# Using a Memory Test to Limit a User to One Account

Vincent Conitzer

Department of Computer Science
Duke University
Durham, NC, USA
conitzer@cs.duke.edu

**Abstract.** In many Web-based applications, there are incentives for a user to sign up for more than one account, under false names. By doing so, the user can send spam e-mail from an account (which will eventually cause the account to be shut down); distort online ratings by rating multiple times (in particular, she can inflate her own reputation ratings); indefinitely continue using a product with a free trial period; place shill bids on items that she is selling on an auction site; engage in false-name bidding in combinatorial auctions; *etc.* All of these behaviors are highly undesirable from the perspective of system performance. While CAPTCHAs can prevent a bot from automatically signing up for many accounts, they do not prevent a human from signing up for multiple accounts. It may appear that the only way to prevent the latter is to require the user to provide information that identifies her in the real world (such as a credit card or telephone number), but users are reluctant to give out such information.

In this paper, we propose an alternative approach. We investigate whether it is possible to design an automated test that is easy to pass once, but difficult to pass a second time. Specifically, we design a memory test. In our test, items are randomly associated with colors (*"Cars are green."*). The user first observes all of these associations, and is then asked to recall the colors of the items (*"Cars are...?"*). The items are the same across iterations of the test, but the colors are randomly redrawn each time (*"Cars are blue."*). Therefore, a user who has taken the test before will occasionally accidentally respond with the association from the previous time that she took the test (*"Cars are...? Green!"*). If there is significant correlation between the user's answers and the correct answers from a previous iteration of the test, then the system can decide that the user is probably the same, and refuse to grant another account. We present and analyze the results of a small study with human subjects. We also give a game-theoretic analysis. In the appendix, we propose an alternative test and present the results of a small study with human subjects for that test (however, the results for that test are quite negative).

# 1 Introduction

Many Web-based applications require a user to sign up for an account first. Because of the anonymity that the Internet provides, it is typically not difficult for a single user to sign up for multiple accounts under fictional identities. Doing so can provide many benefits to the user, including at least the following:

1. The user can send spam e-mail from the fictional accounts. The service provider will typically realize that this is happening fairly quickly and shut down the account, but then the user can simply sign up for another account.
2. In online rating systems, the user can rate the same object many times and thereby distort the aggregate rating. This is especially valuable when the object being rated is a product that the user is selling, or when the object is the user's own reputation (on, say, an auction site).
3. When a product has a free trial period, the user can indefinitely continue using the product at no cost: once the trial period expires, she can simply start using the product under a different account.
4. In an online auction, the user can use another account to place shill bids on the items that she is selling, thereby driving up their selling prices.
5. In more complex economic mechanisms such as *combinatorial auctions*, in which multiple items are simultaneously for sale (for an overview, see [1]), it is often possible to obtain a bundle of items at a lower price by bidding under multiple identities [12, 13]. It is possible to design mechanisms for which using multiple identities is not beneficial [12, 10, 13, 11], but these are less efficient.
6. In online poker, the user can try to play on the same table under two or more distinct identities, allowing her to effectively collude with herself.

While all of the above behaviors are beneficial to the user who engages in them, they reduce the performance of the system as a whole. Users have to deal with potentially large volumes of unwanted e-mail, online ratings become meaningless, companies become reluctant to offer free trial periods, auction mechanisms become less efficient, people become reluctant to play poker online[1], *etc.* As a result, it may well be that *all* users, including those who choose to engage in the behavior, would prefer it if this behavior was impossible.[2]

In some cases, a user would benefit from owning a very large number of accounts. For example, if the accounts are used to send spam e-mail, then the service provider is likely to shut down the account as soon as it realizes that the account is being used to send spam; hence, many accounts are necessary to send out a significant amount of spam. In cases such as these, the user (spammer) may

[1] Given online poker's murky legal status, one may debate whether this is a good or a bad thing.

[2] Game theory (for overviews, see [2, 5, 6]) provides many other examples where agents would prefer it if their most preferred actions were made unavailable, given that those actions are also made unavailable to the other agents: consider defection in the Prisoner's Dilemma, overgrazing in the Tragedy of the Commons, *etc.*

try to use a computer program, or *bot*, that repeatedly registers for an account. This (along with other applications) motivated the development of *CAPTCHAs* (*Completely Automated Public Turing Tests to Tell Computers and Humans Apart*) [8, 9], which are automated tests that are easy to pass for humans, but difficult to pass for computers. A well-known CAPTCHA is *Gimpy*, where the task is to read distorted text. Indeed, Gimpy is now widely used to screen out bots. It should be noted that several variants of Gimpy have been broken, that is, programs have been written that succeed on a large fraction of instances of the test [3, 7, 4]. This arguably represents a significant advance in computer vision. In fact, AI researchers should hope that *every* CAPTCHA that is designed will eventually be broken, since otherwise the CAPTCHA would represent a limit to artificial intelligence (more precisely, to the artificial intelligence that we as humans can create). But when a CAPTCHA is broken, we can in principle switch to using a different CAPTCHA, as long as artificial intelligence does not yet match human intelligence.

Unfortunately, CAPTCHAs are of little use in preventing a *human* from signing up for multiple accounts. Given how little revenue a spammer obtains from a single account, it is perhaps not economically feasible for a spammer to solve sufficiently many CAPTCHAs herself (or to hire people to do it for her). However, for all of the other uses for multiple accounts that we mentioned, only a few accounts are required. So, how can we prevent a human from signing up for multiple accounts? One possibility is to require her to provide information from which her identity in the real world can be established—for example, a credit card number or a phone number.[3] However, users tend to be very unwilling to provide such information, among other reasons because doing so entails giving up the privacy and anonymity that the Internet affords. Another possibility is to charge a price for each account (assuming that payments can be made anonymously), but again, Web users are notoriously unwilling to make payments. Also, if the payment is small enough, then the user may still want to sign up for multiple accounts.

It may seem that if account registrations are completely anonymous, and a user can sign up for one account, then she can always sign up for a second account in the same way. In this paper, we argue that this is not necessarily the case. We investigate whether it is possible to design an automated test that is easy to pass once, but difficult to pass a second time. The idea that we pursue is to have the user be affected by taking the test the first time, in a way that is detectable when she takes the test again. Specifically, we design a memory test. In this test, the user is asked to memorize and then recall a number of (item, color) associations. Across iterations of the test, the items are always the same, but the color associated with each item is randomly redrawn each iteration. Because of this, a user taking the test a second time is likely to get confused and occasionally respond with the association from the first time that she took the

[3] One way to sign up for a Gmail$^{\text{TM}}$account is to submit a mobile phone number, to which an invitation code is then sent. This is explicitly to prevent one person from signing up for many accounts.

test. (This is related to the *proactive interference* phenomenon in psychology, where old memories interfere with the learning of new memories. However, this term typically refers to the decrease in performance on the later iteration of the test, rather than to the overlap in answers with the earlier iteration of the test.) Thus, if there is significant correlation between the user's answers and the correct answers from a previous iteration of the test, then the system can decide that the user is probably the same, and refuse to grant another account. The system must also refuse to grant the account if the user recalls too few associations correctly: otherwise, the user can just respond randomly, and thereby avoid confusion and overlap. Several other minor modifications are necessary to make the system work. For one, memory tests are easy to pass for computers; therefore, the test must be integrated with a CAPTCHA (in such a way that the CAPTCHA cannot simply be separated and given to a human). For example, the (item, color) pairs can be distorted as in Gimpy. Also, the test must be run at a speed that makes it infeasible for the user to write down and look up the associations.

In the remainder of this paper, we first present the specifics of the automated test that we designed. We then present the results of a small, formal study on human subjects. We also present a game-theoretic model of the test, and analyze the effects of different strategies for the user. Finally, we discuss future research. In the appendix, we discuss another (less effective) automated test that we designed, as well as the results of a small study on human subjects for that test.

## 2 Test specifics

The specifics of the test that we designed are as follows. (The source code is available upon request.) There are 100 items in the test, which were chosen with a bias towards items that do not naturally have a color associated with them. (*E.g.*, "cars" was one of the items, "grass" was not. Of course, cars are still more associated with red than with pink; it seems impossible to avoid such association altogether.) There are 8 colors in the test: red, green, blue, yellow, white, black, orange, and pink. At the beginning of an iteration of the test, each item is randomly associated with one of the colors (*e.g.*, "Cars are green."). Each of these associations is then displayed to the user for 4 seconds (in random order). After all of the associations have been displayed, each of the items is displayed to the user for 3 seconds (in a different order[4]), during which the user has to recall the associated color. Thus, the total duration of the test is 700 seconds (11.7 minutes). Clearly, the length of the test makes it somewhat unattractive to take, but with fewer items we are unlikely to be able to recognize correlation with a previous iteration of the test (with statistical significance). In principle, a user would have to take the test only once, to obtain a master account which she can then use to sign up for other accounts. Besides the number of items, the

[4] Changing the order forces users to associate colors with items, rather than just remember a sequence of colors. It also makes it difficult to write down and look up the associations in time.

other parameters are the number of colors, the amount of time each association is displayed, and the amount of time given to recall each item's color. Based on some informal experiments, these parameters were set to make the test difficult but not impossible, as well as to keep its length reasonable.

## 3 A small study with human subjects

We proceeded to conduct a small, formal study with human subjects, whose details and results we describe in this section. Each subject first did a practice run with a version of the system with only 10 items, which do not overlap with the "official" 100 items. Then, the subject did two full iterations of the main test (with the same 100 items, but re-randomized colors for each item). Each subject was compensated US $7, plus US $7 times the percentage of correct answers given in the two iterations of the main test. (Given that the test is somewhat exhausting, it was considered important to reward subjects for good performance, to keep them engaged.) Subjects were recruited by posting flyers. In the end, we obtained data from 7 subjects, all of whom are students at Duke University. The study was approved by the Institutional Review Board. It is very small, but it is enough to illustrate the key phenomena. Earlier, informal tests produced similar results.

Before presenting the results, it is useful to consider what results would indicate that our system is effective. Ideally, we would see: 1. high scores (percentage correct) for the first iteration (so that a user can obtain an account), and 2. either low scores for the second iteration, or significant overlap between the answers given by the user in the second iteration and the correct answers in the first iteration (so that a user will fail to obtain a second account, either because her performance is too poor or because the system can link her to her first attempt). We do *not* want to consider the overlap between given answers in the second iteration and *given* answers in the first iteration, because it is likely that there would be significant overlap between given answers even for two different users—for example, because people tend to answer "white" more often, or because they tend to answer "red" for cars, *etc.* However, the probability of giving the answer that was the *correct* answer for another iteration of the test, *given that the user never saw the correct answers for that iteration*,[5] is exactly $1/n_c$ (where $n_c$ is the number of colors, 8 in our case), because the correct answers are randomly drawn. Thus, if the overlap between the given answers in one iteration of the test, and the correct answers in another iteration, is significantly greater than $1/n_c$, then we can be reasonably sure that the same user was involved in both iterations.

Unfortunately, it appears inevitable that some users will do poorly on their first iteration. If this happens, then we no longer have the same goal for the second iteration: if anything, we would like them to do *better* on their second iteration, since they would have failed to obtain an account the first time. Also, in

[5] ... or was otherwise (indirectly) influenced by the correct answers for that iteration.

this case, it is unreasonable to expect the correct answers from the first iteration to overlap much with the given answers in the second iteration, since these correct answers did not even overlap much with the given answers in the *first* iteration! So, to prevent users from signing up for multiple accounts, the key requirement is that people that perform well in the first iteration either perform poorly in the second iteration, or that their given answers in the second iteration have significant overlap with the correct answers from the first iteration. We are now ready to present the results of the study.

| **subject** | $\mathbf{c_1 = a_1}$ | $\mathbf{c_2 = a_2}$ | $\mathbf{c_1 = a_2}$ | $\mathbf{c_2 = a_1}$ |
|---|---|---|---|---|
| **1** | 53 | 66 | 27 | 13 |
| **2** | 34 | 23 | 14 | 13 |
| **3** | 31 | 46 | 13 | 13 |
| **4** | 50 | 61 | 23 | 14 |
| **5** | 17 | 38 | 15 | 11 |
| **6** | 42 | 43 | 11 | 11 |
| **7** | 60 | 70 | 22 | 12 |

Table 1: Experimental results for human subjects.

In the results, $c_i$ stands for the correct answer in the $i$th iteration, and $a_i$ for the given answer in the $i$th iteration ($i \in \{1, 2\}$). Thus, the sequence $c_1 = a_1$ gives the score for the first iteration, $c_2 = a_2$ gives the score for the second iteration, $c_1 = a_2$ indicates how often the given answer in the second iteration is identical to the correct answer (for that item) in the first iteration (indicating the level of confusion that the subject experienced, and the extent to which the system can identify the subject as the same person that performed the first iteration), and $c_2 = a_1$ indicates how often the given answer in the first iteration is identical to the correct answer (for that item) in the second iteration. For the last sequence, the probability that these answers match is always 1/8 (because the second iteration's correct answer is drawn randomly after the user's answer has been given in the first iteration), so unsurprisingly, this sequence is closely clustered around $12.5 = 100/8$. (Had this not been the case, it could only have been due to a statistical fluke, a mistake in the experimental setup, or a failure of the random number generator.) So we focus on the three remaining sequences.

Unfortunately, not all the subjects do well on the first iteration. Thus, if we require a reasonably high score on the test, some users will be denied an account on their first attempt. However, in all but one case, performance improved on the second attempt. If we look at the three subjects who performed best (1, 4, and 7), we see, encouragingly, that their overlap ($c_1 = a_2$) is very high (27, 23, 22, respectively). The probability that an overlap of at least 22 would have occurred if the two iterations of the test were taken by different people (so that the probability of overlap on any individual answer would be 1/8) is only $\sum_{i=22}^{100} \binom{100}{i} (1/8)^i (7/8)^{100-i} = 0.0056$, so we can reject the second account application in these cases.[6] (Here, we are in some sense evaluating the cutoff of

[6] It should be noted that in a real system, we must compare the answers not just to *one* specific previous iteration of the test, but to *every* previous (successful) iteration

22 on the same data as the data on which we based this cutoff; for a more thorough evaluation, it would be desirable to have a separate training set, on which we base the cutoff, and test set, on which we evaluate the cutoff.) However, the fourth-best performer, subject 6, displayed no overlap at all in spite of performing reasonably well on both iterations. We conjecture that there are different memorization strategies that subjects used, and that while the most successful strategies tend to produce significant overlap, there are other strategies that are still somewhat successful and less prone to cause overlap. For example, a subject can split the colors into two sets of four each, restrict attention to colors in the first set in the first iteration, and to colors in the second set in the second iteration. Fortunately, such strategies will fail if the user is required to recall a large enough number of associations correctly.

## 4 A game-theoretic analysis

Subjects were cautioned that the items in the second iteration would be the same as in the first iteration, with potentially different color associations, so that they should try to take care not to get confused. In reality, however, the reward structure of the study did not penalize subjects for giving an answer in the second iteration that was the correct answer in the first iteration (at least not more than it penalized them for giving any other wrong answer). Since the idea is to deny the request for an account if there is too much overlap with the correct answers from a previous iteration, an ideal study would have penalized subjects for such overlap; this perhaps would have made subjects more careful to avoid it. We chose not to pursue such a design for the study for the following reasons. First, it is *ex ante* not clear by how much to penalize subjects. Perhaps the most convincing design would have been to set strict criteria beforehand for when a subject "passed" the test (*i.e.*, would be awarded an account), and to pay subjects in proportion to the number of accounts that they obtained. However, this would have required us to set the requirements for passing the test before collecting any formal data. Moreover, the lack of any "partial credit" may have made it more difficult to attract subjects. A second reason for the design of our study is that by using a game-theoretic model that we present next, we can use the results of the study to infer what results a subject could have obtained by changing her strategy (assuming correctness of the model). The test designer and the user play a game where the designer sets criteria and the user subsequently tries to obtain multiple accounts.

of the test. If the number of users is large, then the probability of this much overlap occurring by chance in at least one of these comparisons is significant. For example, if nobody is trying to obtain multiple accounts and there have already been 100 iterations of the test with previous users, then the probability that the next user has an overlap of at least 22 with at least one previous iteration is $1-(1-0.0056)^{100} = .43$. That is, if we require that the overlap with *every* one of the previous 100 iterations is less than 22, then an honest agent has a chance of only 57% of getting an account on the first attempt (assuming that this agent is not rejected due to poor performance).

We first introduce a (highly simplified) model of the limitations of human memory. Suppose that when the user is asked to recall the color of an item, one color (not necessarily the right one) pops up into her memory. In game-theoretic terms, this color can be referred to as a *signal* that she receives from her memory. Specifically, suppose that when a user takes the test a second time,

- with probability $p_1$ the signal is the correct answer (from the second iteration),
- with probability $p_2$ the signal is the correct answer from the first iteration, and
- with probability $p_3 = 1 - p_1 - p_2$ the signal is one of the colors drawn at random.

Presumably, $p_1 > p_2$. As for $p_3$, all of the following are reasonably possible: $p_3 \geq p_1$ (a forgetful user), $p_1 > p_3 \geq p_2$ (a user that is somewhat forgetful and does not get confused much), and $p_2 > p_3$ (a user that is not very forgetful but does get confused). Since the probability that the correct answer in the first iteration is the same as the correct answer in the second iteration is $1/n_c$, the (*ex ante*) probability that the correct answer pops up for a given item is $p_1 + p_2/n_c + p_3/n_c = p_1 + (1 - p_1)/n_c$. Similarly, the probability that the correct answer from the first round pops up is $p_1/n_c + p_2 + p_3/n_c = p_2 + (1 - p_2)/n_c$. The user does not receive any other signal from her memory (such as a confidence level that the answer is the correct one).

In this highly simplified model, for each item, the user must choose whether to respond with the color corresponding to her signal, or with some other color. (Since there is no way to distinguish the other colors, we may assume that she chooses one of the remaining colors at random in the latter case.) Thus, the only strategic decision that the user can make is the fraction $q$ of items for which she responds with the signal. If she responds with the signal for an item, the probability that she is right is $p_1 + (1 - p_1)/n_c$. If she responds with a random other color, then the probability that she is right is $(1 - (p_1 + (1 - p_1)/n_c))/(n_c - 1) = (1 - p_1)/n_c$. Thus, the expected fraction of times that she is right is $p_1 q + (1 - p_1)/n_c = p_1(q - 1/n_c) + 1/n_c$. Similarly, it can be shown that the expected fraction of times that she responds with the correct answer from the first iteration is $p_2 q + (1 - p_2)/n_c = p_2(q - 1/n_c) + 1/n_c$. (If we, completely inaccurately, assume that in our experiment, all users had the same $p_1$ and the same $p_2$, and that they all set $q = 1$, then this produces estimates of $p_1 = .42$ and $p_2 = .06$.) If $q = 1/n_c$ (which corresponds to random guessing), both of these expressions are equal to $1/n_c$. Hence, intuitively, if a user wants to increase the first expression beyond $1/n_c$, the second expression must also increase beyond $1/n_c$, and the second increase must be $p_2/p_1$ times the first increase.

This suggests the following metric for evaluating whether a test taker is the same as the taker of a given previous iteration of the test.

- Take the percentage of answers that are correct (for the current iteration), minus $1/n_c$ (the percentage expected for random guessing). Call the resulting fraction $f_1$.

- Then, take the percentage of answers that coincide with the correct answers from the earlier iteration, minus $1/n_c$ (the percentage expected for random guessing). Call the resulting fraction $f_2$.
- Finally, take the ratio $f_2/f_1$.

Then, by the above, if the test taker is the same in both iterations, the resulting ratio must be somewhere close to $\frac{p_2(q-1/n_c)}{p_1(q-1/n_c)} = p_2/p_1$ (assuming that $f_1$ is significantly positive). (With our very rough experimental estimates from above, $p_2/p_1 = .06/.42 = .14$.) The following theorem makes this precise.

**Theorem 1** *Suppose the following are true:*

- *the game-theoretic model proposed above is correct,*
- *the test taker is the same in both iterations,*
- *in the second iteration, the test taker sets $q$ to a value above $1/n_c$ (*i.e.*, she does not guess randomly).*

*Then, for any $\epsilon > 0$, as the number of items $n_i$ goes to infinity, the probability that $|f_2/f_1 - p_2/p_1| \geq \epsilon$ goes to zero.*

*Proof.* For any $\epsilon_1 > 0$, as $n_i \to \infty$, the probability that $|f_1 - p_1(q - 1/n_c)| \geq \epsilon_1$ goes to zero (using the law of large numbers and the fact that $p_1(q - 1/n_c)$ is the expected value of $f_1$). Similarly, for any $\epsilon_2 > 0$, as $n_i \to \infty$, the probability that $|f_2 - p_2(q - 1/n_c)| \geq \epsilon_2$ goes to zero. Because $p_1 > 0$ and $q > 1/n_c$, it must be the case that $p_1(q - 1/n_c) > 0$, and hence, for any $\epsilon > 0$, as $n_i \to \infty$, the probability that $|f_2/f_1 - p_2/p_1| = |f_2/f_1 - \frac{p_2(q-1/n_c)}{p_1(q-1/n_c)}| \geq \epsilon$ goes to zero as well.

By contrast, if the iterations of the test had different test takers, then with high probability, the ratio $f_2/f_1$ is close to 0, because the expectation of $f_2$ must be 0. (This is assuming that performance on the current iteration is significantly better than random guessing, so that the expectation of $f_1$ is positive). Thus, if we require $f_1$ to be significantly above 0 to pass the test (that is, the user should be getting significantly more answers right than random guessing would give, and hence must set $q$ to a value significantly greater than $1/n_c$ to have a good chance of passing), the number of items is sufficiently large, and $p_2 > 0$, then with sufficiently many items we can reliably detect when an applicant has taken the test before (because $p_2/p_1 > 0$).

## 5 Conclusions and future research

In many Web-based applications, there are incentives for a user to sign up for more than one account, under false names. By doing so, the user can send spam e-mail from an account (which will eventually cause the account to be shut down); distort online ratings by rating multiple times (in particular, she can inflate her own reputation ratings); indefinitely continue using a product with a free trial period; place shill bids on items that she is selling on an auction site;

engage in false-name bidding in combinatorial auctions; participate in the same online poker game under multiple identities, allowing her to effectively collude with herself; *etc.* All of these behaviors can be beneficial to the individual user, but are highly undesirable from the perspective of system performance. Users end up receiving tons of unwanted e-mail; online ratings become meaningless; companies become unwilling to offer free trial periods; users become skeptical of online auctions and poker games; *etc.* CAPTCHAs offer a partial remedy in that they can prevent a bot from automatically signing up for many accounts. However, they do not prevent a *human* from signing up for multiple accounts. It may appear that the only way to prevent the latter is to require the user to provide information that identifies her in the real world (such as a credit card or telephone number), but users are typically reluctant to give out such information. In this paper, we proposed an alternative approach. We investigated whether it is possible to design an automated memory test that is easy to pass once, but difficult to pass a second time. Specifically, we designed a memory test. In this test, items are randomly associated with colors (*"Cars are green."*). The user first observes all of these associations, and is then asked to recall the colors of the items (*"Cars are...?"*). The items are the same across iterations of the test, but the colors are randomly redrawn each time (*"Cars are blue."*). Therefore, a user who has taken the test before will occasionally accidentally respond with the association from the previous time that she took the test (*"Cars are...? Green!"*). If there is significant correlation between the user's answers and the correct answers from a previous iteration of the test, then the system can decide that the user is probably the same, and refuse to grant another account. We presented and analyzed the results of a small study with human subjects, in which each subject took the test twice. The results of this study were mixed. On the negative side, about half of the subjects did not perform very well on the tests. On the positive side, for subjects that performed well, there was significant overlap between their answers in the second iteration and the correct answers in the first iteration. Thus, the system may be effective at preventing multiple account registrations from the same person, but not at allowing everyone to obtain an account. To analyze whether there exists some strategy for users that is more successful at signing up for multiple accounts, we introduced a simple game-theoretic model. We showed that under this model, any strategy is likely to fail at signing up for multiple accounts, if the test is large enough.

There are several aspects of the proposed test design that limit its feasibility in practice. First, the test is long and exhausting, which would probably discourage users from signing up for accounts. Second, the study indicates that performance on the test is very variable (even when its takers are restricted to Duke University students). Because in addition, the study suggests that we must require a high percentage of correct answers for passing the test in order to see the desired confusion (that is, overlap) across iterations of the test, this means that some users would have serious difficulty passing the test. Third, while the study suggests that it is difficult to do well on the test twice without getting confused across iterations, the study took place under controlled conditions. In the real world, users may try to pass the test multiple times in different ways (by waiting a longer time between iterations,[7] getting other people to help them, trying to use tools to record the associations (though the speed at which the test is run makes this difficult), *etc.*), and we know little about the test's robustness to such behavior.

While there are many obstacles that need to be overcome for this approach to be truly practical, we feel that the results are encouraging enough, and that the value of having a practical solution would be high enough, that it is very much worthwhile to pursue further research on this topic. Such research should probably investigate other variants of the basic test design. In the appendix, we present results for one alternative design that is based on face recognition by the subjects. Unfortunately, that design did not end up working very well, but the results are informative for future designs.

One can imagine numerous other designs. For example, a test based on procedural ("how-to") memory rather than declarative (fact-storing) memory may be more effective. To find the optimal design, it may be beneficial to reach out to researchers in cognitive psychology and cognitive neuroscience, to exploit known particularities of human memory. However, our approach can introduce incentives for test takers to behave in ways that are not beneficial in more typical memory tests. These incentives and the behavior that they are likely to cause must be rigorously studied, both in theory and through experimental evaluation. Creating a truly practical system is an ambitious goal, but one that, if reached, will make many existing Web-based applications much more efficient, and will probably make new ones feasible.

## References

1. Peter Cramton, Yoav Shoham, and Richard Steinberg. *Combinatorial Auctions.* MIT Press, 2006.
2. Drew Fudenberg and Jean Tirole. *Game Theory.* MIT Press, 1991.
3. Greg Mori and Jitendra Malik. Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 134–141, 2003.
4. Gabriel Moy, Nathan Jones, Curt Harkless, and Randall Potter. Distortion estimation techniques in solving visual CAPTCHAs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 23–28, 2004.
5. Roger Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, Cambridge, 1991.
6. Martin J Osborne and Ariel Rubinstein. *A Course in Game Theory.* MIT Press, 1994.
7. Arasanathan Thayananthan, Bjoern Stenger, Philip H. S. Torr, and Roberto Cipolla. Shape context and chamfer matching in cluttered scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 127–133, 2003.

[7] This would still imply that the rate at which users can sign up for accounts has decreased.

8. Luis von Ahn, Manuel Blum, Nicholas Hopper, and John Langford. CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology - EUROCRYPT 2003, International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311, Warsaw, Poland, 2003.
9. Luis von Ahn, Manuel Blum, and John Langford. Telling humans and computers apart automatically: How lazy cryptographers do AI. *Communications of the ACM*, 47(2):56–60, February 2004.
10. Makoto Yokoo. The characterization of strategy/false-name proof combinatorial auction protocols: Price-oriented, rationing-free protocol. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 733–742, Acapulco, Mexico, 2003.
11. Makoto Yokoo, Toshihiro Matsutani, and Atsushi Iwasaki. False-name-proof combinatorial auction protocol: Groves mechanism with submodular approximation. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1135–1142, Hakodate, Japan, 2006.
12. Makoto Yokoo, Yuko Sakurai, and Shigeo Matsubara. Robust combinatorial auction protocol against false-name bids. *Artificial Intelligence*, 130(2):167–181, 2001.
13. Makoto Yokoo, Yuko Sakurai, and Shigeo Matsubara. The effect of false-name bids in combinatorial auctions: New fraud in Internet auctions. *Games and Economic Behavior*, 46(1):174–188, 2004.

## Appendix: another test based on recognizing faces

In this appendix, we present the experimental results of another test, which turned out not to work as well as hoped. In this test, we used a database of 58 human faces (a subset of the Indian Face Database developed at IIT Kanpur). A subject was first shown 29 faces drawn at random from the 58, one face at a time, for 5 seconds per face. Subsequently, the subject was shown the full set of faces (one at a time, for 4 seconds per face); in this second phase, the subject was asked, for each face, whether she/he had seen the face in the first phase. Each subject took this test twice (using the same database of 58 faces each time, but with a new draw of 29 faces in the first phase of the second iteration). (Each subject also did a practice run beforehand on a few faces not in the 58.) Again, each subject received US $7, plus the percentage of correct answers times US $7.

The hope was that performance in the second iteration of the test would be worse than in the first iteration of the test, due to the fact that, in the second phase of the second iteration, if a face looks familiar to the subject it may be difficult for him/her to decide whether he/she had seen it in the first phase of the second iteration, or only at some point in the first iteration of the test (in the latter case, the correct answer would be "no"). If performance were consistent across subjects, and significantly worse in the second iteration, then perhaps we could set a threshold that everyone can pass the first time but not another time. Unfortunately, this turned out not to be the case, as the results below show.

| **subject** | # correct in iteration 1 | # correct in iteration 2 |
|---|---|---|
| **1** | 53 | 45 |
| **2** | 47 | 49 |
| **3** | 48 | 44 |
| **4** | 43 | 42 |
| **5** | 45 | 45 |
| **6** | 41 | 51 |
| **7** | 46 | 47 |
| **8** | 36 | 34 |

Table 1: Experimental results for human subjects in the face-based test.

While for some subjects, there was a drop in performance in the second iteration, these drops were generally not significant, and some subjects' scores actually increased in the second iteration. It appears that subjects did experience some confusion in the second iteration, but at the same time, there was a learning effect: subjects became generally better at remembering the faces, and this canceled out the confusion effect. Perhaps this learning effect can be removed by making subjects practice beforehand, but this would make the duration of the test unreasonable.

One may also wonder if it is possible to observe correlations across iterations of this test, as we did for the test in the main part of this paper. In the experiment, it was the case that most of the subjects' wrong answers in the second iteration occurred when the correct answer for a face in the second iteration was not the same as the correct answer for that face in the first iteration; however, this effect does not appear strong enough to confidently conclude that two iterations of the test correspond to one person (especially because subjects generally did not have that many wrong answers in this test).