MTR190511
MITRE TECHNICAL REPORT

MITRE

Usability of Biometric Authentication Methods for Citizens with Disabilities

Project No.: 1919MG19-AA

Bedford, MA

This work was presented at the 9th Annual Internal Revenue Service-Tax Policy Center (IRS-TPC) Joint Research Conference on Tax Administration in June 2019. It is slated to be published in the 2019 IRS Research Bulletin.

Author(s): Ronna N. ten Brink
Rebecca I. Scollan

September 2019

APPROVED FOR PUBLIC RELEASE

Abstract

Currently, one out of five adults in the United States has a disability. As the population ages, the number of adults with disabilities will swell. As critical government services move online, the need for accessibility grows. However, poor accessibility and usability in authentication methods can form a barrier to the use of important websites, such as tax and benefit services. Given current commercial trends, biometric authentication methods will be used more widely to ensure secure access to such services. There is currently a dearth of research into both accessibility and usability of authentication modalities, including biometric methods. Thus, we investigated the usability of biometric authentication schemes for users with and without disabilities (vision or hearing). We comparatively evaluated three biometric authentication schemes (fingerprint, eye, and palm recognition) and one non-biometric authentication scheme (PIN) on effectiveness, efficiency, and perceived usability. Traditional and biometric schemes showed some usability differences. Biometric schemes' usability often differed based on whether the interaction required dynamic device positioning (placing and holding the device in relation to specific points on the user's frame). Biometrics that required dynamic device positioning (ex. palm) had lower usability for participants with limited or no vision. We therefore put forth dynamic device positioning as a new consideration for usability evaluations of biometrics.

Keywords: disability, accessibility, usability, human computer interaction, human factors engineering, authentication, ID proofing, biometrics, information interfaces, systems modernization, information services

This page intentionally left blank.

Acknowledgments

The authors wish to thank Katja A. Sednew, Michelle R. Schumaker, Melanie Shere, Jared M. Batterman, Kristen M. Klein, and Erika L. Darling for their support in this work.

Table of Contents

List of Figures

List of Tables

This page intentionally left blank.

1 Introduction

Today, 27.2 percent of people living in the United States experience a disability, which is defined as a functional limitation that affects one or more major life activities [1] [2]. Approximately 17.6 percent of those who report a disability describe it as a severe disability. As we age, our likelihood of having a disability increases. The current percentage of the population with a disability is assumed to be a low assessment because census data is collected from households, which leaves out those who live in nursing or assisted living facilities, the large majority of whom have a disability [1]. Generally, people are living longer both in the United States and across the world. From 2015 to 2030, it is estimated that the elderly population will grow from 9-12 percent of the global population [3]. As our population ages, the number of adults with a disability will grow as well.

Single-factor authentication with a username and password has long been known to be vulnerable to both social engineering and brute-force attacks, as well as a usability challenge due to contradictory advice and the cognitive burden of managing many complex passwords [4]. A smartphone allows for greater use of more convenient methods of authentication. Smartphone ownership increased 42 percent from 2011 to 2018 [5], and 77 percent of U.S. adults now own smartphones. Widespread smartphone use has made two-factor and multi-factor authentication more prevalent [4]. Two-factor authentication combines information that someone knows, such as a password, with something that they own, like a smartphone. Multi-factor authentication provides an additional security factor that is unique to the subject, typically a physical or behavioral biometric [6], and can be inputted on a smartphone. The use of multi-factor authentication will likely continue to grow within the U.S. as e-commerce adopts recommendations from the National Cybersecurity Center of Excellence (NCCoE) at the National Institute of Standards and Technology (NIST) to use multi-factor authentication on online accounts in order to reduce the growing problem of online purchase fraud [7].

We investigated the usability of biometric authentication schemes for users with and without disabilities. We comparatively evaluated three biometric authentication schemes (fingerprint, eye, and palm recognition) and one non-biometric authentication scheme (PIN) on effectiveness, efficiency, and perceived usability. This research contributes to the development of a standardized methodology to evaluate the usability and accessibility of authentication technologies intended for use with public government services. Our initial focus is a comparative usability study on PIN and biometric authentications; these methods were chosen for their current popularity and future usage potential. We worked with the HYPR Corporation, who provided a FIDO Universal Authentication Framework (UAF) client for Android and iOS. HYPR offers an inherently multi-factor, decentralized authentication solution designed to eliminate passwords and shared secrets as a means for authenticating users more securely with an easier user experience. Using a working demonstration application provided by HYPR, we conducted our usability study on a range of popular biometric schemes.

We chose to work with two large populations of adults with disabilities: those who are low vision or blind, and those who are deaf or hard of hearing. In Taylor's Census Report [1] on estimates of disability prevalence based on the Social Security Administration (SSA) Supplement to the 2014 Survey of Income and Program Participation, 12.3 million U.S. adults over the age of 18 had serious difficulty seeing, of which 1.6 million are legally blind. 17.1

million adults reported a serious hearing difficulty, of whom 3.4 million who were deaf. We selected the two populations due to their large size as well as practical and logistical considerations due to time and the research team's familiarity with both populations and assistive technologies used. Ultimately, 30 individuals were recruited; 10 participants who had hearing loss, 10 who were low vision or legally blind, and 10 who reported no disability.

This research contributes to a better understanding of the user experience of smartphone-based biometric authenticators and the eventual increased usability and accessibility of online government services, leading to higher adoption and wider access to these services. Our results can also be generalized to any secured web services, e.g., banking and healthcare services.

1.1 Background

A growing community of people living with one or more disabilities creates a challenge to federal agencies looking to digitize more personalized services. Government services receive low customer satisfaction scores [8] compared to industry for their websites and customer service. Despite this challenge, there is recognition that services must be modernized, personalized, and moved to online channels to reduce costs and improve citizen services [9, 10, 11]. For example, the President's Management Agenda (PMA) CAP goal 4 aims to "provide a modern, streamlined, and responsive customer experience across Government, comparable to leading private-sector organizations" and "improv[e] the experience citizens and businesses have with Federal services whether online, in-person, or via phone" [12]. The 21st Century Integrated Digital Experience Act (21st Century IDEA), passed in December 2018, sets a "minimum accessibility, searchability and security standards for all new and existing government websites, and require agencies to adopt web analytics tools to constantly improve sites' functionality. Organizations would also need to make all sites mobile-friendly and comply with website standards set by the General Service Administration" [13]. As federal agencies work towards meeting this challenge, providing services that are both usable and secure is tantamount, and the design of identity proofing and authentication addresses a critical first user touchpoint.

Federal agencies' digital services face unique usability challenges. Registration for an online service with a federal agency might be the first interaction a citizen has ever had with that agency. Such services might be used only once in a lifetime or be accessed very infrequently. The audience for these services is often diverse, spanning all ages, incomes, geographies and abilities. Additionally, key services may include access to one's own personally identifying information, implying significant risk to both the institution and users. But moving such services online offers federal agencies a great benefit of increased citizen satisfaction and reduced costs. Federal agencies typically have no competition and are the only place to contact when citizens encounter questions or problems. An average business cost for a call-center call is $5.50 versus online services' cost of $0.10 serving those who find answers or resolutions online [14]. Some agencies face even higher call-center costs – the average call to the IRS costs $41 [15]. Agencies must comply with Section 508 and new IDEA Act mandates on accessibility when designing and implementing digital services. Section 508 [16], an amendment made to the 1973 Rehabilitation Act in 1998, mandates that federal agencies provide accessible electronic content and technologies. The 21st Century IDEA Act gives agencies a 180-day deadline to comply with Section 508 for all hardware, software and documentation [17].

The 2017 update to the NIST Special Publication (SP) 800-63-31 Digital Identity Guidelines, which includes SP 800-63B Authentication and Lifecycle Management, now requires two-factor authentication: either a multi-factor authenticator or a combination of two single-factor authenticators to achieve Authentication Assurance Level 2 [18]. Many federal agencies' online services meet the criteria for Authentication Assurance Level 2. Biometrics are growing in popularity [19] and may be used in a multi-factor authentication design. NIST defines biometrics as both physical and behavioral characteristics, and includes them as a factor, as long as they are part of multi-factor authentication with a physical authenticator (with a device like a smartphone meeting security requirements of proving "something you have," and the biometric, "something you are").

But are biometrics captured by smartphones usable and accessible to all citizens? While widespread smartphone ownership has made biometrics more available [20], there is little evidence to support that mobile-based biometrics will be accessible to or usable for all Americans [21]. Federal agencies seeking to leverage multi-factor authentication need more data-driven insight into the usability and accessibility of these technologies. NIST recommends observational usability testing for assessing multi-factor authentication and biometrics [22]. However empirical comparison of authentication schemes, including biometrics, is not common. The historical lack of a standard usability metric in authentication research contributes to difficulty comparing usability across schemes [23, 24].

1.2 Related Work

The body of literature on both accessibility and usability of authentication schemes is growing but currently remains small relative to the amount of existing work on authentication usability. It has been noted that accessibility has not received adequate attention in biometric system design [25]. This section discusses prior research that is relevant to our focus. We build on existing literature by evaluating the relative usability of authentication schemes for users with and without disabilities along effectiveness, efficiency, and perceived usability metrics.

Ruoti, Roberts, & Seamons [23] emphasized the importance of empirical research when evaluating authentication schemes. The authors explored seven web-based authentication systems to determine what was most usable and what features participants valued most. They compared the usability of authentication techniques like email-based and QR-based systems in a tournament-style "championship," measuring usability using the System Usability Scale (SUS) questionnaire. The authors recommend using SUS as a standard metric for future evaluation of new authentication systems. Our usability comparison employed the UMUX-LITE perceived usability scale, which has been shown to be an acceptable alternative to SUS [26, 27, 28, 29].

In 2012, Trewin, Swart, Koved, Martino, Singh, and Ben-David [24] conducted a lab study of three biometric schemes and a password scheme. They observed six experimental conditions: PIN; voice; face; gesture; face and voice together; and gesture and voice together. The authors collected biometric performance, interaction time, error rates, memory recall success rate, and self-reported reactions using modified SUS. They observed that despite the fact that the voice biometric condition resulted in the least errors and performance time, participants found it lacking in usability, gracing it with a SUS score of "D." The authors proposed this might have been due to the volume required for participants to provide an acceptable voice sample. We too

used time and a SUS variant as usability metrics. Trewin, et al. emphasize the importance of providing appropriate feedback to users on achieving proper facial biometric alignment, to reduce errors and reduce the time for biometric recognition to occur; a similar conclusion is discussed later in this paper.

In 2018, Blanco-Gonzalo, Lunerti, Sanchez-Reillo, & Guest [21] performed a comparative study on the usability and accessibility of mobile biometrics. They investigated the accessibility of voice, face, fingerprint, PIN, and pattern schemes and compared the usability and accessibility of the more traditional authentication method of PIN to biometric authentication techniques. They also included multiple groups of participants with disabilities (upper body, lower body, visual, and cognitive) and a control group of participants with no disabilities. The authors measured task time, satisfaction, and errors. Similarly, we compared traditional and biometric authentication schemes (PIN, fingerprint, eye, palm), and worked with participants with low vision, blind participants, and participants with no disabilities. We measured similar metrics, although our error data was ultimately not usable for analysis. Unlike Blanco-Gonzalo, et al., we included participants with hearing loss, and did not examine pattern authentication. Our study also required participants to perform the tasks on their own devices so as to gain a better understanding of usability in the context of personalized assistive technologies. This study's results echo Blanco-Gonzalo, et al.'s findings for their control group. Our participants who had vision loss preferred biometrics that did not require positioning (Section 2.2 explains positioning biometrics), similar to Blanco-Gonzalo, et al.'s finding that participants with visual disabilities disliked the face biometric.

2 Study Design

30 diverse participants were recruited, including participants with limited or no vision and with hearing loss. We evaluated and compared six authentication modalities: PIN, palm, eye, face, face and voice, and fingerprint. Two modalities, face and face and voice, were removed from analysis because technical set-up difficulties caused too small of a sample size for these schemes. The International Organization for Standardization (ISO)'s definition of usability was employed: the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [30]. We did not request nor were we provided performance data on the biometrics from the prototype application partner. This research focuses only on usability measures.

2.1 Mobile Application Prototype

HYPR provided a real, working system and hosting resources to support a prototype of several modes of biometric authentication on iOS and Android devices. HYPR uses Fast Identity Online (FIDO) and a "decentralized" authentication concept. The user's device application allowed six authentication schemes for "unlocking" a private key. Biometric privacy precautions are discussed in Section 3.5.

On installing the application on an iOS or Android, participants were prompted to enroll their biometrics within the application. PIN, palm, face and voice, fingerprint, and eye were available within the iOS app. Android applications contained PIN, palm, face, fingerprint, and eye. Enrollment included text and illustrations on how to position the phone to capture the biometric

best. Some also contained text or visual cues during the enrollment process, such as text suggesting where to move a phone, or green bars that lit up when the user's eyes were properly positioned within a bounding box on the screen or their palm within a red circle on the screen. After enrolling one or more authenticators, a dashboard was enabled for participants, showing icons representing each authenticator enrolled. On selecting an icon on the dashboard, the participant was able to attempt a login using the corresponding scheme.

2.2 Hypotheses

PIN is considered a baseline similar to the most common authenticator, passwords, where users enter characters or numerals through a keyboard input. From observations in pilot sessions and informal interviews with people with disabilities, we created a dynamic positioning versus non-dynamic-positioning categorization for biometrics. We define dynamic positioning as interactions where users are required to position and hold their device in relation to a specific point on their frame (dynamic positioning actions). We define non-dynamic-positioning as interactions where users are not required to position or hold their device in relation to a specific point on their frame (static positioning actions). We predicted three patterns would emerge in our study:

H1: User performance (efficiency and effectiveness) will be different between PIN and biometric schemes;

H2: User performance will be different between positioning biometrics (eye, palm) and non-positioning biometrics (fingerprint); and

H3: For the user group with vision loss, user performance will be better with non-positioning biometrics than with positioning biometrics.

2.3 Task Performance Metrics

2.3.1 Efficiency and Effectiveness

Efficiency was operationalized as response time on an authentication task. Response time was captured by measuring elapsed time on task from the start and end of screen prompt page loads. The mobile authentication application was reviewed to identify common start and end screens for the login task. The task start was considered the first page loaded after selecting a biometric or PIN login icon. The start time was the moment when the mobile application page fully rendered in the session screen recording. On biometric recognition, the mobile application displayed a "success" page, and in fingerprint, "success" was represented by a pop-up message. The app displayed a failure message if authentication was not successful. Task end times were collected on success or failure page or pop-up load. Time in milliseconds was manually captured from video of the mobile screens.

All participants were provided time on each authentication task with no support from the facilitators. Some participants requested assistance mid-task. In these cases, they were given lightweight verbal guidance such as "try that again," or more detailed verbal and/or physical guidance if requested, like frequent verbal directional instructions (ex. "try moving the phone closer to your face"). We therefore categorized completion types (*effectiveness*) as:

- Independent success;
- Success with light guidance (few light verbal prompts);
- Success with heavy guidance (frequent, detailed verbal guidance and/or physical guidance); and
- Failure.

Independent success and success with light guidance were grouped as *trial success* in our analysis. Success with heavy guidance and failure, including instances when participants chose to end the trial, are both considered *trial failure*. Generally, choosing to end the trial only happened after a number of errors had occurred.

2.3.2 Perceived Usability

The 10-item long System Usability Scale (SUS) is an industry standard method of assessing a user's *perceived usability* of a system and has been recommended as a standard metric for comparing usability of authentication systems [23]. We deemed requiring participants to complete the 10-item questionnaire several times as too cumbersome for participants in our specific study design. Due to long task set up, short task times, and rapid switching between tasks, we selected a shorter perceived usability questionnaire, the UMUX-LITE. Figure 2-1 shows the two questionnaire items.

Item 1. This system's capabilities meet my requirements.
Item 2. This system is easy to use.

Figure 2-1. UMUX-LITE questionnaire items.

UMUX-LITE [31] is a two-item questionnaire based on the Usability Metric for User Experience (UMUX) questionnaire. It has been shown to have high reliability and validity. A regression adjusted version called the UMUX-LITEr has been found to correspond closely with the SUS in assessing user satisfaction in a given system [26, 27, 28, 29]. We report results in UMUX-LITE format but interested readers may use this adjustment to transform perceived usability data into SUS equivalency scores, which combine results of both UMUX-LITE items. The conversion is:

$$\text{SUS equivalency score} =$$

$$.65 * (((\text{UMUX LITE Item1} - 1) + (\text{UMUX LITE Item2} - 1)) * (100/12)) + 22.9$$

2.4 Ensuring Accessibility

Because we examined usability for populations with specific disabilities, it was especially important to ensure test materials and environments were accessible for people with limited or no vision and people who have hearing loss. Lab environments and building entrances were checked for accessibility prior to sessions. All equipment that was not the subject of testing was accessible to and comfortable for participants. We confirmed that all elements in the prototype application could be read by a screen reader. Signature guides were provided for users with low

or no vision to use on consent forms. Consent forms were provided digitally ahead of the session to participants with vision loss to give them time to review the information themselves. Upon scheduling, participants with hearing loss were asked if they desired American Sign Language interpretation. If requested, an ASL interpreter was present to facilitate communication during the study as well as during introductions, consent discussion, debriefing, and other immediate pre- and post-session interactions. Participants who used hearing aids in everyday life used them during the study.

Based on informally received advice within the usability community on working with people with disabilities, we chose to select participants who were willing to use their personal smartphones and install the application required for the study. Personal devices help ensure that the hardware used in research is easily accessible to participants as it enables participants to use their personal assistive technology configurations. This method also allows facilitators to observe users with audio and visual disabilities' individual approaches to using a smartphone and how the authentication methods in question interact with participants' everyday assistive technologies. Using personal devices provides privacy advantages as well (see Section 3.5).

3 Methodology

We conducted a lab study comparatively evaluating the usability of three biometric authenticators (fingerprint recognition, eye recognition, palm recognition) and one non-biometric authentication scheme (PIN). Participants completed a pre-session survey, described in Section 3.1, before the session. After giving informed consent at the start of the session, a facilitator assisted participants through prototype set-up. During the session, participants used each authentication scheme to perform login tasks on the smartphone application. After the task portion, facilitators engaged the participant in structured interviews to gather their opinions on the accessibility and usability of each authentication mode; their general preferences between the schemes; and their thoughts regarding personally using the technology to authenticate into online services. The structured interviews are not described further in the methodology as they are beyond the focus of this paper. The study ran for two and a half weeks during the summer of 2018. 29 participants took part in the study.

3.1 Pre-Session Survey

A survey on authentication use and behaviors was developed to ascertain participants' technical acumen and security awareness, and to surface meaningful relationships between experimental results, demographics, and technology perspectives. Survey analysis is outside of this paper's current focus. It is only discussed here for transparency and insight into performance data results. The survey first gathered the types of technologies participants regularly use, including assistive technology. It then evaluated their awareness and use of authentication technologies such as passwords, patterns, and biometrics. Finally, it attempted to identify how security-minded participants were by including questions about password-sharing practices, software update habits, and types of sensitive accounts they access from their devices. The behaviors surveyed were constrained to best practices for securing a sensitive application on a personal smartphone. Participants were offered the option to complete the study online in advance of the session, or on paper or verbally at the beginning of their scheduled session appointment.

3.2 Study Set-up

Sessions took place in conference rooms. Environments were accessible, and light levels were controlled to ensure minimal interference with camera-based authentication actions.

Audio recordings and top- and side-view video recordings centered on the participant's interactions with the prototype were captured. If participants had iOS devices and agreed to it, their screens were captured using iOS screen sharing to a researcher's laptop. Recordings started after the participant had provided informed signed consent and explicitly consented to being recorded. Video and audio recordings were later used to manually calculate response times and to double-check live session notes. When an ASL interpreter was present, they sat in full sight of the participant and aided communication between participant and facilitator.

Participants provided their personal mobile devices for use in the study. The facilitator guided the participant through downloading and installing the mobile application and enrolling their authentication information to the prototype, providing aid if needed. Enrolling included performing each authentication action and thus served as an introduction to unfamiliar authentication schemes and a practice for all schemes. Before enrolling schemes, the participant was instructed not to use any passwords or PINs they had used before or planned on using outside of the study.

Participants were informed that facilitators could answer questions related to the study at any time during the session and answer questions related to using the authentication schemes during set-up and after the tasks but not during experiment trials. Since the prototype used unlabeled icons as elements to navigate to authentication tasks, the icons were explained to participants and a visual cheat-sheet of the icons was provided during registration and tasks.

Participants took 60-90 minutes to complete the study and were compensated $100 USD in cash. Regardless of completion of the session, participants with disabilities received an additional $25 USD incentive to compensate for added travel time and expense.

3.3 Tasks

Tasks began at the home screen of the prototype application. The facilitator described a fictional scenario in which the user's goal was to use their mobile device to log in to a government service called MyUSA Account in order to download a digital copy of their latest tax returns. Participants were aware that the service was not real but were asked to place themselves in the scenario. It was used to ground experiences in real-life application and introduce using biometric authentication for digital government services.

The facilitator directed the participant to authenticate using a particular scheme. To start a task, the participant tapped the corresponding authenticator icon. A "trial" began when the app instructed the participant to attempt the authentication interaction. The trial continued until a Success or Fail was achieved. Before each PIN trial, the participant was reminded not to enter any passwords during tasks that they had used before or planned on using outside of the study. Tasks consisted of two sequential trials using the same scheme. A trial was an individual attempt to authenticate using the task scheme.

Trials could contain multiple authentication interactions if errors occurred. If an error occurred (known by the appearance of an error message), the participant was told to try authenticating again. The participant completed the trial by achieving a success or failure (criteria in Section 2.3.1). After successful trials, the participant was returned to the app's home screen. Task order was counterbalanced to control for the possibility of task ordering patterns influencing results. After the first or second trial ended, the facilitator asked the participant to rate their agreement with each UMUX-LITE item on a scale of 1 (strongly disagree) to 7 (strongly agree). What trial the ratings were collected after was randomized to reduce the risk of repetitive questioning influencing participant responses.

Face recognition and face/voice combination were tested during the sessions by all participants who had registered those schemes. However, unexpected updates to the prototype application during the weeks the experiment took place caused technical difficulties with registration. Not enough participants were able to successfully register the two schemes to achieve a useful sample size, so face recognition and face/voice recognition are excluded from this paper's analysis.

Participants with disabilities were encouraged to use their normal assistive technologies during the study. Participants with limited or no vision used VoiceOver, screen magnification, and color filtering assistive technologies to complete the tasks, depending on their needs. Participants with hearing loss did not use assistive technology on their mobile devices, but some made use of ASL interpretation.

3.4 Participants

We worked with a professional usability recruitment firm to recruit 30 participants who were U.S. citizens in the Northern Virginia and Baltimore region. We aimed to balance the sample overall for gender and include participants across the following age groups: 18-24 years old; 25-34 years old; 35-44 years old; 45-54 years old; 55-64 years old; and 65 years old or over. All participants were required to be fluent in English or ASL.

One of thirty participants did not show for their session and could not be rescheduled, giving an overall participant count of 29 (13 women, 16 men). Two participants were unable to set up the prototype due to technical difficulties, giving a final count of 27 participants providing task performance data. These two participants still took part in the survey and structured interview. Participant ages skewed older. Table 3-1 gives the number of participants in each age range.

Table 3-1. Number of participants in each age range, 29 participants total.

	Age range (years)				
	25 34	35 44	45 54	55 64	65+
Number of participants	1	6	8	5	9

All participants were required to own a smartphone and agree to install a mobile phone application for the duration of the study. Smartphones were Android OS 4.4+ or iOS models 5s and above or iOS 9.1+ and had operational fingerprint sensors and operational front-facing cameras. Participants were requested to bring all assistive technology they use regularly with

their mobile devices to their study session. Six participants owned an Android device and 23 owned an iOS device.

Participants were grouped into those with no disabilities (control), participants with hearing loss, and participants with limited or no vision. We aimed for recruitment of 10 participants with moderate to profound hearing loss (phrased as "hearing impairment") with no more than 5 who required an ASL interpreter, and 10 participants with moderate to profound vision loss (phrased as "vision impairment"). An additional requirement was that these participants not have any other disabilities. All disabilities and levels were self-reported by participants to the usability recruitment firm. The following definitions were provided to the firm for recruitment guidance:

Visual impairment (at the participant's presenting corrected vision level) [32]:

- Low vision, consisting of partially sighted, moderate visual impairment or severe visual impairment; and
- Profound visual impairment, legally blind or totally blind.

Hearing impairment [33, 34, 35]:

- Moderate impairment or hearing loss, or hard-of-hearing;

Detail: Someone with a moderate level of hearing loss has difficulties hearing regular conversational speech, even at close distances. This includes people who use technology that allows them to operate at a less severe hearing loss level, ex. cochlear implants, hearing aids.

- and, Severe to profound impairment or hearing loss, deaf, or total hearing loss.

Detail: Someone with a severe or profound level of hearing loss does not hear conversational speech. Someone with a severe level may hear very loud speech or loud sounds in the environment, such as a fire truck siren or a door slamming. Someone with a profound level or someone who is deaf does not hear conversational speech and may perceive loud sounds as vibrations. They cannot understand speech (with or without hearing aids or other devices) using sound alone (i.e., no visual cues such as lipreading).

Table 3-2 details participants per group and level, as reported by the recruitment firm. It also presents how many participants completed enrollment in each authentication scheme.

Table 3-2. Participant demographics and enrollments in authentication schemes.

		Disability Type and Level						
		Visual		Hearing			None	All participants
		Total	Moderate	Total	Severe	Moderate		
	Total participants	6	3	4	2	5	9	29
		9		11			9	
Participants who enrolled in…	PIN	7		11			9	27
	Finger print	7		11			9	27
	Eye print	7		9			9	25
	Palm print	5		9			6	20
	Face	1		2			3	6
	Voice / Face	1		6			6	13

3.5 Ethics & Privacy

The experimental design was approved by The MITRE Institutional Review Board (IRB). At the start of each session, the participants were given a consent form to sign, detailing the study and their rights as participants. Consent forms were provided in accessible formats and with longer review times when appropriate. We took care to treat all participants with respect and performed accessibility checks of materials and lab settings before sessions (see Section 2.4). Participants were informed that, among other participant rights, they would receive a pro-rated incentive if they chose to end the session early.

Facilitators reminded participants frequently during the study not to use any past or future personal passwords or PINs. The simulation prototype did not include any identity verification steps. All passwords, PINs and biometric data created during the study were stored locally to the participant's personal smartphone and were not transmitted off of the device or out of the application. Facilitators supervised participants securely installing and uninstalling the prototype at the start and finish of each research session. Participants were made aware of these precautions.

4 Results

This section reports the quantitative data gathered, organized by metric. We also show participants' prior exposure to biometric authentication schemes, as reported in pre-session questionnaires. As this paper focuses on task performance data, analysis of qualitative interviews is deferred to future publications. Tables A-1 through A-6 in the appendix present further results details.

4.1 Perceived Usability

Perceived usability was measured through responses to UMUX-LITE items, shown in Figures 4-1 and 4-2. Note that 1 corresponds to the "strongly disagree" response and 7 to "strongly agree."

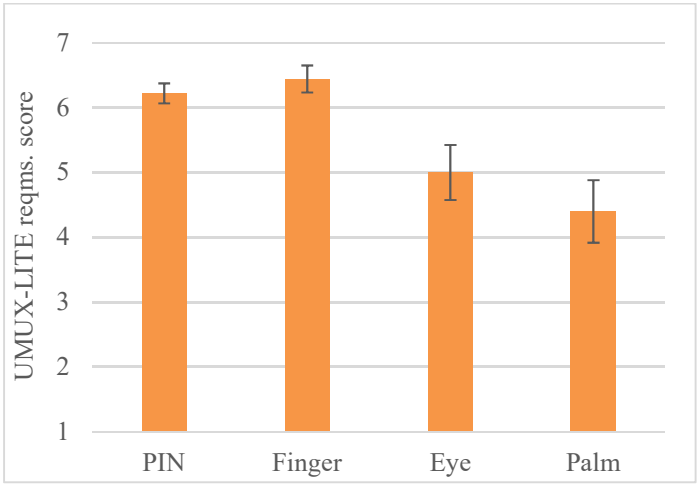

Figure 4-1. Mean UMUX-LITE requirements item scores across authentication schemes and all populations, with standard error shown.

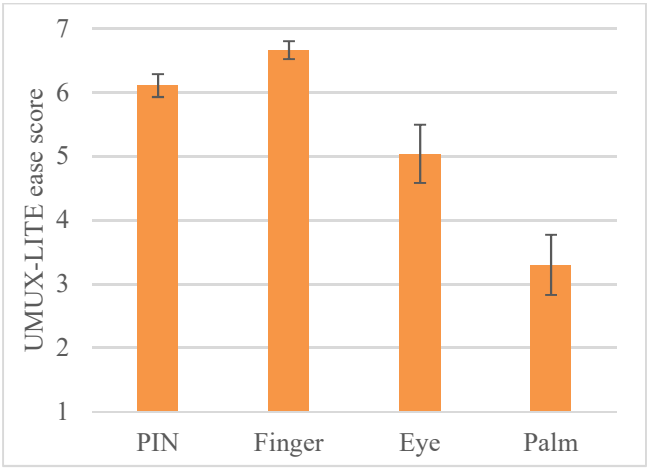

Figure 4-2. Mean UMUX-LITE ease item scores across authentication schemes and all populations, with standard error shown.

UMUX-LITE data were not normally distributed, therefore non-parametric tests were needed. Due to the interval nature of the data, k independent samples analysis was performed.

A Kruskal-Wallis H test, a one-way ANOVA on ranks for non-parametric data, showed that there were no statistically significant differences in requirements item scores between the different populations (χ^2 (2) = 2.000, p = 0.368, with a mean rank score of 45.17 for no disability, 54.25 for hearing loss, and 49.62 for vision loss); nor in ease of use item scores between the different populations (χ^2 (2) = 0.415, p = 0.813, with a mean rank score of 49.33 for no disability, 52.01 for hearing loss, and 47.75 for vision loss).

There were significant differences in the UMUX-LITE requirement ratings between schemes; χ^2 (3) = 19.000, p = 0.000, with a mean rank score of 55.56, 42.36, 64.54, and 32.43 for PIN, eye, fingerprint, and palm, respectively. Mann-Whitney post-hoc tests, the non-parametric alternative

to the independent sample t-test, found significant differences in requirements item scores between several schemes. Median requirements ratings were significantly higher for PIN (6) than palm (5); (U = 135.000, p = 0.003). Median requirements ratings were significantly higher for fingerprint (7) than eye (6); U = 197.500, p = 0.005. Finally, median requirements ratings were significantly higher for fingerprint (7) than palm (5); U = 45.000, p = 0.000.

A Kruskal-Wallis H test showed significant differences in ease of use item scores between schemes; χ^2 (3) = 33.048, p = 0.000, with a mean rank score of 54.50, 45.76, 68.94, and 23.65 for PIN, eye, fingerprint, and palm. According to a Mann-Whitney post-hoc test, median UMUX-LITE ease ratings were significantly higher for fingerprint (7) than PIN (6) (U = 228.500, p = 0.007), eye (6) (U = 187.000, p = 0.002), and palm (3) (U = 100.000, p = 0.000). Median ease scores were significantly higher for PIN than palm (3); U = 75.000, p = 0.000. They were also significantly higher for eye than palm; U = 143.000, p = 0.013.

A Mann-Whitney post-hoc test was run to test the third hypothesis about the experiences of participants with vision loss. It determined that median requirements item scores were significantly higher for PIN (7) than eye (4), (U = 5.500, p = 0.011); and palm (3), (U = 5.500, p = 0.036). Median requirements scores for fingerprint (7) were significantly higher than eye, (U = 5.500, p = 0.011); and palm, (U = 5.500, p = 0.036). Finally, median ease item scores were significantly higher for fingerprint (7) than for eye (3) (U = 9.000, p = 0.033); and for palm (2) (U = 5.500, p = 0.036).

4.2 Effectiveness

Effectiveness was assessed through measuring completion rate. Task completion rate is the number of successful task completions out of the number of attempted task completions (which are also the number of successful scheme registrations). Note that each participant had two task attempts (one per trial). Figure 4-3 shows completion rate results.

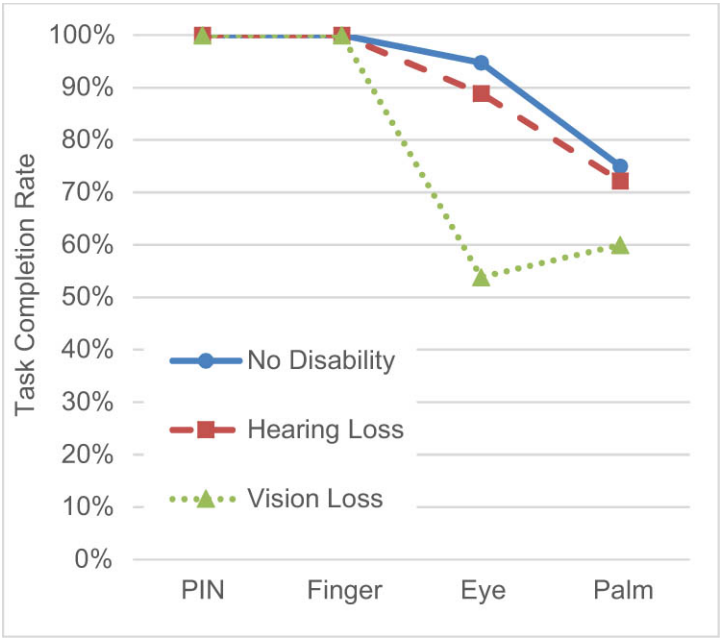

Figure 4-3. Mean completion rates across authentication schemes and participant groups.

A logistic regression was performed to ascertain the effects of population on the likelihood that participants successfully completed the tasks. The model explained 5.4 percent (Nagelkerke R2) of the variance in completion rate and correctly classified 89.4 percent of cases. Population was found to have an effect, with participants with no disability being 3.690 times more likely to be successful than those with vision loss; χ^2 (1) = 4.372, p = 0.037.

Every participant who registered PIN and fingerprint was able to successfully complete PIN and fingerprint tasks, regardless of participant group. No participant group had 100 percent task completion rates for eye and palm tasks. However, a logistic regression performed to examine the effects of the authentication scheme on the likelihood that participants successfully completed the tasks found no significant differences between completion rates due to mechanism. The model explained 35.5 percent (Nagelkerke R2) of the variance in completion rate and correctly classified 89.4 percent of cases.

To ensure no learning effects were at play, a logistic regression was used to ascertain the effects of number of trials (1 or 2) on the likelihood that participants successfully completed the tasks. The model explained 0.2 percent (Nagelkerke R2) of the variance in completion rate and correctly classified 89.8 percent of cases. There were no significant differences between completion rates due to trial number and thus no learning effect due to number of trials experienced; χ^2 (1) = 0.185, p = 0.667.

A logistic regression was performed to ascertain, specifically for the vision loss group, the effects of scheme on likelihood that participants successfully completed the tasks. The model explained 47.6 percent (Nagelkerke R2) of the variance in completion rate and correctly classified 80.8 percent of cases. There were no significant differences for this group between completion rates due to scheme.

We planned to examine error rate as a component of effectiveness, with an error defined as an instance when the participant does not fail the task but must redo the authentication action. The prototype gave error prompts such as "Incorrect Match, This palm does not match the saved value," "Authentication Aborted, The eye authenticator timed out," and "Unable to authenticate, Eye verification not matched." However, prompts also included descriptions like "An Error Occurred, Unexpected HTTP status code received" and simply "Authentication Failed" with no explanation. Since some error messages were opaque and the prototype was created and managed by a third party, we were unable to accurately diagnose the genesis of each participant error or to guarantee that all errors were user-caused and never the result of a technical glitch (as the HTTP status code message implied). Therefore we consider error rate data unfit for the same degree of scrutiny as completion rate, and do not report it here.

4.3 Efficiency

Efficiency is operationalized as response time, specifically, the time elapsed from when the prototype app instructed the participant to attempt the authentication interaction until the interface's indication of task success or failure (overall, length of a trial). Data from all success task trials are reported here. Figure 4-4 presents response time results.

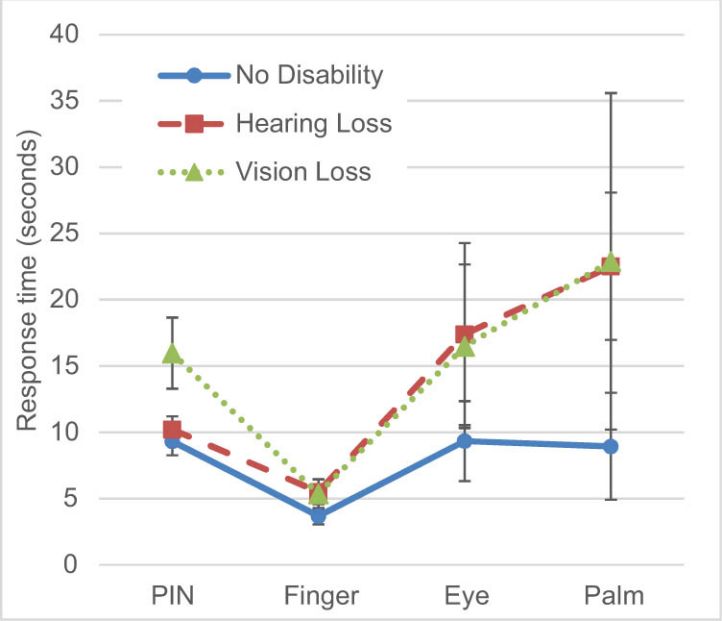

Figure 4-4. Mean response time from all success trials across authentication schemes and participant groups, with standard error shown.

Mauchly's Test of Sphericity, which tests the assumption that the variance between the levels of independent variables are equal, indicated that the assumption of sphericity had been violated, ($\chi 2$ (5) = 44.308, p = 0.000). A Greenhouse-Geisser correction, typically used when the assumption of sphericity is violated, was used. A repeated measures ANOVA found that population had no significant effect on response time; F(2, 20) = 2.246, p = 0.132. A repeated measures ANOVA also found that scheme had no significant effect on response time; F(1.741, 34.823) = 3.260, p = 0.057.

However, the lack of power ($\eta = 0.546$) may have limited the ability to find a significant effect. Because differences between schemes were hypothesized, post-hoc tests were still performed on scheme comparisons. Additionally, many post-hoc procedures are designed to control familywise error rates in the absence of a significant prior omnibus analysis. Simple contrast post-hoc tests with Bonferroni correction, a correction made to p-values when several statistical tests are performed on a single data set, found significant differences in response times. Specifically, the mean response time for fingerprint (4.86s) was significantly faster than mean response times for all other schemes (PIN (11.41s), $F(1, 20) = 37.520$, $p = 0.000$; eye (13.71s), $F(1, 20) = 5.339$, $p = 0.032$; and palm (18.24s), $F(1, 20) = 10.421$, $p = 0.004$).

To test specifically within the vision loss participant group, a one-way repeated measures ANOVA was performed to ascertain the effect of scheme on reaction time, with planned pairwise comparisons. No significant differences between schemes were found $F(3,12) = 1.154$, $p = 0.367$. The lack of power ($\eta = 0.236$) may have limited the ability to find a significant effect. Because differences within the vision loss group were hypothesized, post-hoc tests were performed, but planned pairwise comparisons found no significant differences between schemes on reaction time for the group with vision loss.

4.4 Biometric Authentication Scheme Experience

In the pre-session survey, 30 participants reported on the authentication schemes they had previously used to secure both their device and secure any personal accounts (such as a banking account). Items were phrased: “Do you have experience with the following ways to__?” Illustrative examples were included, like banking account for personal account and RSA token for digital key. Table 4-1 shows their responses. Password, PINs, two-factor with email and SMS, and fingerprint biometrics were all widely used. All participants reported experience using passwords to secure both devices and individual accounts, and over 80 percent reported experience using a PIN or pattern. Most participants had experience with some form of two-factor authentication, with the majority of the experience with a code received by email or SMS or with using a security question. A majority (83 percent) had used a biometric fingerprint to unlock their smartphone, and 60 percent had used fingerprint to unlock a personal account. A small number reported experience with face or voice biometrics to secure their phone (3 with voice, 2 with face). None had used palm or eye biometrics before for securing devices or accounts for web services.

Table 4-1. Participant responses to questionnaire items about prior experience with authentication methods.

Authentication method	Number of "yes" responses to the following questionnaire items:	
	… secure your personal devices to access a web service?	… secure your personal accounts to access a web service?
Passwords	30	30
Pin or pattern	25	24
2-factor using code received by email	23	22
2-factor using security question	21	22
2-factor using code received by personal cellphone or smartphone	20	19
Two-factor using standalone device with digital key	7	5
Two-factor using a code received by landline phone	6	8
Two-factor using an online or software digital key (e.g., Google Authenticator, Duo)	4	4
Biometric – fingerprint	25	19
Biometric – voice	3	2
Biometric – face	2	1
Biometric – iris	0	0
Other	0	0

5 Discussion

5.1 Traditional Authentication & Biometric Authentication

Performance data partially supported the hypothesis that PIN and biometric authentication schemes would differ in the metrics we collected. PIN had significantly lower perceived usability (specifically, ease of use) and lower efficiency (slower response time) than fingerprint. PIN had significantly higher perceived usability than palm (both items). Counter to expectations, no significant differences were seen between PIN and eye in any metric, and no significant differences in completion rate were seen for any scheme.

PIN and Fingerprint

The PIN/fingerprint difference could be caused by the two schemes' different memory requirements and their required target acquisition actions. To use PIN successfully, participants had to recall a six-digit pattern, while they did not have to remember anything for fingerprint. For PIN, users performed six input actions in selecting six digits in the keypad entry interface; for fingerprint, they simply had to touch one input location (the touch sensor). In both the recall and the physical input differences, PIN's actions have a longer inherent time burden than do fingerprint's, which could explain the response time difference. When using PIN with a screen reader during sessions, participants often had to cycle through digits listening for the correct one

before selection – again, a possible time sink. Recall also brings in a cognitive element that fingerprint does not require. Preferences against needing to create and remember PINs could have affected perceived usability ratings.

The added cognitive component and the speed differences might have contributed to participants rating PIN and fingerprint differently for ease of use. The lack of difference in the "meets my requirements" aspect of perceived usability could indicate that participants held expectations of a minimum threshold of usability required to fulfill their needs, and that both schemes met such a threshold, causing a ceiling effect. Participants may also have viewed PIN and fingerprint similarly in terms of security the schemes provide.

PIN & Palm

PIN and palm's perceived usability difference could again stem from different cognitive requirements and different time burdens. The palm authentication interaction of positioning the palm parallel to the phone's screen-side camera and adjusting does not easily compare time-wise to PIN's classic target acquisition and selection movements of selecting numbers in an on-screen keypad. That said, palm had a longer mean response time (18.24s) than PIN. Basic times for both actions could be assessed, for example with Goals, Operators, Methods, Selection Rules (GOMS) model analysis [36], to delve deeper into comparisons of the schemes' inherent time burdens. PIN and palm's cognitive actions differ as well; remembering a number sequence is a one-time recall, while reaching and maintaining a correct relative hand position involves continuous spatial monitoring and adjustment.

Differences in time to authenticate and in cognitive actions required, as well as in perceived security provided by each scheme, could have contributed to the differences in perceived usability between the schemes. Prior exposure could have had an effect as well, since a majority of participants reported having used PIN or pattern before the session and no participants reported using palm authentication before the session.

The palm print condition had the smallest sample size since fewer participants were able to successfully enroll palm print than other schemes. The sample shrunk further for response time data as only results from successful trials were included in efficiency analysis. The lack of a significant efficiency difference does not align with expectations, but it may have been caused or affected by the lack of power and the high variability in palm response time results.

PIN & Eye

Counter to expectations, there were no significant differences between PIN and eye schemes. Eye's low sample size could have impacted the ability to find a significant difference if there was one, although eye's sample size was larger than palm's. From observation, eye seems to be more similar to palm than to PIN. Like palm, there is no recall needed, and the user continuously monitors and adjusts their relative hand positions. Unlike palm, in eye, a hand containing the mobile device is positioned relative to the user's head, and authentication requires assuming a specific head posture and face configuration (eyes open, gaze on the phone). In fact, eye and palm differed significantly in their ease of use item scores.

Within sighted participants, the prototype app feedback for eye seemed easier for users to monitor than did feedback for palm. During palm authentication, some sighted users shifted their hand away from and back over the screen as well as tilted their hand to peek under it in order to

view the screen more fully. Some users remarked on these actions. No such actions or comments on ability to perceive feedback were observed during eye authentication sessions with sighted participants (perception of feedback being different from understanding of feedback).

We are ultimately unsure as to why participant performance did not differ significantly between the PIN and eye. There were no statistically significant effects of participant group on perceived usability or efficiency, but participants with no disability were 3.69 times more likely to complete tasks successfully than were participants with vision loss. This suggests that vision loss participants' different experiences of PIN and eye bear further study.

Overall

PIN-fingerprint and PIN-palm comparison differences were supported by a subset of performance data, though not by completion rate (addressed in Section 5.3). A PIN-eye difference was unsupported. This mixed bag suggests that there might not be a clear usability divide between traditional authentication methods and biometric schemes. Another possibility is that traditional methods may indeed have distinct usability differences from some biometrics, but that grouping the biometrics examined here into a single usability category is an overreach.

Biometrics offer many advantages over traditional authentication schemes like PIN and password. They do not require recall, which cuts down on cognitive burden as well as time. However, some biometrics, such as palm and eye, require additional monitoring of spatial information. This comparison merits further research to empirically evaluate the usability of PIN and other biometrics that can be captured by smartphone cameras or sensors. Future studies could explore: comparisons with use over time, for example authenticating several times over the course of months; comparing with stringent PIN or password creation requirements; use in field settings; larger sample sizes; and users with other single or concurrent disabilities.

5.2 Dynamic Positioning Interactions in Authentication

Fingerprint, the non-dynamic-positioning biometric authentication scheme, had significantly higher perceived usability (both items) and better efficiency than eye and palm, the dynamic-positioning biometrics. This supports the hypothesis that biometric authentication schemes' performance results would divide along the dynamic positioning aspect. Counter to expectations, no scheme showed significantly different completion rates.

As discussed earlier, eye and palm share similarities – no need for recall, and a continuous spatial information monitoring by the user. Fingerprint also does not require recall, but neither does it need hand and/or head position perception and adjustment. It simply requires the user to locate and select a single, non-moving target with tactile edges. In cases where the user is holding the phone in one hand, they can even brace their fingerprint-input hand against their phone-holding hand. Dynamic positioning actions require more granular and frequent monitoring and adaptation of the body part's location as well as movement and pausing in space, generally with no physical bracing or tactile breakpoints. This difference in the use of dynamic positioning – positioning one body part relative to another, whether hand to hand or hand to head – is a likely cause for the performance differences seen between schemes.

There were no significant differences in completion rates between either comparison (finger and eye, finger and palm). This lack of significance could stem from a small sample size, or from

differing levels of familiarity with the schemes. Most participants had previous experience with fingerprint and none had used eye or palm before their sessions.

Results partially support the prediction that biometric schemes would exhibit a usability split along dynamic positioning lines. Further research is needed to confirm this split and to explore its nature – are there important distinctions within the types of biometrics captured by smartphone cameras and sensors? Are there meaningful groupings within the dynamic positioning conglomeration? Do individuals with certain disabilities experience disproportionally better or worse usability from positioning biometrics? Might different feedback channels (ex. audio tone, audio text, haptic vibration) of positioning guidance mitigate the effects of the split, so much so as to erase the dynamic positioning performance difference?

Results gave some support to the third hypothesis that the user group with vision loss would experience better performance with non-positioning biometrics than with positioning biometrics. Low vision and blind participants reported significantly better perceived usability (both items) with fingerprint than with eye or palm. Also, participants with vision loss were far less likely to complete tasks successfully with given schemes than were control group participants (3.69 times). Since all enrolled participants had 100 percent completion rates only with PIN and fingerprint, this lower-success effect is likely occurring with eye and palm. No significant differences were found between completion rates due to scheme within the vision loss group, but this pattern is noteworthy and should be explored further in future. These results suggest that dynamic positioning is an important aspect of biometric usability and accessibility for users with low or no vision.

However, there were no significant efficiency differences between schemes for the group with vision loss. This could be affected by the lack of power.

It should be noted that the palm sample size of users with visual disabilities was small at 5 participants (other participants in the group were unable to enroll the scheme successfully). While 5 is not considered out of the ordinary for usability testing, it is a very small sample to support statistical analyses. Palm's sample size may have impacted results.

5.3 Effectiveness Metric

Completion rate did not vary significantly due to scheme. This was surprising, as PIN and fingerprint had 100 percent task completion rate and eye and palm had lower rates (mean 82.35 percent and 70 percent respectively, over all participants). It could be that there was not enough power to see a significant effect. Levels of prior exposure to the schemes might have impacted completion rate results; 83 percent of participants had used fingerprint and PIN or pattern before, while no participants reported experience with eye or palm before the study. There was no learning effect due to trial number, but familiarity could have had an impact larger than what the experience from registration and two trials could correct for.

Population significantly affected effectiveness, with participants with no disability being 3.69 times more likely to be successful than those with vision loss. All unsuccessful vision loss participants had been able to register the schemes and could technically access the app content, but baseline access did not mean they could successfully use the schemes. Therefore, we

recommend completion rate as a consideration in assessing technology usability and accessibility for low vision and blind users.

6 Limitations & Future Research Directions

The response time measurement method was prone to human error. As described in the Methodology, researchers manually calculated response times from videos of the prototype screen. Though care was taken to move through videos at low frame rates, measurements may have gained errors during this process. We recommend automating task time capture instead.

As detailed in the Results, useful error data could not be captured consistently due to prototype limitations. We believe error rate and diagnosis would be useful for future work.

Enrollment, or registration, performance was outside the scope of this study. Enrollment performance data, such as how difficult the participant found enrollment in a scheme and how many registration fails they caused, could give interesting insights.

What trial the perceived usability ratings were collected after was randomized to reduce the risk of repetitive questioning influencing participant responses. In retrospect, the risk of question repetition influence may have been lower than risk of effects due to uneven experience with the system. To address this, we recommend gathering self-report ratings after every trial or after the same number of trials, and/or building in more participant interactions with the system in order to pursue a high enough level of familiarity that lack of experience does not have an effect. The latter is the better option, as it would also combat difference in general levels of familiarity with particular schemes, as participants' prior exposure to authentication schemes could have had an effect, especially on results that showed high variability. Prior experience with the tested technology has been shown to affect SUS scores [37]. Previous exposure should be examined in future studies for possible impact on perceived usability or other performance results, or should be further controlled for.

Some metrics may be better suited to testing *across* disabilities and some to testing *between* disabilities. Response time might not be a useful metric for comparisons between groups where groups have different disabilities. It could be a more useful metric in within-group situations, since the functional effects of the disability on response time (ex. effects of poor fine motor control) would be standardized. Assistive tech may additionally influence task time and would also be better standardized within groups. Completion rate and self-reported reactions (ex. SUS scores), on the other hand, can more easily be compared across groups.

It is possible that slower response time does not always indicate inferior usability. Users might consider a scheme usable as long as it meets a minimum response time threshold and might at that point not be concerned with what scheme is faster.

Our findings should be validated through replication of the experiment with larger participant pools. Our study size was small due to the difficulty of recruiting participants with disabilities and to resource limitations and technical difficulties. We hope this work is expanded further by studying more types of disabilities and by investigating the effects of severity levels within disabilities. More biometrics should be compared in order to expand authentication design guidance to other schemes that will become more common in the future, as well as to the face and voice biometrics for which not enough data could be collected. More research into the

directionality of usability differences for people with disabilities would also be valuable as it could contribute to clear, evidence-based guidance toward selecting certain schemes over others.

We recruited participants into groups based on their self-reported disabilities. During the study, there was confusion over the definition of disability severity levels ("moderate," "severe," "total"). Many users did not describe their disabilities with this terminology. We recommend instead including assistive technology use when forming participant groups, as that may be more indicative of the type and degree of a hearing or vision loss. We also recommend a focus on testing authentication schemes with populations whose disabilities map to the scheme's interaction requirements, as these may have more immediate value. We observed usability decrements for participants with vision loss using schemes with a greater reliance on visual feedback, while users with hearing loss and control participants did not seem to have markedly different experiences with our analyzed schemes, none of which involved audio or speech-based interactions.

This work prompts ideas for future pursuit. Considering how the specific interactions that a biometric requires relates to the abilities of the user could surface more accessibility considerations like dynamic positioning that can be used to guide accessible authentication design. Further, it is not uncommon for people to have more than one disability. Usability for participants with multiple disabilities should be investigated.

We are also interested in how learnability may play a role in biometric accessibility. Participants with vision loss often expressed excitement and interest in eye and palm authentication during the study but sometimes could not employ them without verbal and occasionally physical assistance from facilitators. However, these participants said they were optimistic about their ability to learn to use the schemes over time. During informal background interviews, several technology users who had vision loss indicated that they frequently used iOS FaceID to secure their smartphones. They reported that the interactions were difficult at first, but that after some practice, they were highly satisfied with face recognition authentication and used it regularly. With repeated, possibly guided practice, certain authentication schemes that are initially difficult for participants with a disability to use may become easy and even preferred.

7 Conclusion

Our study found that there is not a clear usability divide between the traditional authentication method and all biometric schemes as a group. There may be no marked usability distinction, or it may be that fingerprint, eye, and palm are too distinct to consider together. The question of differences between traditional authentication schemes, like PIN or password, and biometric authenticators that can be captured by smartphones merits further exploration.

The results of our study partially supported a "dynamic positioning" split among the biometrics tested, with participants showing markedly different usability experiences between fingerprint and eye and between fingerprint and palm. The non-positioning fingerprint scheme seemed somewhat more usable for participants with visual disabilities than the positioning eye and palm. Findings add weight to the positioning split. We propose research questions to further probe this categorization and other questions raised during the study, share thoughts on the metrics deployed in this usability evaluation, and discuss limitations in the experiment.

Based on the evidence collected, we propose dynamic device positioning as a new consideration for biometric usability evaluations. This new principle is operationalized as two actionable recommendations, to be used in authentication process design. Our recommendations were created with the accessibility and usability needs of citizen-facing federal agencies in mind. Our work also contributes empirical findings on the usability of biometric authentication schemes for users with disabilities, expanding the body of work and demonstrating methods for comparative biometric usability evaluation with an accessibility focus.

7.1 Dynamic Positioning as an Accessibility Consideration

Smartphones offer a wide range of biometric capture, from fingerprint, eye, iris, face, and voice to emerging biometrics like ear shape. They offer conveniences to all users, including those with disabilities, but based on our research we feel that a better understanding of the accessibility of different biometrics is needed. There is little in-depth usability guidance for designers to consult when integrating multi-factor authentication into their services. Decision-makers at federal agencies with accessibility mandates need to choose authentication techniques relatively early in the design process. They typically do not have the resources nor the time to perform rigorous experimentation on their web service's usability for people with disabilities. We seek to provide evidence-based knowledge to guide them in evaluating authentication options for people with disabilities and propose dynamic device positioning as a new consideration for usability evaluations of biometrics.

Participants with vision loss were far less likely to successfully complete tasks with given schemes than were control group participants. With this in mind, we suggest that completion rate is a key metric to consider when populations with disabilities are involved.

The fingerprint/eye and fingerprint/palm perceived usability and efficiency differences suggest that dynamic positioning could have an impact on biometric accessibility for users with low or no vision, though the relationship should be studied further and with larger participant pools.

We see positioning used alongside accessibility principles such as text alternatives for non-text content [38]. Based on our findings, we offer the following recommendations to guide decision-makers in selecting biometric authentication techniques:

- A dynamic positioning biometric should never be the sole authentication scheme.
- Multi-factor authentication using biometrics should offer at least one non-dynamic positioning biometric. Fingerprint is a good option until other schemes are empirically shown to be more accessible.

8 References

[1] D. M. Taylor, "Americans With Disabilities: 2014," November 2018. [Online]. Available: https://www.census.gov/content/dam/Census/library/publications/2018/demo/p70-152.pdf. [Accessed 1 March 2019].

[2] "Disability Language Style Guide," 29 March 2019. [Online]. Available: https://ncdj.org/style-guide/.

[3] A. W. Roberts, S. U. Ogunwole, L. Blakeslee and M. A. Rabe, "The Population 65 Years and Older in the United States: 2016," October 2018. [Online]. Available: https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-38.pdf. [Accessed 1 March 2019].

[4] J. Bonneau, C. Herley, F. M. Stajano and P. C. v. Oorschot, "Passwords and the Evolution of Imperfect Authentication," 2014.

[5] Pew Research Center, "Mobile Fact Sheet," 5 February 2018. [Online]. Available: http://www.pewinternet.org/fact-sheet/mobile/. [Accessed 1 March 2019].

[6] A. Ometov, S. Bezzateev, N. Mäkitalo, S. Andreev, T. Mikkonen and Y. Koucheryavy, "Multi-factor authentication: A survey.," *Cryptography 2, no. 1,* p. 1, 2018.

[7] W. Newhouse, B. Johnson, S. Kinling, B. Mulugeta and K. Sandlin, "Multifactor Authentication for E-Commerce Risk-Based, FIDO Universal Second Factor Implementations for Purchasers," NIST; NCCOE, 2018.

[8] J. Comer, "21st Century Integrated Digital Experience Act Floor Speech," 29 November 2018. [Online]. Available: https://votesmart.org/public-statement/1306374/21st-century-integrated-digital-experience-act#.XHjELlNKjfb. [Accessed 1 March 2019].

[9] "H.R.5402," [Online]. Available: https://www.congress.gov/bill/115th-congress/house-bill/5402. [Accessed 1 March 2019].

[10] "IT Modernization Centers of Excellence," [Online]. Available: https://coe.gsa.gov/. [Accessed 1 March 2019].

[11] F. Konkel, "Bill Would Create Federal Customer Service Standards," 4 April 2018. [Online]. Available: https://www.nextgov.com/cio-briefing/2018/04/bill-would-create-federal-customer-service-standards/147197/. [Accessed 1 March 2019].

[12] "President's Management Agenda," [Online]. Available: https://www.performance.gov/PMA/Presidents_Management_Agenda.pdf. [Accessed 1 March 2019].

[13] J. Corrigan, "The IDEA Act would require agencies to upgrade their websites to meet basic security and usability standards, but lawmakers did make some changes.," 29 November 2018. [Online]. Available: https://www.nextgov.com/it-modernization/2018/11/house-passes-bill-improve-governments-digital-services/153162/. [Accessed 1 March 2019].

[14] J. Cardello and S. Farrell, "HealthCare.gov's Account Setup: 10 Broken Usability Guidelines," 27 July 2017. [Online]. Available: https://www.nngroup.com/articles/affordable_care_act_usability_issues/. [Accessed 1 March 2019].

[15] F. Konkel, "It Costs Taxpayers $41 Per Phone Call To IRS," 9 February 2018. [Online]. Available: https://www.nextgov.com/emerging-tech/2018/02/it-costs-taxpayers-41-phone-call-irs/145870/. [Accessed 1 March 2019].

[16] [Online]. Available: https://www.section508.gov/. [Accessed 1 March 2019].

[17] "S.3050 - 21st Century IDEA," [Online]. Available: https://www.congress.gov/bill/115th-congress/senate-bill/3050/text. [Accessed 1 March 2019].

[18] P. A. Grassi, J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, N. B. Lefkovitz, J. M. Danker, Y.-Y. Choong, K. K. Greene and M. F. Theofanos, "NIST Special Publication 800-63B," June 2017. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-63b. [Accessed 1 March 2019].

[19] L. Kessem, "IBM Study: Consumers Weigh in on Biometrics, Authentication and the Future of Identity," 29 January 2018. [Online]. Available: https://securityintelligence.com/new-ibm-study-consumers-weigh-in-on-biometrics-authentication-and-the-future-of-identity/. [Accessed 1 March 2019].

[20] R. L. German and K. S. Barber, "Consumer Attitudes About Biometric Authentication," The University of Texas at Austin , 2018.

[21] R. Blanco-Gonzalo, C. Lunerti, R. Sanchez-Reillo and R. M. Guest, "Biometrics: Accessibility challenge or opportunity?," *PLOS ONE,* 2018.

[22] M. Theofanos, B. Stanton and C. A. Wolfson, "Usability & Biometrics Ensuring Successful Biometric Systems," 11 June 2008. [Online]. Available: https://www.nist.gov/sites/default/files/usability_and_biometrics_final2.pdf. [Accessed 1 March 2019].

[23] S. Ruoti, B. Roberts and K. Seamons, "Authentication Melee: A Usability Analysis of Seven Web Authentication Systems," *Proceedings of the 24th International Conference on World Wide Web,* pp. 916-926, 2015.

[24] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh and S. Ben-David, "Biometric authentication on a mobile device: a study of user effort, error and task disruption," *Proceedings of the 28th Annual Computer Security Applications Conference,* pp. 159-168, 2012.

[25] M. A. Sasse and K. Krol, "Usable biometrics for an ageing population," in *Age Factors in Biometric Processing*, Security, IET Digital Library, 2013, pp. 303-320.

[26] J. Lewis, "Measuring perceived usability: The CSUQ, SUS, and UMUX," *International Journal of Human–Computer Interaction, 34(12),* pp. 1148-1156, 2018.

[27] J. R. Lewis, B. S. Utesch and D. E. Maher, "Investigating the correspondence between UMUX-LITE and SUS Scores," in *International Conference of Design, User Experience, and Usability*, 2015.

[28] S. Borsci, S. Federici, S. Bacci, M. Gnaldi and F. Bartolucci, "Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience," *International Journal of Human-Computer Interaction 31 no. 8,* pp. 484-495, 2015.

[29] M. I. Berkman and D. Karahoca, "Re-assessing the usability metric for user experience (UMUX) scale," *Journal of Usability Studies 11, no. 3,* pp. 89-109, 2016.

[30] "ISO 9241-11:2018(en) Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts," [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en. [Accessed 1 March 2019].

[31] J. R. Lewis, B. S. Utesch and D. E. Maher, "UMUX-LITE: when there's no time for the SUS," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[32] World Health Organization, "Change the Definition of Blindness," [Online]. Available: https://www.who.int/blindness/Change%20the%20Definition%20of%20Blindness.pdf?ua=1. [Accessed 29 March 2019].

[33] World Health Organization, "Grades of hearing impairment," World Health Organization, 21 November 2017. [Online]. Available: https://www.who.int/pbd/deafness/hearing_impairment_grades/en/. [Accessed 29 March 2019].

[34] Canadian Association of the Deaf - Association des Sourds du Canada, "Definition of "Deaf"," Canadian Association of the Deaf - Association des Sourds du Canada, 3 July 2015. [Online]. Available: http://cad.ca/issues-positions/definition-of-deaf/. [Accessed 29 March 2019].

[35] D. Clason, "Degrees of Hearing Loss," Healthy Hearing, 10 April 2015. [Online]. Available: https://www.healthyhearing.com/report/41775-Degrees-of-hearing-loss. [Accessed 29 March 2019].

[36] D. E. Kieras, "A guide to GOMS model usability evaluation using GOMSL and GLEAN3," University of Michigan, 1999.

[37] S. McLellan, A. Muddimer and S. C. Peres, "The effect of experience on System Usability Scale ratings," *Journal of Usability Studies,* vol. 7, no. 2, pp. 56-67, 2012.

[38] "Accessibility Principles," [Online]. Available: https://www.w3.org/WAI/fundamentals/accessibility-principles/. [Accessed 1 March 2019].

This page intentionally left blank.

Appendix A Details on Usability Performance Results

This appendix presents additional details on usability performance results.

Table A-1. Perceived usability results for all participant groups combined.

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	27	5	7	6	6.2222	.15408	.80064
PIN ease	27	4	7	6	6.1111	.17969	.93370
Finger reqms	27	2	7	7	6.4444	.20901	1.08604
Finger ease	27	4	7	7	6.6667	.14122	.73380
Eye reqms	25	1	7	6	5.0000	.42426	2.12132
Eye ease	25	1	7	6	5.0400	.45636	2.28181
Palm reqms	20	1	7	5	4.40	.483	2.162
Palm ease	20	1	7	3	3.30	.471	2.105

Table A-2. Perceived usability results for the control participant group.

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	9	5	7	6	5.88889	0.26058	0.78174
PIN ease	9	5	7	6	6.11111	0.26058	0.78174
Finger reqms	9	2	7	7	6.11111	0.53863	1.61589
Finger ease	9	4	7	7	6.55556	0.33793	1.01379
Eye reqms	9	2	7	5	4.77778	0.57198	1.71594
Eye ease	9	2	7	6	5.22222	0.57198	1.71594
Palm reqms	6	3	7	5.5	5.33333	0.55777	1.36626
Palm ease	6	1	7	3.5	3.66667	0.88192	2.16025

Table A-3. Perceived usability results for the hearing loss participant group.

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	11	5	7	6	6.18182	0.26348	0.87386
PIN ease	11	4	7	6	6	0.35675	1.18322
Finger reqms	11	5	7	7	6.54546	0.2473	0.8202
Finger ease	11	5	7	7	6.72727	0.19498	0.64667
Eye reqms	9	3	7	7	6.33333	0.44096	1.32288
Eye ease	9	1	7	7	6	0.66667	2
Palm reqms	9	1	7	5	4.11111	0.78959	2.36878
Palm ease	9	1	6	3	3	0.60093	1.80278

Table A-4. Perceived usability results for the vision loss participant group.

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	7	6	7	7	6.71429	0.18443	0.48795
PIN ease	7	5	7	6	6.28571	0.28571	0.75593
Finger reqms	7	6	7	7	6.71429	0.18443	0.48795
Finger ease	7	6	7	7	6.71429	0.18443	0.48795
Eye reqms	7	1	7	4	3.57143	0.97241	2.57275
Eye ease	7	1	7	3	3.57143	1.04328	2.76026
Palm reqms	5	1	7	3	3.8	1.15758	2.58844
Palm ease	5	1	7	2	3.4	1.28841	2.88097

Table A-5. Completion rate results for all participant groups and schemes.

Scheme	Trial Group	N	Mean (%)	Std Error	St Dev
PIN	All participants	54	1.0000	.00000	.00000
	Control	18	1	0	0
	Hearing Loss	22	1	0	0
	Vision Loss	14	1	0	0
Fingerprint	All participants	53	1.0000	.00000	.00000
	Control	18	1	0	0
	Hearing Loss	22	1	0	0
	Vision Loss	13	1	0	0
Eye	All participants	51	.8235	.05391	.38501
	Control	18	0.94444	0.05556	0.23570
	Hearing Loss	18	0.88889	0.07622	0.32338
	Vision Loss	14	0.57143	0.13725	0.51355
Palm	All participants	40	.70	.073	.464
	Control	12	0.75	0.13056	0.45227
	Hearing Loss	17	0.70588	0.11391	0.46967
	Vision Loss	11	0.63636	0.15212	0.50453

Table A-6. Response time results from success trials for all participant groups and schemes.

Scheme	Trial Group	N	Min (sec)	Max (s)	Median (s)	Mean (s)	Std Error	St Dev
PIN	All participants	54	5.02	35.81	8.9225	11.4128	.93534	6.87335
	Control	18	5.016	23.486	8.399	9.32933	1.06617	4.52339
	Hearing Loss	22	5.365	24.497	9.103	10.21586	0.9795	4.59426
	Vision Loss	14	5.731	35.809	15.22	15.96979	2.68522	10.0472
Fingerprint	All participants	54	.82	17.07	3.3765	4.8594	.52495	3.85755
	Control	18	1.47	11.724	2.6625	3.68361	0.62436	2.64894
	Hearing Loss	22	0.815	17.07	3.5935	5.50532	0.93728	4.39621
	Vision Loss	14	1.201	12.299	3.494	5.35679	1.11421	4.16898
Eye	All participants	42	1.74	108.67	7.686	13.7045	3.10427	20.11794
	Control	17	1.741	56.357	6.746	9.33606	3.00964	12.40908
	Hearing Loss	16	3.737	108.669	8.1175	17.40663	6.87365	27.49460
	Vision Loss	8	3.435	56.526	9.931	16.4855	6.17114	17.45461
Palm	All participants	28	1	81	8.346	18.24	3.976	21.039
	Control	9	0.928	33.094	3.981	8.94544	4.04064	12.12193
	Hearing Loss	13	3.483	57.069	12.058	22.51715	5.54965	20.00953
	Vision Loss	6	2.298	80.993	8.2025	22.89633	12.68421	31.06984

Appendix B Abbreviations and Acronyms

21st Century IDEA	21st Century Integrated Digital Experience Act
FIDO	Fast Identity Online
GOMS model	Goals, Operators, Methods, Selection Rules model
IRB	Institutional Review Board
ISO	International Organization for Standardization
NCCoE	National Cybersecurity Center of Excellence
NIST	National Institute of Standards and Technology
PMA	President's Management Agenda
SP	Special Publication
SSA	Social Security Administration
SUS	System Usability Scale
UAF	Universal Authentication Framework
UMUX	Usability Metric for User Experience

This page intentionally left blank.

This page intentionally left blank.