for example, users with poor or limited vision. It may also be appropriate where application development resources are limited, as Google’s voice assistant applications can work on both Android and IOS devices. Ultimately though, voice applications may be a better fit in answering health surveys or questionnaires. Those applications may benefit from a more conversational form this modality provides. This avenue would be a useful area to study in the future.

Smartwatches and voice assistants can never fully replace all the things a smartphone can do. However, there is a good reason to think they can work as an augmentation or companion to smartwatches. For simple tests and surveys, having multiple different ways to respond might help with compliance. Users reported they were almost as likely to use the smartwatch application to take the test as the smartphone. As smartwatches and voice assistants become more ubiquitous, interacting with these devices for tests and health surveys will become more commonplace. With the right interface design we can hope to get similar acceptance rates when compared to smartphones.

References

1. de Ridder DT, Lensvelt-Mulders G, Finkenauer C, Stok FM, Baumeister RF. Taking Stock of Self-Control: A Meta-Analysis of How Trait Self-Control Relates to a Wide Range of Behaviors. Personality and Social Psychology Review. 2011 Aug;16(1)76-99. Available from: https://doi.org/10.1177/1088868311418749
2. Mischel HN, Mischel W. The development of children’s knowledge of self-control strategies. Motivation, intention, and volition 1987 (pp. 321-336). Springer, Berlin, Heidelberg.
3. http://researchstack.org/
4. Rawassizadeh R, Price BA, Petre M. Wearables: Has the Age of Smartwatches Finally Arrived? Communications of the ACM. Volume 51 Issue 1, January 2017 Pages 45-47. Available from: https://doi.org/10.1145/2629633
5. Sharma V, Mankodiya K, De La Torre F, Zhang A, Ryan N, Ton TG, et al. SPARK: Personalized Parkinson Disease Interventions through Synergy Between a Smartphone and a Smartwatch. In International Conference of Design, User Experience, and Usability 2014 Jun 22 (pp. 103-114). Springer, Cham.
6. Arsand E, Muzny M, Bradway M, Muzik J, Hartvigsen G. Performance of the first combined smartwatch and smartphone diabetes diary application study. Journal of diabetes science and technology. 2015 Jan 14;9(3):556-63.
7. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, Strong DR, Brown RA. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). Journal of Experimental Psychology: Applied. 2002 Jun;8(2)75-84. PubMed ID 12075692