How Old Are You?: Towards Identifying Measurable Cognitive Phenomena for Online Age Verification

Ahmad Alsaleem
University of Utah
ahmad.alsaleem@utah.edu

Roger Altizer, Jr.
University of Utah
roger.altizer@utah.edu

Tamara Denning
University of Utah
tdenning@cs.utah.edu

ABSTRACT

Electronic age verification is necessary for keeping children out of adult content online and keeping adults out of areas such as children-only chatrooms. There is a need for more spoof-resistant, logistically feasible, and privacy-respecting methods for age verification. In this paper we identify a series of cognitive phenomena that have been shown to demonstrate stratified performance based on age: interference, multisensory responses, and working memory. One or more tests based on these phenomena—or variations on these tests—could be used as a new form of age verification online.

1. INTRODUCTION

Age verification is the process of answering a basic question about an individual's identity: whether or not they are over or under a certain age threshold. Age verification is widely used throughout the web. One common application is verifying the age of the user before they enter an adult site on the internet (e.g., social networks, sexual content, violent content, gambling).

Common age verification tools ask the user to confirm that they are age-appropriate for the content of the site, to provide their birth date, or to submit information on a working credit card. Such techniques result in many problems with unauthorized access to content. The Internet Safety Technical Task Force [32] report that in 2006 13% of youth online received sexual solicitations. A recent NY Times article [21] describes multiple incidents of sexual assault that resulted from adults entering social groups intended to be used only by minors. The authors linked such problems to the fact that age verification tools are easy to fool. They cited multiple attempts to create fool-proof age verification tools. Such proposals have included establishing a (United States) national identity database that can be used to identify the age of the user. Such an approach, however, requires significant resources and effort to implement and is a major concern in terms of user privacy. Other approaches such as hand radiography exist; however, the hardware required for such a method is logistically infeasible for online age verification.

We seek to improve the state of age verification tools by identifying candidate approaches that are **resistant to spoofing**, **logistically feasible**, and **privacy-respecting**. In particular, we studied the literature from other disciplines to identify cognitive phenomena that have demonstrated stratified performance based on age.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

In the following sections we discuss related work and then describe the candidate tests. The tests are based on the following phenomena:

- Interference.
- Multisensory responses.
- Working memory and cognitive load.

While the supplied tests have all demonstrated differences in average performance based on age, further research remains to determine whether the tasks are suitable for age verification on a per-user basis. It is our hope that this work will help advance the state of the art for age verification.

2. RELATED WORK

Below we discuss both deployed methods of age verification and methods for determining age from academic research.

Age Verification in Practice. As mentioned previously, current deployed age verification systems use approaches such as showing checkboxes or confirmation dialogues, asking users to confirm their birth dates or birth years in order to calculate age, or requiring users to enter valid credit card information, thereby implying that they are either old enough to use a credit card or have the consent of someone who does. Systems like AgeID collect name, address, telephone number, and date of birth in order to further verify that the identity actually exists [1]. Other systems attempt to verify that the user belongs to the supplied identity by sending a code via SMS to their phone or by asking Knowledge-Based Authentication (KBA) questions [33]. Some sites instead require that users submit copies of their passports or driver's licenses.

All of the above approaches are spoofable and/or require the user to supply private information. Furthermore, the above systems are all about a user demonstrating that they are above a certain age threshold (rather than under it).

There are also methods for verifying that adults are excluded from (or detected in) children's areas; however, they suffer from similar weaknesses. For example, just as we can verify an adult's age by collecting enough information to check their identity, we can also verify a minor's identity. Some of the databases used for such purposes, however, do not contain minors (e.g., voter registration databases, credit bureau databases). Instead, we would either need to use a preexisting database suitable for the purpose (e.g., some countries' identity databases, school databases [7]) or construct a new one for the purpose. Privacy (and in some cases logistics) concerns remain relevant here.

According to Facebook's submission to the ISTTF [32], they implemented a peer verification system to establish high school affiliation for any new Facebook user under the age of 18. The system works by sending questions to other Facebook users who are already verified and belong to the same high school as the new user. Based on the answers the user account will be verified or disabled. Depending on the details of how such a system is handled, it is potentially susceptible to Sybil or social engineering attacks.

Another approach, like that taken by the app Spotafriend, is to use computer vision to determine that a user is under 18 [31]; however, there have been reports that this system is either ineffective or unenforced (e.g., [28]).

All of the above methods are age verification for server-side blocking. Content can also be controlled via client-side blocking such as from parental controls and web filters. We choose not to focus on this model, however; it only prevents children from accessing adult content on pre-configured machines, and it does not prevent adults from accessing areas intended only for minors.

Age Verification in Research. There is a related body of research that demonstrates different potential ways to determine age. For example, researchers have used the different acoustic and prosodic features in the speech to determine the speaker's age (e.g., [11, 14]). There are tools such as the one from Microsoft that determine age from a photograph [5]; however, reports from users have been mixed (e.g., [16, 20]). A similar system, Amazon Rekognition, has been introduced by Amazon [2]. We note that the tool output can produce a vast range (e.g, 17 years) for the user's predicted age [3]. Other research can determine age from hand radiography (i.e., the bone structure of the hand) [24]. Such a system has logistically infeasible hardware requirements for online age verification.

3. CANDIDATE AGE VERIFICATION TESTS

We considered the literature from other disciplines in order to locate candidate phenomena/tests that could potentially be used for electronic age verification. In order to be used for age verification, the tests need to produce results that are accurately stratified based on users' ages. Additionally, we were looking for tests that are:

- Resistant to spoofing.
- Logistically feasible.
- Privacy-respecting.

Our particular interest in cognitive phenomena is because we are seeking tests whose results cannot be faked. Being logistically feasible means, for example, that an age verification technique does not take an excessive amount of time and can be performed on reasonable hardware in a normal usage context. The different proposed tests vary in terms of their logistical requirements. For example, some tests might take longer (e.g., up to 2 minutes) and others have more hardware requirements (e.g., speakers or headphones). Table 1 gives an overview of example performance numbers based on age for some of the described tests.

3.1 Interference: The Stroop Test

The Stroop test measures selective attention capacity and processing speed skills [13]. The standard Stroop test consists of three sections, in which the participant is presented with: (A) cards showing the English words for colors (e.g., `blue`), (B) cards showing colors (e.g,), and (C) color-word cards [9]. The color-word cards are a mix between incongruent and congruent cards. On congruent cards, participants are presented with the words for colors, which are printed in their respective color (e.g., `blue`). The participant is asked to name the printed color of the word (e.g., "blue"). On incongruent cards, participants are presented with the words for colors which are printed in colors that *do not match* the printed word (e.g., `blue`). The participant is asked to name the *printed color* of the word (e.g., "red") and to ignore the *printed word*. Participants are asked to work through the color-word cards as quickly as possible while still retaining accuracy. Unsurprisingly, participants perform more quickly and more accurately on the congruent cards than on the incongruent cards; the incongruency between the printed word and the printed color creates interference that negatively impacts performance on the task.

There are many variations of the Stroop test: for example, variations where the stimuli are presented in grids rather than on individual cards, variations with different numbers of colors, variations where drawings are presented instead of words, and electronic variations; however, many of the result trends persist across different Stroop versions.

Many publications show that there are differences in performance on the Stroop test between children and adults, both in terms of speed and accuracy (e.g., [15, 22, 23, 27, 34]). Table 1 shows the mean response time (in ms) per item and the standard error for congruent items and incongruent items in task C (see above) of a computerized version of the Stroop test [34]. The performance numbers are given for children (ages 9–13, N=11) and adults (M=29.35, SD=5.45, N=20). Table 1 also gives the mean response time for another publication's results on a (standard) paper version of task A, task B, and the incongruent items from task C of the Stroop test [22]. The results, which are for the time required to complete a sequence of 48 consecutive items, are given in seconds. The paper breaks down the children's results by ages 7 (N=24), 8 (N=20), 9 (N=20), 10 (N=25), 11 (N=29), 12 (N=25), and 13 (N=29). In Table 1 we show the results for ages 17–19 (N=18) in the 'adult' column; the paper also reports response times for ages 25–34 (N=14), ages 35–44 (N=16), and ages 65–80 (N=15). As with other studies, the authors found that children do not perform as quickly as adults and that incongruent task C items cause the greatest difference between childrens' and adults' response times.

The results from these publications show that older participants perform more quickly on average than younger participants, particularly on the incongruent task. An electronic age verification test could require the user to complete incongruent (and potentially congru-

Table 1: Example reported differences from Stroop tests and multisensory response tests based on age. Ages in the 'adult' column vary by publication; more details are supplied in the text. For the Stroop C, Computer tests (Congruent and Incongruent) we report the mean response time per item and the standard error. For the Simple RT Task, the Discrimination RT Task, and the Choice RT Task we report the mean response time and the standard deviation. The reported numbers for the Simple RT, Audio Only task are approximated based on the publication's graph.

Age (years)	4	5	6	7	8	9	10	11	12	13	adult
Stroop C, Computer, Congruent Only, Per Item (ms) [34]	-	-	-	-	-	613 (15.92)					582.4 (11.81)
Stroop C, Computer, Incongruent Only, Per Item (ms) [34]	-	-	-	-	-	814.36 (22.8)					711 (16.91)
Stroop A, Paper, Per 100 Items (secs) [22]	-	-	-	89.8	77.6	68.5	62.3	55.6	59.3	54.1	40.5
Stroop B, Paper, Per 100 Items (secs) [22]	-	-	-	126.9	108.3	100.9	92.8	82.1	86.4	79.5	56.1
Stroop C, Paper, Incongruent Only, Per 100 Items (secs) [22]	-	-	-	264.7	208.3	191.4	184.3	160.8	157.9	147.6	103
Simple RT Task (ms) [12]	740 (162)	580 (144)	467 (85)	-	-	-	-	-	-	-	270 (31)
Discrimination RT Task (ms) [12]	1790 (581)	1198 (254)	949 (139)	-	-	-	-	-	-	-	449 (51)
Choice RT Task (ms) [12]	2485 (783)	1652 (437)	1346 (319)	-	-	-	-	-	-	-	704 (132)
Simple RT Task, Audio Only (ms) [6]	-	-	-	~400			~350			~320	~300

ent) Stroop tasks and compare their performance to thresholds based on age. If such a test were used as the basis for an age verification system, minors would potentially be unable to perform as if they were adults.

3.2 Multisensory Responses

Humans perceive information from the world using multiple senses, which provide the brain with the required information to construct meaningful internal representations. The senses go through multiple development cycles, which in general result in faster response times with age [17]. Additionally, multiple publications have noted that reactions to information from a single sense are faster than when people interpret information from multiple senses (e.g., [8, 18, 26]).

There are multiple cognitive tests to assess multisensory integration. They help monitor mental function across different ages. We can classify the tests of interest into the following categories [12]:

- Simple Response Time (RT) tasks: Participants are asked to perform a certain motor function (e.g., press a key) in response to any auditory, visual, or audio-visual stimuli. For example, a participant is asked to press the spacebar as fast as they can when they hear something (e.g., a babbling brook or a chirping bird), see something on the screen (e.g., a picture of a bird or a giraffe), or simultaneously hear and see something.
- Discrimination Response Time (RT) tasks: Participants are asked to complete a task very similar to the Simple RT task; however, participants are *only* to respond to *particular* stimuli. For example, the participant presses the spacebar only when they hear a babbling brook (but not a chirping bird), when they see a picture of a bird (but not a giraffe), or when they hear and see both simultaneously.
- Choice Response Time (RT) tasks: Participants are asked to perform different motor actions based on the presented stimuli. For example, participants are asked to press 'A' when they hear a babbling brook and 'D' when they hear a bird chirping.

Researchers have reported differences in response times and accuracy for different ages (e.g., [12]). For a Simple RT task, kids (ages 4–6, N=166) had slower response times than adults (M=25.3, N=35) by at least 190ms (on average). Different forms of the Discrimination Response Time tasks had different results, but in the spatial orientation condition, the kids had slower response times than adults by at least 500ms (on average). The Choice Response Time tasks also had different results with different forms of the tests, but for the four-choice task, the kids had slower responses than adults by at least 650ms (on average). Table 1 gives mean response times per age (ms) and standard deviation.

Table 1 also shows the results from auditory stimuli only for a Simple RT Task [6]. The numbers in the table are estimated from the publication's graph. The participants were from the following groups: 7–9 years (N=17), 10–12 years (N=15), 13–16 years (N=17), and adults (M=23.11, N=13). The paper reported a significant improvement in the reaction time between the first group (7–9) and the second group (10–12). The improvement continues with subsequent age groups. Adults in the study were also on average more accurate when responding to all types of stimuli.

It is also relevant that the response times for auditory stimuli are faster than the response times to visual stimuli and that the differences between response times to auditory and visual stimuli are more pronounced among younger populations [12, 29].

The results from these publications show that older participants perform more quickly on average than younger participants on Simple RT tasks, Discrimination RT tasks, and Choice RT tasks (most particularly on Choice RT tasks). Furthermore, children exhibit a larger difference between their response to auditory and visual stimuli than adults. Using some combination of these tasks, an electronic age verification test could compare a user's performance (absolute or relative) against thresholds based on age. If such a test were used as the basis for an age verification system, minors would potentially be unable to perform as if they were adults.

3.3 Working Memory: Digit Span

The digit-span test addresses the short-term working memory of participants. Participants are presented with a series of digits (e.g., "8, 3, 4") and must immediately repeat them back. If they do this successfully, they are given a longer list (e.g., "9, 2, 4, 0"). The length of the longest list a person can remember is that person's *digit span*.

There are multiple variations of the test. For example, the participant might be presented with pictures instead of digits (and asked to recall the names of the objects in the pictures) or might be asked to write the digits down instead of verbally repeating them [10, 25].

In a study of particular interest, the participant is equipped with an eye tracker during their digit-span test in order to use the changes in their pupil dilation to measure their fluctuating cognitive load [10]. The paper shows that adults (ages 18.3–60.8, N=54) recalled 12% more digits than children (ages 7.5–14.0, N=69). Adults showed an increase in pupil dilation with each increase in digit sequence length up until length 7, while children stopped showing significant increase in dilation at sequence of length 5. At sequences of lengths 7, 8, and 9, adults' pupils either continued to dilate or reached their maximum dilation; at those sequence lengths, childrens' pupils began to constrict.

The results from these papers show that adult participants have higher digit spans than children; additionally, the pupils of adults and children exhibit different dilation behavior during a digit-span task. A digit-span test can potentially work as an age verification tool to keep children out of adult content; a digit-span test with an eye tracker would potentially have better performance. A digit-span test with an eye tracker would also potentially be able to *keep adults out of children-only areas.*

4. DISCUSSION

The tests that we describe in this paper are from disciplines like psychology. While we had several criteria when identifying candidate tests, the primary criterion was that prior publications using the test reported differences in results based on age. The typical purpose of a study we cite was to help clinicians set standards for normal cognitive performance for a given age group. As a result, the studies tend both not to comprehensively cover all age groups and to report their results as averages for a given age group. Unfortunately for our purposes, the publications do not give much data on how individual performance varies. While we have identified tests with potential, more data needs to be gathered in order to understand whether such tests are suitable for age verification in practice—most particularly in order to understand whether such systems would have acceptable false positive/false negative rates with users who are outliers. Additionally, such systems would need to account for older adults; on some cognitive tests the performance of older adults starts to lower, and would thus start to resemble the performance of a minor. Similarly, a robust age verification system would need to consider how to accommodate individuals with varying levels of cognitive performance that are not neurotypical.

Some variations of the digit-span test that we describe—more particularly the variation that could potentially block adults from children-only areas—rely upon the use of an eye tracker for measuring pupil dilation. With current technology this would most likely either be achieved via a dedicated eye tracker, which is logistically unlikely, or, more likely, via a VR headset (e.g., [4]). In the future such a measurement might be achieved via a laptop webcam or a front-facing cell phone camera (e.g., [30]).

While we identified multiple tests that rely upon different underlying cognitive phenomena, our set of tests is not comprehensive. There are certainly variations on the identified tests that have potential for applications to age verification. For example, the Corsi block-tapping test, which bears some resemblance to the digit-span test, is a cognitive test that assesses visuospatial short-term working memory [19]. The computerized version of the Corsi task requires users to select, in order, squares that have been highlighted on the screen. The longest sequence that a participant can repeat is their *CB span.* Multiple studies have reported different scores between different ages; however, it remains to be seen whether these results persist when the positioning of the squares and the sequences to repeat are randomized (instead of fixed). Furthermore, the results of such a test are not as granular as, for example, the timing results on the Stroop or RT tasks. The Corsi block-tapping test, like the digit-span test, would most likely work best when combined with measurement of cognitive load via pupil dilation.

We acknowledge that such tests are not fully immune to cheating or fooling by users. Potential outliers outside, in most of our identified tests minors *should* be incapable of performing as well as adults. That being said, not all of the phenomena/tests have been studied to determine whether training can improve performance. Additionally, for most tests (digit span with pupil dilation measurement aside), adults could potentially intentionally downgrade their performance in order to resemble a child in order to gain access to children-only areas. Conversely, it might be difficult to downgrade performance just enough to resemble a realistic child without setting off any spoofing alerts, which would be looking for unrealistic performance. This is a matter that requires further study. Pupil dilation, as a physiological response, is generally involuntary, although it can be controlled in some circumstances by conditioning or drugs.

We are hopeful that it will be useful to combine multiple tests in order to increase the performance of an age verification tool, both in terms of discriminating between different age users and in terms of detecting users attempting to fool the system.

5. CONCLUSION

Age verification is an important problem that has real-world consequences, whether the goal is to keep children out of adult content or adults out of children-only areas. Age verification is a complex problem that has no current good solution. More particularly, we have no current solutions that are resistant to spoofing, logistically feasible, and privacy-respecting.

In this paper we present multiple tests as potential candidates off of which to base an age verification system. These tests are from disciplines such as psychology and have demonstrated stratification in performance based on age. The tests are based off of the cognitive phenomena of interference, multisensory responses, and working memory/cognitive load.

We describe the Stroop tests, the Simple, Discrimination, and Choice Response Time tasks, and the Digit Span task. We provide some references to and summaries of results from prior publications which demonstrate differences in performance based on age. We discuss related issues such as the need for reporting data at the level of individuals, the hardware requirements for measuring pupil dilation, and the potential for spoofing the tests.

We hope that this work serves as a foundation for future work into using cognitive (and physiological) tests as the basis for spoof-resistant, logistically-feasible, and privacy-respecting systems for age verification.

6. REFERENCES

[1] AgeID. https://www.ageid.com/.

[2] Amazon rekognition. https://aws.amazon.com/rekognition/.

[3] Amazon rekognition update – estimated age range for faces. https://aws.amazon.com/blogs/aws/amazon-rekognition-update-estimated-age-range-for-faces/.

[4] Fove headset. https://www.getfove.com/.

[5] Microsoft how-old tool. https://how-old.net/.

[6] A. B. Brandwein, J. J. Foxe, N. N. Russo, T. S. Altschuler, H. Gomes, and S. Molholm. The development of audiovisual multisensory integration across childhood and early adolescence: A high-density electrical mapping study. *Cerebral Cortex*, 21(5), 2011.

[7] Eguardian. Eguardian site. http://eguardian.com.

[8] B. Forster, C. Cavina-Pratesi, S. Aglioti, and G. Berlucchi. Redundant target effect and intersensory facilitation from visual-tactile interactions in simple reaction time. *Experimental brain research*, 143, 2002.

[9] C. Golden. Stroop color and word test.

[10] E. Johnson, A. M. Singley, A. Peckham, S. Johnson, and S. Bunge. Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology*, 5, 2014.

[11] F. Kelly, A. Drygajlo, and N. Harte. Speaker verification in score-ageing-quality classification space. *Computer Speech & Language*, 27(5), 2013.

[12] S. Kiselev, K. A. Espy, and T. Sheffield. Age-related differences in reaction time task performance in young children. *Journal of Experimental Child Psychology*, 102(2), 2009.

[13] M. J. M. Lamers, , A. Roelofs, and I. M. Rabeling-Keus. Selective attention and response set in the stroop task. *Memory & Cognition*, 38(7), 2010.

[14] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech Language*, 27(1), 2013.

[15] C. Ludwig, E. Borella, M. Tettamanti, and A. de Ribaupierre. Adult age differences in the color stroop test: A comparison between an item-by-item and a blocked version. *Archives of Gerontology and Geriatrics*, 51(2), 2010.

[16] A. Lynch. How-old.net is a new website that guesses people's ages and it's really, really bad at it, 2015. https://metro.co.uk/2015/05/01/how-old-net-is-a-new-website-that-guesses-peoples-ages-and-its-really-really-bad-at-it-5176246/.

[17] J. Mahoney, C. Li, M. Oh-Park, J. Verghese, and R. Holtzer. Multisensory integration across the senses in young and old adults. *Brain research*, 1426, 2011.

[18] S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder, and J. J. Foxe. Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, 14(1), 2002.

[19] K. F. Pagulayan, R. M. Busch, K. L. Medina, J. A. Bartok, and R. Krikorian. Developmental normative data for the corsi block-tapping task. *Journal of Clinical and Experimental Neuropsychology*, 28, 2006.

[20] S. Panwar. Proof: Microsoft's how old do i look app is wildly inaccurate, 2015. https://www.hindustantimes.com/tech-reviews/proof-microsoft-s-how-old-do-i-look-app-is-wildly-inaccurate/story-43ppsz3CJ8ejYiVFYaWyWM.html.

[21] N. Perlroth. Verifying ages online is a daunting task, even for experts, 2012. https://nyti.ms/2rLJFBk.

[22] S. W. Peter E., Comalli Jr. and H. Werner. Interference effects of stroop color-word test in childhood, adulthood, and aging. *The Journal of Genetic Psychology*, 100(1), 1962.

[23] V. Pritchard and E. Neumann. Avoiding the potential pitfalls of using negative priming tasks in developmental studies: Assessing inhibitory control in children, adolescents, and adults. *Developmental psychology*, 45, 2009.

[24] F. Remy, G. Hossu, R. Cendre, E. Micard, L. Mainard-Simard, J. Felblinger, L. Martrille, and L. Lalys. Development of a biometric method to estimate age on hand radiographs. *Forensic Science International*, 271, 2017.

[25] A. Roman, D. Pisoni, and W. Kronenberger. Assessment of working memory capacity in preschool children using the missing scan task. *Infant and Child Development*, 23, 2014.

[26] L. A. Ross, D. Saint-Amour, V. M. Leavitt, S. Molholm, D. C. Javitt, and J. J. Foxe. Impaired multisensory processing in schizophrenia: Deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophrenia Research*, 97(1), 2007.

[27] R. Salo, A. Henik, and L. Robertson. Interpreting stroop interference: An analysis of differences between task versions. *Neuropsychology*, 15, 2001.

[28] L. Schmidt. The danger of teenage dating apps: How i posed as a 15-year-old girl online, 2015. https://globaldatinginsights.com/2015/10/14/14102015-the-danger-of-teenage-dating-apps-how-i-posed-as-a-15-year-old-girl-online/.

[29] J. Shelton and G. Kumar. Comparison between auditory and visual simple reaction times. *Neuroscience & Medicine*, 1, 2010.

[30] J. Y. Shin, D. H. Jo, S. L'Yi, S. Y. Moon, J. H. Bae, and T. S. Lee. Smartphone-based pupillary light reflex test system. In *ICCAS*, 2013.

[31] Spotafriend. Finally, an app that your parents can't use to spy on you, 2018. http://blog.spotafriend.co/2016/12/05/finally-an-app-that-your-parents-cant-use-to-spy-on-you/.

[32] I. S. T. T. F. Team. Enhancing child safety and online technologies, 2009. http://cyber.harvard.edu/pubrelease/isttf/.

[33] E. verification systems. Keeping adult products to adults. https://www.electronicverificationsystems.com/age-verification.

[34] B. C. Wright and A. Wanley. Adults' versus children's performance on the stroop task: Interference and facilitation. *British Journal of Psychology*, 94(4).