CESifo WORKING PAPERS

10239
2023

January 2023

Can You Spot a Scam? Measuring and Improving Scam Identification Ability

Elif Kubilay ⓡ, Eva Raiber ⓡ, Lisa Spantig ⓡ, Jana Cahlíková ⓡ, Lucy Kaaria ⓡ

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10239*

Can You Spot a Scam? Measuring and Improving Scam Identification Ability

Abstract

The recent expansion of digital financial products leads to severe consumer protection issues such as fraud and scams. As these potentially decrease trust in digital services, especially in developing countries, avoiding victimization has become an important policy objective. In an online experiment, we first investigate how well individuals in Kenya identify phone scams using a novel measure of scam identification ability. We then test the effectiveness of scam education, a commonly used approach by banks and institutions for fraud and scam prevention. We find that common tips on how to spot scams do not significantly improve individuals' scam identification ability, i.e., the distinction of scams from genuine messages. This null effect is driven by an increase in correctly identified scams and a decrease in correctly identified genuine messages. We interpret this as an increase in caution. In addition, we find suggestive evidence that genuine messages which contain scam-like features are more likely to be misclassified, highlighting the importance of a careful design of official communication.

JEL-Codes: D140, D180, G530, O120.

Keywords: consumer protection, consumer fraud, digital financial services, scam susceptibility, scam education, Kenya.

Elif Kubilay ⓡ
University of Essex / United Kingdom
elif.kubilay@essex.ac.uk

Eva Raiber ⓡ
Aix-Marseille University / France
eva.raiber@univ-amu.fr

Lisa Spantig ⓡ
RWTH Aachen / Germany
lisa.spantig@rwth-aachen.de

Jana Cahlíková ⓡ
University of Bonn / Germany
jana.cahlikova@uni-bonn.de

Lucy Kaaria ⓡ
University of Nairobi / Kenya
lucy.kaaria@hopawi.com

January 18, 2023

For their valuable comments and suggestions, we thank seminar and conference participants at the IPA Researcher Gathering 2021, the University of Essex, the University of Groningen, and the Ludwig-Maximilian-University of Munich. We are grateful to Lyne Chahed for her outstanding research assistance and to Nendo for excellent research support. Funding from Innovations for Poverty Action Consumer Protection Research Initiative (BMG-19-10001-X5) is gratefully acknowledged. Jana Cahlíková acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866. Eva Raiber acknowledges funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from the Excellence Initiative of Aix-Marseille University - A*MIDEX. Eva Raiber is a Research Affiliate at the Centre for Economic Policy Research (CEPR). The study has received IRB approval from IPA (16014), and ethical approval from the University of Essex (ETH2021-1858), and is registered with the AEA registry (AEARCTR-0008754). Declarations of interest: none.

1 Introduction

The expansion of digital financial services (DFS) has increased access to financial services, both in developed and developing countries (e.g., Pazarbasioglu et al., 2020; Balyuk, 2022). With this increase in DFS, consumer protection issues are also on the rise (Garz et al., 2021). One major issue is fraud. Fraud is detrimental to consumers both in terms of direct monetary costs and indirect costs such as erosion of trust in financial services (Guiso et al., 2008; Gurun et al., 2017; Johnson et al., 2019), loss of confidence in financial matters (Brenner et al., 2020), and mental health problems including depression and stress (DeLiema et al., 2020; Financial Institution Regulatory Authority, 2015). One common type of fraud is phone scams using text messages or calls. The goal of scammers is to trick consumers into sending money or revealing private information such that their accounts can be accessed. Since scammers often target random phone numbers, all segments of society who have a phone and use basic DFS are at risk.

The existing recipe for avoiding consumers' scam victimization is to pursue education and awareness campaigns. Yet, do educational campaigns indeed improve people's ability to detect scams and do they influence how genuine messages from e.g. banks or telecommunication providers are perceived? An important obstacle to evaluating the effect of education campaigns is quantifying the relevant outcome metrics. Consumers under- or misreport fraud attempts and victimization: They might not be able to recognize all types of fraud, differentiate genuine offers from scams, or remember all instances of fraud attempts (Chen et al., 2018). Moreover, victims often feel shame and guilt and do not report scams to avoid potential stigma (Burke et al., 2022). Therefore, we argue that a policy-relevant metric is the ability to identify fraud attempts and confidence in this ability. Even if only a few individuals are direct victims of fraud, the inability to recognize fraud or the lack of confidence in this ability may impede market participation.

In this paper, we study susceptibility to scams and the effectiveness of a light-touch scam education in Kenya. First, we develop a novel measure for an individual's scam identification ability (SIA) and confidence in their ability. For this, we collect actual scams and official communication that circulate in Kenya. Second, we test experimentally if common tips for scam detection improve SIA and confidence. We focus on Kenya, Africa's leader in digital infrastructure and mobile money use (Koyama et al., 2021). At the same time, the country

suffers from increasing rates of phone scams, which by now represent the most often cited consumer protection issue (Blackmon et al., 2021).

In an online survey (N=1,000) we show respondents 12 different messages and ask them to indicate whether these messages are scam or not. Each classification decision is followed by a confidence rating. The messages include both common scams and genuine messages sent by, e.g., banks or telecommunication companies in Kenya. After having classified the first six messages, a random half of the respondents receive tips on fraud prevention that are commonly provided by banks or telecommunication companies. These tips warn consumers about "scam markers," which include i) typos and grammar mistakes, ii) an unknown sender, iii) a shortened link, and iv) requests for private information such as pin codes or passwords. Ideally, these tips help respondents become better at distinguishing scams from genuine messages. However, it is also possible that tips about scams make respondents more cautious and hence more likely to classify any given message as scam. The latter would make it harder for service providers to communicate with their clients.

We find that on average tips do not increase scam identification ability. This null effect arises because while respondents in the treatment group are more likely to correctly identify scams, they are also less likely to correctly identify genuine messages. On average, tips appear to make consumers more cautious, i.e., more likely to classify any given message as scam. Moreover, receiving tips makes respondents significantly more confident in their classification decisions. The increase in confidence could be concerning as average SIA does not increase in our study and overconfidence has been found to be correlated with victimization (McAlvanah et al., 2015). However, we find suggestive evidence that higher confidence is associated with better SIA at the individual level.

Looking deeper, we find a more nuanced result depending on whether a given message contains a "scam marker." First, tips increase the number of correctly identified scams, irrespective of whether a scam marker is present in the message. This suggests that tips indeed make people more cautious. Second, a scam marker in a genuine message increases the likelihood of this message being classified as a scam. Part of the null effect of tips thus seems to be driven by official messages that look like scams. This highlights that tips need to be specific enough to unambiguously increase SIA and that non-scam communication should avoid features that are commonly cautioned against in educational campaigns.

To test whether the results change when money is at stake, a random half of our sample receives incentives for each correct classification. Results show that incentives do not lead

to better SIA, and there is no interaction effect with tips. While we find some indications that those who receive both tips and incentives exert more effort, these individuals are not performing better. This implies that our measure can be used as an unincentivized survey measure. We also investigate treatment effect heterogeneity and illustrate an important shortcoming of such education interventions. Tips appear to be effective in increasing SIA only for more experienced DFS users and those with higher education. Less educated participants do not benefit from tips, suggesting that it is difficult to design universally-helpful communication.

This study relates to several strands of literature. First, we contribute to the nascent literature on financial fraud in developing countries. For example, Ensminger and Leder-Luis (2022) and Andersen et al. (2022) study the detection of fraud in foreign aid, whereas Garz et al. (2021) summarize the consumer protection challenges of the expansion of DFS. Different types of fraud have been documented in various settings: fraudulent smartphone apps in India (Fu and Mishra, 2022), phone scams as the most prominent consumer protection issue in Kenya (Blackmon et al., 2021), and agent misconduct in Ghana (Annan, 2022a,b). Here, we focus on phone scams, develop a measure of SIA based on actual scams and official messages, and show that information makes individuals more careful on average but does not increase their SIA.

Second, we contribute to the literature studying the causal effects of educational interventions on fraud susceptibility. This literature has mostly focused on phishing attacks (e.g., Sheng et al. (2007)) but also studied telemarketing schemes (Scheibe et al., 2014), and investment scams (Burke et al., 2022). In general, tips and information may decrease fraud susceptibility, especially among better-educated individuals (Burke et al., 2022). Our study suggests that previous findings, albeit focusing on different types of scams, appear to hold for phone scams and in a developing country setting. Additionally, we make a methodological contribution by making mistakes costly for half of our sample and show that this does not alter results. Participants' intrinsic motivation to correctly classify messages appears to be high enough.

Third, we contribute to a large literature documenting correlates of fraud susceptibility and victimization (see Moustafa et al., 2021; Norris et al., 2019, for recent reviews). This literature studies samples from developed, Western countries. The most common demographic characteristics that have been found to matter are gender and age. Additionally, financial knowledge (Engels et al., 2020), as well as risk aversion, curiosity, and the level of trust

(Chen et al., 2018) are associated with fraud susceptibility. We find similar results in the Kenyan context. Women and less experienced DFS users have lower SIA. Additionally, we show that these groups do not differentially benefit from tips. These results imply that unless information provision is more targeted at specific groups of the population, such policy interventions are unlikely to close existing gaps in SIA.

2 Background

Kenya is a leading market for digital financial products and services (Koyama et al., 2021). Often, solutions are tested in Kenya and then rolled out to other countries in the region. With near-universal phone penetration and use of DFS in Kenya, almost all adults are at risk of phone scams. In a representative survey of active DFS users, 56% reported they had been contacted by scammers in the past six months, most commonly by phone (Blackmon et al., 2021). Scam reports are prevalent among all demographic groups. Interestingly, 68% of users with tertiary education reported scam attempts, compared to only 50% among those with at most secondary education. This suggests that less educated consumers might not recognize all scam attempts and/or might be less willing to report them.

Given the high prevalence of scams, it is not surprising that 90% of the adult population is concerned about fraud when using digital services (Koyama et al., 2021). In terms of direct costs of victimization, recent numbers show a positive correlation between the depth of digital services use and the amount lost due to fraud.[1] This implies that with increased use of DFS, more people will be at risk of suffering from unexpected losses that might be difficult for them to absorb. Regarding indirect costs, 71% of the self-employed report limiting their usage of DFS due to concerns about fraud (Koyama et al., 2021), indicating loss of trust.

Qualitative interviews and scam examples from social media that we collected show that scammers try to trick individuals into transferring money or to obtain personal information to either access accounts or steal the identity of the victim. Scammers often impersonate bank and telecommunication agents, relatives or friends. A variety of different scams exist, from fake loan or investment offers to prizes for which money has to be sent upfront to take advantage of these "opportunities." "Erroneous contact" is another common scam in which the sender pretends to have sent money or sensitive information and either asks for

[1] Koyama et al. (2021) find that over the past three years, more advanced users lost more than twice as much as the basic digital services users due to fraud.

the money to be transferred back or for the enticing information to be ignored. In the latter cases, the primary goal is to start a conversation for more sophisticated social engineering.

While Kenya has passed digital safety policies and laws, and has established the office of the Data Protection Commissioner, the problem of fraud cannot be solved by regulation alone. Technological innovations such as biometric identification can help protect identities and accounts, but the human factor also needs to be addressed. In other contexts, it appears that financial knowledge is associated with lower susceptibility to fraud (Engels et al., 2020), which might explain the general popularity of educating consumers to raise awareness of and resilience to fraud (DeLiema et al., 2020; Engels et al., 2020).

3 Measuring Scam Identification Ability

To build our measure of scam identification ability (SIA), we obtained information about ongoing scams from different sources. First, using a social media analytic tool, we collected public posts from Twitter sent between January 2020 and June 2021 from a Kenyan location. We kept the posts that were sent from an individual account and related to phone scams based on topic clustering. We further restricted the sample to contain screenshots of text messages which, after removing duplicates, left us with 116 tweets. Additionally, we conducted a survey in the largest Kenyan fraud-detection Facebook group in September 2021. Members of the group were asked to submit examples of both scam and official messages and calls. Participants submitted 922 examples, of which about 62% were scams. As the type of messages, i.e., scam or official, is self-reported and might be subject to error, we hired two research associates to independently classify 516 messages (including 116 from Twitter) and assert their confidence. In cases where two coders' classification did not match, a third research associate was asked to make a classification.

We focus on SMS scams and non-scam text messages to use examples verbatim.[2] To generate variation in our SIA measure, we construct a database of ambiguous messages.[3] From this set of ambiguous messages, we randomly select 13 scams and seven official messages, stratified by topic. We turn these messages into vignettes by equalizing the visual appear-

[2]Recalled protocols of calls were incomplete, similar to examples of SMS that were not copy-pasted or submitted as a screenshot. Administering the vignettes in a written context (in our online survey) allows us to keep the mode of perception close to real life.

[3]We discuss the implications of this choice in Section 6.2 and describe the process of building the measure in more detail in Appendix C.

ance and pilot the 20 vignettes in two small convenience samples (N=39). We select the 12 final vignettes based on the classification decisions and confidence of pilot participants.

Our measure consists of two blocks with four scam vignettes and two official messages each.[4] The blocks are presented in random order, and the messages are randomized within each block. We refer to the block shown first as "block 1" and the one shown second as "block 2." For each block, we measure SIA as the share of correctly classified messages. We also examine separately whether individuals classify scams and non-scams correctly. As we are more interested in the former, we decided to include more scam than non-scam messages in our measure. For each vignette, participants indicate whether this is a scam or not (binary choice). Afterwards, a scale appears on the same page and asks participants to rate their confidence in their classification on a five-point Likert scale where the higher values indicate higher confidence.

4 Experimental Setting

We measure SIA in an online survey in which we also administer an education treatment to estimate the causal effect of scam tips on the ability to distinguish fraudulent from genuine messages.

Tips treatment

Educational campaigns aim at raising awareness and providing tips on how to distinguish scam and non-scam communication (e.g., "Safaricom will only SMS you from MPESA and Safaricom") or on how to behave (e.g., "never share your PIN"). These campaigns are often run visually on billboards or social media. Therefore, to capture available information on fraud prevention, we collected examples of tips using Twitter and qualitative data. We condense the five most common pieces of information into one infographic (see Figure 1). To avoid information overload and ensure that all tips are read, we animate the graphic, such

[4]For non-scams, we focus on official communication by banks, Safaricom as the provider of MPESA, and other telecom providers. As we exclude circumstantial clues from our design, personal messages from family and friends cannot be unambiguously classified as non-scam. As an unknown sender is the most obvious clue for a scam, we vary whether the sender is shown in the vignette. See Table A1 for an overview of all vignettes and Figure A1 for a visual example.

that the participants see one bullet point at a time. Participants go through this animation at their own speed. On average, they spent 1.12 minutes (SD=0.67) reviewing the tips.

We randomize scam education at the individual level and provide it to 50% of our sample. We administer the treatment between the two blocks of vignettes, which allows us to assess individuals' SIA level prior to tips treatment. It is important to note that, as in real life, we do not distinguish between information being new or serving as a reminder.

Incentive treatment

In contrast to real life where mistakes can be costly, our participants may exert less effort. We hence cross-randomize a robustness treatment in which we pay 10 KES for each correctly classified message. Half of our sample receives incentives in both blocks. Different from the tips treatment, incentives may thus influence all classifications. We opted to pay incentives from the beginning such that participants who receive both tips and incentives can focus on understanding the main treatment between the two blocks. Finally, the incentive treatment allows us to explore whether using incentives is essential to elicit scam identification ability.

Online survey and sequence of events

After written consent and questions on demographics, phone ownership and usage, participants are shown a definition of scams and told that their task is to identify scam messages. They do not receive information about the number of vignettes or the fraction of fraudulent messages. Before starting the first block, participants in the incentive treatment learn about the payment for correct classification. After the first block, participants in the tips treatment go through the animated infographic. Nobody receives feedback on their SIA measured in the first block. Afterwards, everyone proceeds with the second block, followed by questions regarding the use of DFS, scam experiences, and an attention check. At the end of the survey, participants learn the number of correctly identified messages and those in the incentive treatment also see the corresponding bonus payment.

Procedures

We programmed the survey in Qualtrics and recruited 1,000 Kenyan respondents from a consumer panel of Geopoll, implementing quotas for gender, age, and county of residency.

On average, respondents took 21 minutes to complete the entire survey, and each participant received a completion payment of 500 KES (4.40 USD at the time of the experiment), in addition to any eventual incentive payments.

5 Results

We randomly allocated 1,000 participants to the four treatments, which resulted in 256 individuals in Control, 259 in Tips, 246 in Incentives, and 239 in Tips and Incentives. Baseline characteristics are balanced across treatments (see Table A2).[5]

5.1 Descriptive statistics

Due to our quotas, half of our sample is female, 32% between 18 and 24 years, 27% between 25 and 34 years, and 41% 35 years and above. This implies that with 32 years on average our sample is older than the general Kenyan population but relatively comparable to the adult population (see Table A3). While respondents come from all over Kenya and are representative in terms of residency at the county level, urban participants are over-represented (50% as compared to 31% of the population in urban areas). Table A4 presents further descriptive statistics: Our sample is comparatively well-educated (73% have a post-secondary education), 78% self-classify as low-income and 36% have formal employment. As the design of the survey requires access to internet, it is not surprising that 99% have internet access and use social media on their phone. Almost all participants (96%) have recently used DFS on their phone and on average, participants use five different services with the most frequent ones being sending and receiving mobile money (89%), paying bills (71%), and conducting transactions involving an agent (55%).

In our sample, 96% report that they have been contacted by a scammer in the past.[6] Of those, 14% state having been contacted in the past week. The most common way of

[5]The data collection proceeded as planned and there were no changes to the pre-registered experimental design. In a few instances, we deviate from the pre-analysis plan, mostly for expositional clarity. We discuss all these changes in Appendix E.

[6]These numbers are substantially higher than the ones reported in the phone survey by Blackmon et al. (2021). This may be explained by several differences. First, in our survey, we provide participants with visual examples of scams that might make recall easier. Second, our sample is more educated than theirs, and they find that reports of scam contacts are positively correlated with education. Third, if reporting is influenced by social image concerns, online survey mode might increase reporting rates.

contact is reported to be SMS, followed by phone calls. Consistently with our findings from the social media and qualitative analysis (see Appendix C), the top three asks by scammers were to send money, to reverse a payment, and to share personal information. More than half of our sample report having ever been victimized.

5.2 Scam identification ability

We first present descriptive statistics from block 1, i.e., prior to the tips treatment. On average, participants correctly identified 71% of the six messages. Panel A in Figure A2 illustrates the distribution of SIA. Only 12% of all respondents correctly identified all six messages. Participants can make two kinds of identification mistakes: They might misclassify a scam (as a non-scam message), or they might misclassify a non-scam message (as a scam). On average, individuals classified 74% of scams and 66% of non-scams correctly. Confidence in SIA is high on average, at 4.23 out of 5 in block 1. Seventeen percent of participants always indicate the highest confidence score (see Panel B in Figure A2). SIA and confidence are positively correlated (Spearman's rho=0.179, $p<0.001$).

Table 1 shows the correlates of SIA and confidence in block 1. Gender is the most robust and significant correlate of both SIA and confidence, with women having a 3 percentage point lower SIA score (equivalent to classifying 0.2 fewer messages correctly) and being less confident in their ability. These results are consistent with the well-documented gender gap in financial literacy (Lusardi and Mitchell, 2014). Other demographic characteristics are at most weakly correlated with SIA. Age and having more than secondary education are positively correlated with confidence. Those who use a larger variety of DFS have a 3 percentage point better SIA score (they classify 0.2 more messages correctly). Low trust in DFS is associated with lower confidence. We find no significant association between individuals' scam experience (i.e., being contacted or victimized) and SIA or confidence.

Lastly, we assess the effect of incentives in block 1 on our four main outcome variables: SIA (the share of correctly identified messages), the share of correctly identified scams, the share of correctly identified non-scams, and the confidence level. Panel 1 in Table A5 shows that incentives have no significant effect on any of the outcomes. While we control for the incentive treatment in all the following analyses, we will focus on the two tips treatments for ease of exposition.

5.3 Effects of scam education

To test the null hypotheses that i) tips (unincentivized) and ii) tips (incentivized) have no effect on our main outcome variables, we estimate the following model:

$$y_i = \alpha_0 + \alpha_1 Tips_i^U + \alpha_2 Tips_i^I + \gamma_1 y_{0i} + X_i'\gamma_2 + Other_i\delta + \epsilon_i$$

, where y_i is our outcome variable measured in block 2. $Tips_i^U$ indicates that individual i received the tips treatment without the incentives. $Tips_i^I$ indicates that individual i received both the tips and incentives. y_{0i} controls for the baseline levels of the outcome variable from the first block. X_i is a set of individual characteristics for respondent i. These include gender, age, income, and education level. $Other_i$ captures additional controls, such as the order of the two blocks and whether individual i received incentives (with no tips). We use robust standard errors ϵ_i. Our coefficients of interest are α_1 and α_2, i.e., the effect of tips without the incentives and with the incentives, respectively.

Column 1 of Panel 1 in Table 2 shows that tips do not increase SIA relative to the control group (no tips and no incentives). The same holds for tips with incentives. Columns 2 and 3 help explain why tips have no overall effect. While tips are helpful in increasing the share of correctly identified scams (Column 2), they decrease the share of correctly identified non-scams (Column 3). These effects do not depend on incentives.

Columns 4 to 6 present the treatment effects on confidence. Column 4 shows that, on average, individuals who received tips become more confident in their classifications. This increase is driven by participants becoming more confident in the classification of scams (Column 5). In contrast, the confidence in the classification of non-scams does not change with tips (Column 6), despite the worse performance (Column 3).[7]

Panel 2 in Table 2 shows the effect of our treatments on secondary outcomes. First, we find no significant effect of tips on trust in digital financial services.[8] Tips increase the time participants spend on the classification task in comparison to the control group only

[7]These averages might mask substantial heterogeneity. We hence try to assess to what extent changes in confidence coincide with improvements in SIA. Table A6 provides suggestive evidence that, on average, increases in confidence occur together with increases in SIA (Column 1). This association of SIA and confidence is particularly strong for scams (Column 2), but reversed for non-scams (Column 3): confidence does not increase while performance decreases.

[8]Note that we measure trust in DFS only once after all messages have been classified. We hence cannot control for a baseline level of trust.

when the incentives are provided. Note, however, that we cannot statistically distinguish the effect of tips with and tips without the incentives. The former may induce higher effort (proxied by longer response times), but this does not lead to better outcomes. The last two columns show treatment effects on classifying all scams and all non-scam messages correctly, confirming the results from Panel A. Our results are not driven by a lack of attention or a specific set of control variables (see Appendix B).

5.4 Heterogeneity

We investigate who benefits from tips. Specifically, we explore treatment effects for respondents separately by the following characteristics: gender, age, education, income level, rural and urban areas as well as experience with DFS and scams. Figures 2a and 2b plot the coefficients of SIA and confidence, respectively, for each subgroup.

First, we note that the directions of effects in most subgroups are consistent with our main results and most subgroups react equally to the tips treatments. In terms of SIA, tips, irrespective of the presence of incentives, appear to work better for those with post-secondary education and a more diverse use of DFS (using 5 or more different services). Recall that those with more DFS experience are also better at identifying scams in the baseline (see Table 1). This suggests that tips further increase the gap in SIA between inexperienced and experienced DFS users. Confidence increases for most subgroups. Those who are older, with lower education, and higher income do not become more confident in the absence of incentives. In contrast, when incentives are present, only those who are younger, with higher education, lower income, and living in urban areas become more confident with tips.

6 Discussion

We find no significant average effect of tips on SIA, but differential effects of tips on scams and non-scams. In this section, we present potential explanations for these results and underlying mechanisms. Additionally, we discuss how to interpret our effect sizes. Note that this section is exploratory in nature.

6.1 Exploring the effect of tips on SIA

Our light-touch scam education in the form of scam tips does not improve SIA. However, tips improve the identification of scams, while they worsen the identification of non-scams. This pattern could emerge due to two reasons. First, tips may increase caution, such that participants are more likely to classify any given message as a scam. Second, not only scams but also non-scam messages may contain "scam markers," such that tips "apply" to both scam and non-scam messages. In the former case, policymakers may want to weigh the benefits of improved scam identification against the costs of heightened classification mistakes for genuine communication. In the latter case, it should be discussed whether tips can be refined and whether official communication can distinguish itself better from scams.

We analyze the effects of our treatments at the vignette level to shed light on potential mechanisms. To account for the fact that not all tips are helpful for all vignettes, we construct an indicator, $ScamMarker_m$, which captures whether at least one of the tips is helpful for correctly identifying the message as a scam. Only one scam message does not contain a scam marker while the other seven do. Yet, two out of the four official messages also contain a scam marker making them look like scams.

Figure 3 plots the average marginal effects obtained from our estimates for the control group and the tips without incentives treatment in block 2.[9] In the left panel, we include all messages, in the center panel, we only include scam messages, and in the right panel non-scams.[10] Similar to our main results, we find no differential effect of our treatment on the share of correctly identified messages in block 2, irrespective of scam markers (left panel).

Focusing only on scams (center panel), tips significantly increase the share of correctly identified messages, independent of whether the message contains a scam marker or not. This is in line with the interpretation that participants become more cautious and hence more likely to classify any given message as scam when they receive tips. There is one caveat worth mentioning here. We only have one scam message without a scam marker. For non-scams, we see that tips do not increase the share of correctly identified messages for messages without a scam marker. However, if a scam marker is present, tips significantly

[9]We focus on this comparison for ease of exposition; the effects for the tips with incentives treatment are qualitatively similar. We provide more detail in Appendix D.

[10]Note that the magnitudes cannot be compared across the panels as the share of correctly identified messages relies on six (left panel), four (center), and two messages (right panel), such that one mistake has a different magnitude in the three panels.

reduce the share of correctly identified messages. This highlights the challenge of designing educational campaigns in a setting in which genuine communication contains scam markers.[11] We conclude that if non-scams can avoid scam markers, tips can be unambiguously beneficial in increasing scam detection irrespective of scam markers while not decreasing the correct classification of non-scams.

6.2 Interpretation of effect sizes

Our setting differs in several ways from the "real life." For one, we abstract away from situational circumstances that may help classify messages. We also focus on messages that may be harder to classify than the average SMS individuals receive in Kenya. In general, without knowing all messages and the frequency at which they are being received, it is hard to interpret the absolute levels of our SIA measure. Thus, we mainly focus on differences in SIA between different groups, either defined by our treatments, or by demographics.

As to our treatment effects, we are primarily interested in their directions, and less so in the magnitudes. There are several reasons to believe that we estimate an upper bound of the effect of tips. First, our sample is literate and relatively educated and hence able to understand and apply the tips. In line with this, more educated and experienced DFS users appear to benefit more from the tips. Second, we provide tips when they are needed, in a more salient way than in real life. Additionally, as participants are aware they might face scams, they may pay more attention to tips than they would otherwise.

However, other points speak toward a lower bound of the effect. Being alert also means that the awareness-raising potential of tips is weakened, if not muted. As we find tips to be effective even when attention is incentivized, this argument seems to have less bite. In addition, since we use common tips, participants may know them already. This is especially likely given that our sample is more educated and uses the internet more than the average Kenyan population. Finally, if average scams are less challenging to identify than our vignettes, we might estimate a lower bound, as the following analysis suggests. Using vignette-level data from block 1, we create a measure of difficulty and analyze treatment heterogeneity at the vignette level in block 2, analogously to the analysis of "scam markers." For easy vignettes, we find a slight increase in SIA with tips for all messages, a positive

[11] Note that scam markers in official communication are not specific to our experiment. Anecdotally, we were surprised to find other scam markers such as urgency, all caps or shortened links in several of the official communication messages sent by banks and Safaricom.

and significant effect for scams, and no effect for non-scams. For difficult vignettes, tips significantly increase the correct classification of scams but significantly decrease the correct classification of non-scams (see Appendix D and Figure D2 for more details). Assuming that most official entities manage to communicate in easy-to-classify messages, we rather estimate a lower bound of the effectiveness of tips.

Lastly, we note that a limitation of our approach is that the effects of tips are examined using vignettes. While we make classification mistakes costly for half of our sample, this does not take into consideration that the costs of misclassifying scams and non-scams are likely different in practice. Moreover, similar to other studies in the literature, we are not able to assess how SIA translates into fraud detection in practical settings and the likelihood of victimization (Burke et al., 2022). Our results suggest that tips can decrease classification errors for scams, but further testing and quantifying effects, also in terms of potential downsides for non-scam communication, remains an important question for further research.

7 Conclusion

We study a progressive DFS market in which phone scams are highly prevalent, develop a measure of scam identification ability, and experimentally test the effect of scam education in the form of tips. On average, we find no significant effect of tips on SIA. We explain this null effect by an increase in correctly identified scams, and a decrease in correctly identified genuine messages. Further analyses reveal that these differential effects appear to be driven by scam markers that are also present in some of the non-scam communication by banks or telecommunication companies. If such communication could be distinguished more easily from scams, tips on how to spot scams may have an unambiguously positive effect on SIA. Moreover, we show that tips lead to an increase in confidence, driven by higher confidence in classifying messages that are indeed scam. We also find suggestive evidence that tips do not make individuals overly confident. This is in line with specific subgroups, namely the more educated and more experienced, benefiting from the treatment and becoming more confident.

Our analyses reveal several reasons why scam tips, despite of being a commonly used approach, might not be the silver bullet in addressing the human factor in scam victimization. First, it is challenging, if not impossible, to provide tips that benefit all. Our findings suggest

that tips, for example, benefit only the highly educated which potentially leads to a further increase in gaps between groups. Therefore, a more targeted approach may be necessary to reach everyone, and in particular, populations who may be more susceptible to scam victimization. Importantly, targeting is not only about the content, but also the medium used to educate consumers. For example, Burke et al. (2022) find that text-based messaging may work better for more educated populations, potentially explaining why our written texts work better for this subgroup. Second, it is difficult to communicate tips that apply to all kinds of scams. Tips in our setting seem to increase scam detection irrespective of scam markers, potentially due to an increase in caution. Moreover, as scams evolve dynamically, tips and guidance provided by authorities need to be revised regularly. Notifying consumers of these updates poses an additional challenge. Therefore, identifying new strategies for fraud prevention and scam awareness remains an important endeavor for future research.

References

Andersen, Jørgen Juel, Niels Johannesen, and Bob Rijkers (2022) "Elite Capture of Foreign Aid: Evidence from Offshore Bank Accounts," *Journal of Political Economy*, 10.1086/717455.

Annan, Francis (2022a) "Gender and Financial Misconduct: A Field Experiment on Mobile Money," January, 10.2139/ssrn.3534762.

——— (2022b) "Misconduct and Reputation under Imperfect Information," January, 10.2139/ssrn.3691376.

Balyuk, Tetyana (2022) "FinTech Lending and Bank Credit Access for Consumers," *Management Science*, 10.1287/mnsc.2022.4319.

Blackmon, William, Rafe Mazer, and Shana Warren (2021) "Kenya Consumer Protection in Digital Finance Survey Report,"Technical report.

Brenner, Lukas, Tobias Meyll, Oscar Stolper, and Andreas Walter (2020) "Consumer Fraud Victimization and Financial Well-Being," *Journal of Economic Psychology*, 76, 102243, 10.1016/j.joep.2019.102243.

Burke, Jeremy, Christine Kieffer, Gary Mottola, and Francisco Perez-Arce (2022) "Can Educational Interventions Reduce Susceptibility to Financial Fraud?" *Journal of Economic Behavior & Organization*, 198, 250–266, 10.1016/j.jebo.2022.03.028.

Chen, Yan, Iman YeckehZaare, and Ark Fangzhou Zhang (2018) "Real or Bogus: Predicting Susceptibility to Phishing with Economic Experiments," *PLOS ONE*, 13 (6), e0198213, 10.1371/journal.pone.0198213.

CrowdTangle, Team (2020) "CrowdTangle," *Facebook, Menlo Park, California, United States.*

DeLiema, Marguerite, Martha Deevy, Annamaria Lusardi, and Olivia S. Mitchell (2020) "Financial Fraud Among Older Americans: Evidence and Implications," *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, 75 (4), 861–868, 10.1093/geronb/gby151.

Engels, Christian, Kamlesh Kumar, and Dennis Philip (2020) "Financial Literacy and Fraud Detection," *The European Journal of Finance*, 26 (4-5), 420–442, 10.1080/1351847X.2019.1646666.

Ensminger, Jean and Jetson Leder-Luis (2022) "Detecting Fraud in Development Aid," December, 10.3386/w30768.

Financial Institution Regulatory Authority, Investor Education Foundation (2015) "The Non-Traditional Costs of Financial Fraud: Report of Survey Findings,"Technical report, Applied Research and Consulting: New York, NY.

Fu, Jonathan and Mrinal Mishra (2022) "Fintech in the Time of COVID-19: Technological Adoption during Crises," *Journal of Financial Intermediation*, 50, 100945, 10.1016/j.jfi.2021.100945.

Garz, Seth, Xavier Giné, Dean Karlan, Rafe Mazer, Caitlin Sanford, and Jonathan Zinman (2021) "Consumer Protection for Financial Inclusion in Low- and Middle-Income Countries: Bridging Regulator and Academic Perspectives," *Annual Review of Financial Economics*, 13 (1), 219–246, 10.1146/annurev-financial-071020-012008.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2008) "Trusting the Stock Market," *The Journal of Finance*, 63 (6), 2557–2600, 10.1111/j.1540-6261.2008.01408.x.

Gurun, Umit G, Noah Stoffman, and Scott E Yonker (2017) "Trust Busting: The Effect of Fraud on Investor Behavior," *The Review of Financial Studies*, 31 (4), 1341–1376, 10.1093/rfs/hhx058.

Johnson, Eric J, Stephan Meier, and Olivier Toubia (2019) "What's the Catch? Suspicion of Bank Motives and Sluggish Refinancing," *The Review of Financial Studies*, 32 (2), 467–495, 10.1093/rfs/hhy061.

Koyama, Naoko, Swetha Totapally, Shruti Goyal, Petra Sonderegger, Priti Rao, and Jasper Gosselt (2021) "Kenya's Digital Economy: A People's Perspective,"Technical report.

Lusardi, Annamaria and Olivia S. Mitchell (2014) "The Economic Importance of Financial Literacy: Theory and Evidence," *Journal of Economic Literature*, 52 (1), 5–44, 10.1257/jel.52.1.5.

McAlvanah, Patrick, Keith Anderson, Robert Letzler, and Jack Mountjoy (2015) "Fraudulent Advertising Susceptibility: An Experimental Approach,"Technical report.

Moustafa, Ahmed A., Abubakar Bello, and Alana Maurushat (2021) "The Role of User Behaviour in Improving Cyber Security Management," *Frontiers in Psychology*, 12.

Norris, Gareth, Alexandra Brookes, and David Dowell (2019) "The Psychology of Internet Fraud Victimisation: A Systematic Review," *Journal of Police and Criminal Psychology*, 34 (3), 231–245, 10.1007/s11896-019-09334-5.

Pazarbasioglu, Ceyla, Alfonso Garcia Mora, Mahesh Uttamchandani, Harish Natarajan, Erik Feyen, and Mathew Saal (2020) "Digital Financial Services," *World Bank Symposium*, 54.

Scheibe, Susanne, Nanna Notthoff, Josephine Menkin, Lee Ross, Doug Shadel, Martha Deevy, and Laura L. Carstensen (2014) "Forewarning Reduces Fraud Susceptibility in Vulnerable Consumers," *Basic and Applied Social Psychology*, 36 (3), 272–279, 10.1080/01973533.2014.903844.

Sheng, Steve, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge (2007) "Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish," in *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, 88–99, New York, NY, USA: Association for Computing Machinery, 10.1145/1280680.1280692.

Tables

Table 1: Correlates of Scam Identification Ability and Confidence

	SIA			Confidence in SIA		
	(1)	(2)	(3)	(1)	(2)	(3)
Demographics:						
Female	-0.03***	-0.03***	-0.03***	-0.11***	-0.10***	-0.12***
	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)
Age in Years	0.00	0.00	0.00	0.01**	0.01**	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Post-Seconday Education	0.03*	0.02	0.01	0.11**	0.10*	0.13**
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.06)
Low Income	0.02	0.02	0.02	0.04	0.03	0.04
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.05)
Formal Employment	-0.00	-0.01	-0.00	0.06	0.04	0.03
	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)
DFS Use:						
Low Trust in DFS		0.01	0.01		-0.11**	-0.09**
		(0.01)	(0.01)		(0.05)	(0.04)
Above mean use of different DFS		0.03**	0.03**		0.05	0.03
		(0.01)	(0.01)		(0.04)	(0.04)
Scam Experience:						
Contacted less than 1 week ago			-0.01			-0.01
			(0.02)			(0.06)
Victim of a Scammer			-0.01			-0.05
			(0.01)			(0.04)
N	997	997	956	997	997	956
R-Squared	0.05	0.05	0.05	0.03	0.04	0.04

Notes: Dependent variables are the SIA score in block 1 and average confidence ratings in block 1. *Female, Post-Secondary Education, Formal Employment, Low Trust in DFS, Contacted less than 1 week ago*, and *victim of a scammer* are binary indicators, *Low Income* and *Above mean use of different DFS* are binary indicators for median splits. All variables rely on self-reports. All specifications control for the order of the two blocks and failing the attention check. The displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

Table 2: Treatment Effects

Panel 1: Main Outcomes

	Correctly Identified Messages			Confidence		
	SIA	Scams	Non-scams	SIA	Scams	Non-scams
Tips (unincentivized)	0.02	0.08***	-0.09***	0.13***	0.16***	0.07
	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)	(0.06)
Tips (incentivized)	0.03*	0.08***	-0.07**	0.09**	0.09*	0.09
	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)	(0.06)
Control Mean	0.70	0.69	0.71	4.20	4.20	4.33
p-value ($Tips^U = Tips^I$)	0.56	0.89	0.40	0.43	0.18	0.78
N	956	956	956	956	956	956
R-Squared	0.04	0.10	0.16	0.46	0.40	0.26

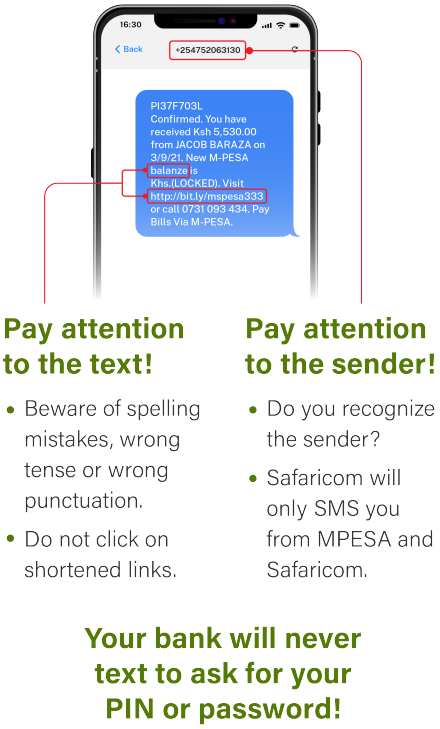
Panel 2: Secondary Outcomes

	Trust in DFS	Response Time SIA	All Scams Identified	All Non-scam Identified
Tips (unincentivized)	0.01	0.12	0.10**	-0.11***
	(0.07)	(0.08)	(0.04)	(0.04)
Tips (incentivized)	-0.00	0.23**	0.11**	-0.08*
	(0.07)	(0.10)	(0.04)	(0.05)
Control Mean	2.02	2.21	0.30	0.52
p-value ($Tips^U = Tips^I$)	0.92	0.26	0.99	0.42
N	956	956	956	956
R-Squared	0.03	0.34	0.07	0.11

Notes: In Panel 1, the dependent variables are the scam identification ability (SIA) score in block 2, the share of correctly identified scams in block 2, the share of correctly identified non-scams in block 2, and the average confidence ratings in block 2 for all messages (confidence in SIA), for the scam messages, and for the non-scam messages. In Panel 2, the dependent variables are Trust in DFS, the time spent on SIA in block 2, a binary indicator for classifying all scams correctly in block 2, and a binary indicator for classifying all non-scams correctly in block 2. All specifications include an indicator for the incentives treatment, the value of the outcome variable in block 1 (except for trust which was only measured after block 2), and the full set of controls, i.e., variables displayed in Table 1 (female, age, post-secondary education, low income, formal employment, low trust in DFS (except for the effect on trust), above mean use of different DFS, contacted less than one week ago, victim of a scammer), as well as indicators for the order of the two blocks and failing the attention check. $Tips^U$ and $Tips^I$ refer to Tips (unincentivized) and Tips (incentivized), respectively. The displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

Figures

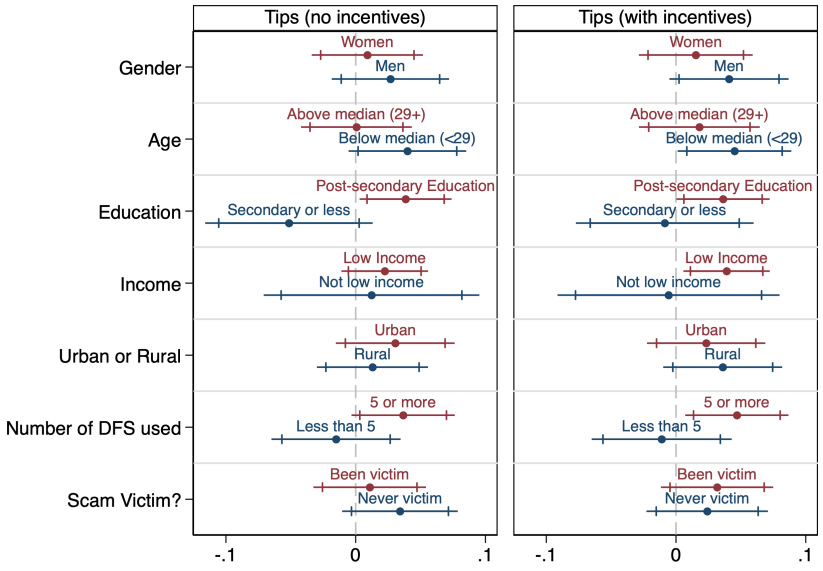
Figure 1: Tips treatment



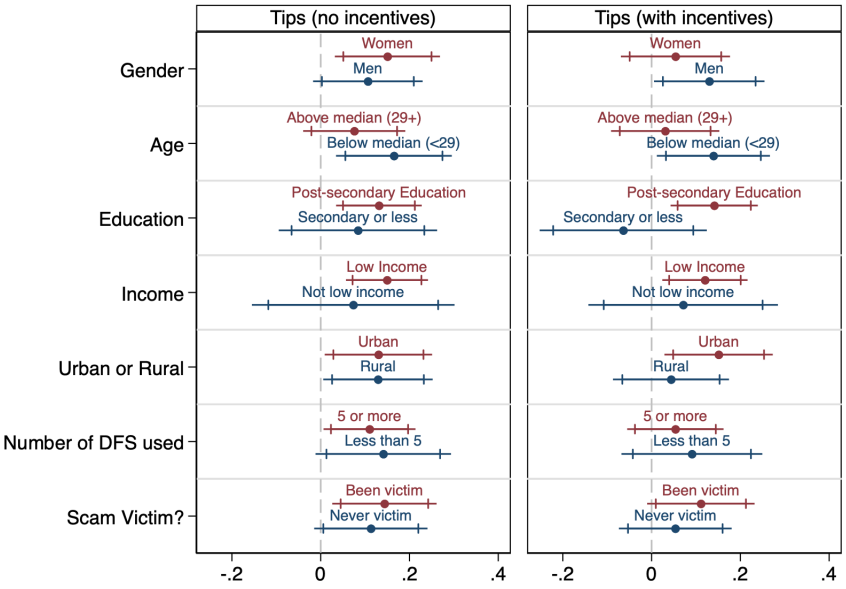
Notes: Tips treatment was designed based on commonly communicated tips in Kenya. The graphic was "animated," such that the pieces of information would be shown step-by-step. Participants clicked through this animation at their own speed, i.e., they hit the "continue" button five times before they see the overall graphic.

Figure 2: Treatment Effect Heterogeneity

(a) Scam Identification Ability (SIA)

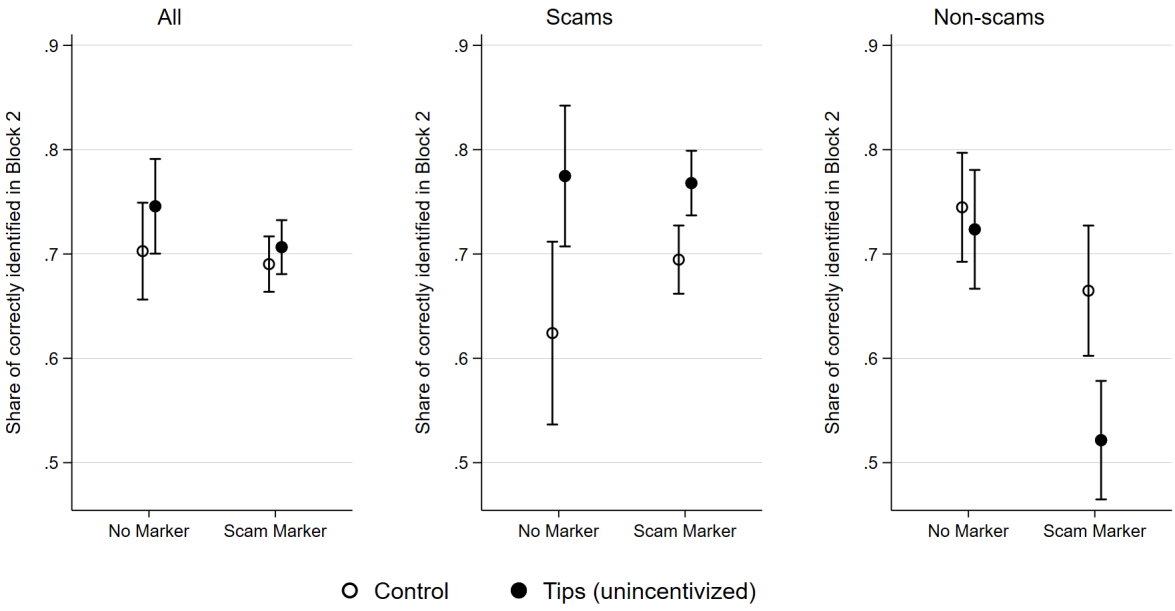


(b) Confidence



Notes: Figures plot the OLS coefficients and the 90% and 95% confidence intervals from the estimating regressions in Panel 1, Table 2 (Column 1 for SIA and Column 4 for Confidence) separately for the different subcategories.

Figure 3: Vignette-level effects by whether the message contains a scam marker



Notes: Figures plot the average marginal effects of triple-differences estimation with 95% confidence intervals based on standard errors clustered at the respondent level (see also Appendix D). Scam Marker is an indicator for whether the message contains at least one of the scam markers the tips warn about. The left panel contains all vignettes, the center panel focuses on scams, and the right panel on non scams. For ease of exposition, only the control and the Tips (unincentivized) treatment are displayed. The empirical specification contains the full set of interactions and individual-level controls.

Online Appendix

A Additional Tables

Table A1: Overview of vignettes

	Content	Intention	Sender
Block A	M-PESA transfer receipt	Genuine	Displayed
	Offer to use the new M-PESA app and get cash back	Genuine	Not displayed
	Random message to encourage contact	Fraudulent	Displayed
	Investment opportunity	Fraudulent	Displayed
	Suspended bank account	Fraudulent	Not displayed
	Notification as emergency contact	Fraudulent	Not displayed
Block B	M-PESA reversal request	Genuine	Displayed
	Notification of new registered SIM	Genuine	Displayed
	Job offer	Fraudulent	Displayed
	Lottery win	Fraudulent	Displayed
	Covid-19 relief fund	Fraudulent	Not displayed
	Notification as loan grantor	Fraudulent	Displayed

Notes: Vignettes kept the original wording of the screenshots and were visually harmonized, i.e., all vignettes were displayed on the same phone with the same signal strength, battery level, etc. (see also Figure A1). The order of the blocks was randomized at the individual level, as was the order of vignettes within a block. As blocks might be different for various reasons, we always control for the order of the blocks in the analysis.

Table A2: Balance at Baseline

	(1) Control		(2) Tips (unincentivized)		(3) Incentives		(4) Tips (incentivized)		T-test P-value						F-test for joint
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)	orthogonality
Female (0/1)	256	0.52	259	0.50	246	0.50	239	0.48	0.76	0.80	0.44	0.96	0.64	0.62	0.90
		(0.03)		(0.03)		(0.03)		(0.03)							
Age	256	32.23	259	32.55	246	32.07	239	32.26	0.72	0.85	0.97	0.58	0.76	0.83	0.96
		(0.61)		(0.62)		(0.59)		(0.67)							
Urban (0/1)	256	0.47	259	0.47	246	0.52	238	0.53	0.97	0.21	0.18	0.20	0.17	0.91	0.32
		(0.03)		(0.03)		(0.03)		(0.03)							
Post secondary education (0/1)	256	0.78	259	0.74	246	0.71	239	0.70	0.34	0.09*	0.05**	0.45	0.29	0.76	0.18
		(0.03)		(0.03)		(0.03)		(0.03)							
Low income (0/1)	256	0.80	259	0.81	246	0.74	239	0.77	0.69	0.09*	0.40	0.03**	0.22	0.38	0.15
		(0.03)		(0.02)		(0.03)		(0.03)							
Formal employment (0/1)	256	0.36	258	0.38	245	0.37	238	0.35	0.77	0.92	0.74	0.84	0.53	0.67	0.94
		(0.03)		(0.03)		(0.03)		(0.03)							
Internet on phone (0/1)	256	0.99	258	0.99	246	1.00	239	0.99	0.66	0.58	0.95	0.33	0.71	0.55	0.78
		(0.01)		(0.01)		(0.00)		(0.01)							
Social media on phone (0/1)	256	0.99	258	0.99	246	0.99	239	1.00	0.66	0.62	0.60	0.95	0.35	0.33	0.69
		(0.01)		(0.01)		(0.01)		(0.00)							
Recent DFS use (0/1)	251	0.95	253	0.94	240	0.95	236	0.97	0.70	0.92	0.30	0.63	0.16	0.36	0.50
		(0.01)		(0.01)		(0.01)		(0.01)							
Number of DFS used	256	4.85	259	4.79	246	4.68	239	4.79	0.80	0.47	0.81	0.63	0.99	0.62	0.90
		(0.16)		(0.16)		(0.16)		(0.16)							
Above mean use of DFS (0/1)	256	0.62	259	0.60	246	0.61	239	0.63	0.60	0.72	0.81	0.87	0.45	0.55	0.87
		(0.03)		(0.03)		(0.03)		(0.03)							

Notes: Asterisks indicate that the difference is statistically significant at the 1% ***, 5% **, and 10% * levels.

Table A3: Sample and Kenyan Population

	Online Survey	**Kenya National Bureau of Statistics (2019)**
	Fraction	Fraction of total population (adults)
Gender		
Female	50.1%	50.5%
Male	49.9%	49.5%
Age		
18-24	32.1%	13.4% (24.8%)
25-34	27.0%	15.6% (29.0%)
35+	40.9%	24.9% (46.3%)
County		
Baringo	1.4%	1.4%
Bomet	1.9%	1.8%
Bungoma	3.7%	3.5%
Busia	2.0%	1.9%
Elgeyo-Marakwet	1.0%	1.0%
Embu	1.3%	1.3%
Garissa	1.6%	1.8%
Homa Bay	2.5%	2.4%
Isiolo	0.4%	0.6%
Kajiado	1.7%	2.4%
Kakamega	4.4%	3.9%
Kericho	1.9%	1.9%
Kiambu	4.4%	5.1%
Kilifi	3.0%	3.1%
Kirinyaga	1.4%	1.3%
Kisii	2.9%	2.7%
Kisumu	2.5%	2.4%
Kitui	2.6%	2.4%
Kwale	1.7%	1.8%
Laikipia	1.0%	1.1%
Lamu	0.3%	0.3%
Machakos	2.8%	3.0%
Makueni	2.4%	2.1%
Mandera	2.5%	1.8%
Marsabit	0.8%	1.0%
Meru	3.4%	3.2%
Migori	2.4%	2.3%
Mombasa	2.4%	2.5%
Murang'a	2.4%	2.2%
Nairobi	8.3%	9.2%
Nakuru	4.2%	4.5%
Nandi	2.0%	1.9%
Narok	2.2%	2.4%
Nyamira	1.5%	1.3%
Nyandarua	1.5%	1.3%
Nyeri	1.8%	1.6%
Samburu	0.7%	0.7%
Siaya	2.1%	2.1%
Taita-Taveta	0.7%	0.7%
Tana River	0.6%	0.7%
Tharaka-Nithi	0.9%	0.8%
Trans Nzoia	2.1%	2.1%
Turkana	2.1%	1.9%
Uasin Gishu	2.3%	2.4%
Vihiga	1.4%	1.2%
Wajir	1.6%	1.6%
West Pokot	1.3%	1.3%
Location		
Urban	49.6%	31.2%
Rural	50.3%	68.8%

Table A4: Sample characteristics

	N	Mean	SD	Min.	Max.
Demographics					
Female (0/1)	1000	0.50	0.50	0	1
Age	1000	32.28	9.84	18	67
Urban (0/1)	999	0.50	0.50	0	1
Post secondary education (0/1)	1000	0.73	0.44	0	1
Low income	1000	0.78	0.41	0	1
Formal employment (0/1)	997	0.36	0.48	0	1
Internet on phone (0/1)	999	0.99	0.09	0	1
Social media on phone (0/1)	999	0.99	0.09	0	1
Financial transactions w/ phone in the past 90 days	980	0.96	0.21	0	1
DFS Use					
Number of DFS used	1000	4.78	2.52	0	9
Scam Experience					
Have you ever been contacted by a scammer?	999	0.96	0.18	0	1
Ever been a victim of a scammer?	960	0.56	0.50	0	1
Anyone you know ever been a victim of a scammer?	1000	0.85	0.35	0	1
Scam Identification Ability (Block 1)					
Share of correctly identified messages (SIA)	1000	0.71	0.18	0	1
Share of correctly identified scams	1000	0.74	0.24	0	1
Share of correctly identified non-scams	1000	0.66	0.35	0	1
Average confidence in SIA	1000	4.23	0.63	1	5

Table A5: The Effects of Incentives

Panel 1: Outcomes in Block 1

	SIA	Scams Identified	Non-scams Identified	Confidence
Incentives	-0.01	-0.02	0.00	0.06
	(0.02)	(0.02)	(0.03)	(0.06)
Control Mean	0.71	0.75	0.63	4.23
N	956	956	956	956
R-Squared	0.05	0.04	0.09	0.04

Panel 2: Outcomes in Block 2

	SIA	Scams Identified	Non-scams Identified	Confidence
Incentives	0.02	0.02	0.03	0.08*
	(0.02)	(0.02)	(0.03)	(0.04)
Control Mean	0.70	0.69	0.71	4.20
N	956	956	956	956
R-Squared	0.04	0.10	0.16	0.46

Notes: In Panels 1 and 2, dependent variables are the SIA score, the share of correctly identified scams, the share of correctly identified non-scams, and the average confidence ratings in block 1 and 2, respectively. All specifications include indicators for the tips treatments, and the full set of controls, i.e., variables displayed in Table 1 (female, age, post-secondary education, low income, formal employment, low trust in DFS, above mean use of different DFS, contacted less than one week ago, victim of a scammer), as well as indicators for the order of the two blocks and failing the attention check. In Panel 2, the additional controls are baseline values of the outcome variables in block 1. The displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

Table A6: Confidence weighted by SIA

	All messages	Scams	Non scams
Tips (unincentivized)	0.05*	0.16***	-0.17***
	(0.03)	(0.04)	(0.05)
Tips (incentivized)	0.06**	0.15***	-0.12**
	(0.03)	(0.04)	(0.05)
Control Mean	.36	.34	.38
p-value ($Tips^U = Tips^I$)	.81	.85	.41
N	956	956	956
R-Squared	.042	.12	.17

Notes: The dependent variable are weighted confidence in block 2 for all messages, scams, and non-scams, respectively. Weighted confidence ranges from -1 to 1, where 1 means fully confident and perfect SIA score, whereas -1 means fully confident and no correctly classified vignette. All specifications include an indicator for the incentives treatment, the value of the outcome variable in block 1, and the full set of controls, i.e., variables displayed in Table 1 (female, age, post-secondary education, low income, formal employment, low trust in DFS (except for the effect on trust), above average use of different DFS, contacted less than one week ago, victim of a scammer), as well as indicators for the order of the two blocks and failing the attention check. $Tips^U$ and $Tips^I$ refer to Tips (unincentivized) and Tips (incentivized), respectively. The displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

Additional Figures

Figure A1: Example Vignette

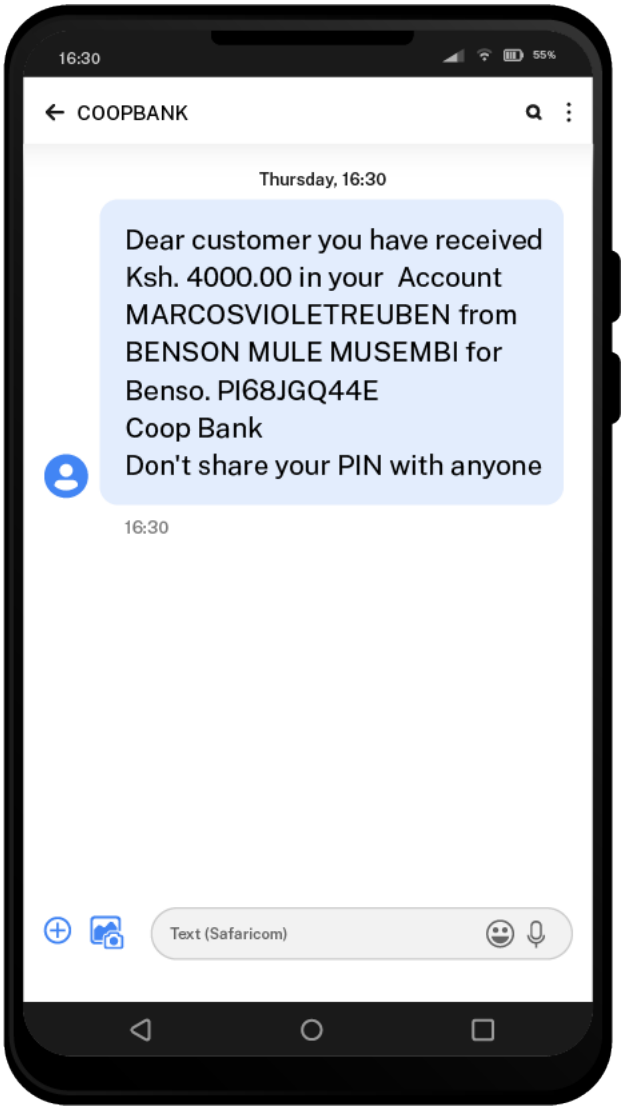
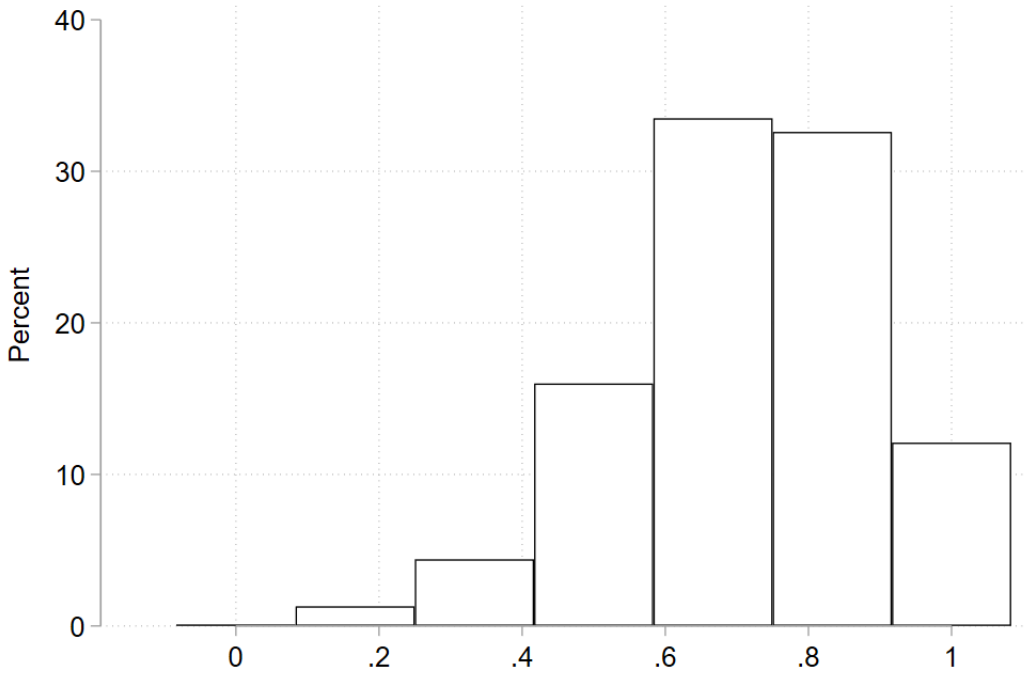
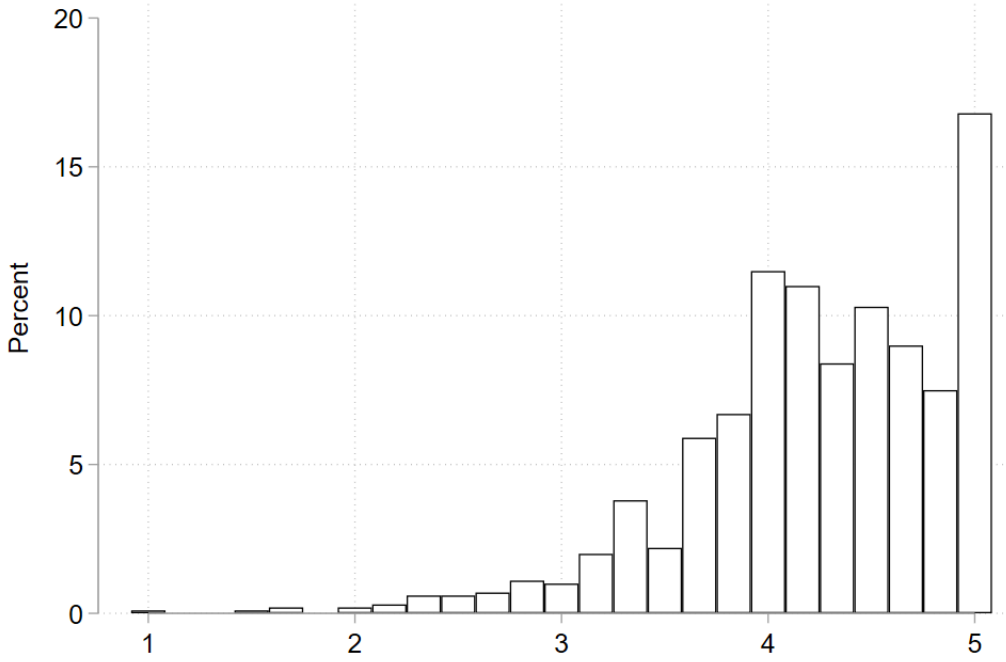


Figure A2: SIA and Confidence in Block 1

(a) Distribution of scam identification ability (SIA)



(b) Distribution of Confidence in SIA



B Robustness

We briefly address robustness with a focus on attention. One might be worried that participants of the online survey did not pay attention to the classification task or in general. We address a potential lack of attention in the classification task with the incentive treatment. We show that incentives have no significant effect on our primary outcomes (see Table A5), and they make no difference in the Tips treatments (see Tables 2 and A6). We interpret this as evidence that participants pay attention to the classification tasks. Paying participants for accuracy does not improve their SIA, which we interpret as evidence that participants in the non-incentivized classification tasks provide their best effort. From a methodological point, this implies that our measure of SIA can be used without incentives.

To mitigate the concern regarding general attention, we implemented an attention check, which was administered after the classification task. About 27% of the participants failed the attention check. Columns 1-4 of Table B1 exclude those who fail the attention check and replicate the main results from Table 2. Results are similar in terms of direction, magnitude, and significance. In the last column of Table B1, we show that those who receive both tips and incentives are slightly less likely to fail the attention check. In that sense, the first four columns show results for a selected sample. Regarding further robustness checks, we note that our sample is balanced (see Table A2) and that our results are robust to the inclusion of different sets of control variables or none at all (see Table B2).

Table B1: Attention check and main treatment effects

	Those who passed attention check				All
	SIA	Scams Identified	Non-scams Identified	Confidence	Failed Attention
Tips (unincentivized)	0.20*	0.10***	-0.09***	0.16***	-0.01
	(0.11)	(0.03)	(0.03)	(0.05)	(0.04)
Tips (incentivized)	0.18*	0.09***	-0.09**	0.07	-0.10**
	(0.11)	(0.03)	(0.03)	(0.05)	(0.04)
Control Mean	4.17	0.69	0.71	4.20	0.30
p-value ($Tips^U = Tips^I$)	0.85	0.83	0.90	0.09	0.03
N	699	699	699	699	956
R-Squared	0.05	0.11	0.18	0.44	0.05

Notes: Columns 1-4 include only those who passed the attention check. Column 5 includes all participants. Dependent variables are the SIA score in block 2, the share of correctly identified scams in block 2, the share of correctly identified non-scams in block 2, the average confidence ratings in block 2, and an indicator for failing the attention check. All specifications include an indicator for the incentives treatment, the value of the outcome variable in block 1 (except for the attention check which was only administered once), and the full set of controls, i.e., variables displayed in Table 1 (female, age, post-secondary education, low income, formal employment, low trust in DFS, above mean use of different DFS, contacted less than one week ago, victim of a scammer), as well as indicators for the order of the two blocks and failing the attention check (except for the last specification where the attention check is the dependent variable). T1 and T2 refer to Tips (unincentivized) and Tips (incentivized), respectively. The displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

Table B2: Treatment Effects: Varying Control Variables

Panel 1: Scam Identification Ability

	(1)	(2)	(3)	(4)
Tips (unincentivized)	0.02	0.02	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Tips (incentivized)	0.02	0.03*	0.03*	0.03*
	(0.02)	(0.02)	(0.02)	(0.02)
Demographics		✓	✓	✓
Design Controls			✓	✓
Scam Experience				✓
Control Mean	0.70	0.70	0.70	0.70
p-value ($Tips^U = Tips^I$)	0.81	0.59	0.70	0.56
N	1000	997	997	956
R-Squared	0.01	0.03	0.04	0.04

Panel 2: Confidence

	(1)	(2)	(3)	(4)
Tips (unincentivized)	0.12***	0.11***	0.12***	0.13***
	(0.04)	(0.04)	(0.04)	(0.04)
Tips (incentivized)	0.08*	0.08*	0.08*	0.09**
	(0.04)	(0.04)	(0.04)	(0.04)
Demographics		✓	✓	✓
Design Controls			✓	✓
Scam Experience				✓
Control Mean	4.20	4.20	4.20	4.20
p-value ($Tips^U = Tips^I$)	0.36	0.39	0.37	0.43
N	1000	997	997	956
R-Squared	0.45	0.46	0.46	0.46

Notes: Dependent variables in Panel 1 and Panel 2 are scam identification ability and average confidence in block 2, respectively. Demographic controls include gender, age, post-secondary education, low income, formal employment), design controls include indicators for order of blocks and a dummy for attention check, controls for scam experience include low trust in DFS, above average use of different DFS, contacted less than one week ago, and victim of a scammer. Dependent variable is the SIA score in block 2. All specifications include an indicator for the incentives treatment and the baseline value of the outcome variable. $Tips^U$ and $Tips^I$ refer to Tips (unincentivized) and Tips (incentivized), respectively. Displayed coefficients are from OLS regressions. Robust standard errors are in parenthesis. Asterisks indicate that the estimate is statistically significant at the 1% ***, 5% **, and 10% * levels.

C Scam Data Collection and SIA Measurement

C.1 Collecting examples of scam messages

We used Brandwatch to obtain public posts mainly from Twitter between January 2020 and June 2021. Posts included at least one scam-related keyword (see Appendix C.4) and the location was identified as Kenya. Of the overall 427,121 posts, we focus on the 58,804 original tweets (not replies or retweets) from individual accounts on Twitter. Of those 11,640 include a picture. We use topic clustering of the post text to identify different subjects. We download 800 pictures: 75% are drawn from the topics that we identify as relevant, and 25% from other topics.[12] Out of the 800 pictures, we select those which display a screenshot of a text message independent of the text of the message.[13] After removing duplicates, we are left with 116 screenshots of potential scam messages.

We also used Crowdtangle (CrowdTangle, 2020) to obtain public posts from Facebook that contain at least one scam-related keyword and were posted by a profile located in Kenya between June 2020 and June 2021. Overall, we obtained 18002 posts of which 4328 included the keyword "fraud" and 2244 the keyword "scam." However, most posts were sent from official company accounts (often media companies) and only 1089 (6%) were from individuals. Manually verifying the posts from individuals and pictures revealed that the scam messages shared were related to Facebook scam attempts and therefore were out of the scope of this paper.

In September 2021, we launched a survey in the largest Kenyan fraud-detection Facebook group with the consent of the administrators. In this anonymous survey, we asked participants a couple of questions about their phone scam experience. We then encouraged them to submit examples of scam messages and calls, as well as messages and calls from official sources. Within three days, 919 people had completed the questionnaire.

Only around 8% indicated that they had not received a scam message within the last month. Most participants stated to have received between 1 and 5 scam messages, and

[12] We cluster for 6 topics: Three seem more related to political or organizational fraud as they include words such as "BBI", "court", "business" as frequent words. The three topics identified as relevant included words such as "mpesa", "cash", "win", "paybill" as frequent words. The topic clustering algorithm computes for each tweet a score of how much it speaks to each topic. We classify each tweet by its main topic - the topic with the highest score.

[13] We first check using a text scraping function if there is any text on the picture and keep only those with a text of more than three words. We then manually select the screenshots.

around 12% received more than 10. Roughly half of the participants stated that they received one scam call in the past month with 24% not receiving any, and around 10% more than 5. Official messages and calls were much rarer: 25% of participants did not receive any official message within the past month, and nearly 50% did not receive an official call. Overall, participants submitted 567 examples of (supposed) scams and 355 examples of (supposed) official messages.

While this data collection leaves us with real examples of scams and official communication, collected examples are likely not representative, potentially missing both obvious scams and very good scams.

C.2 Creation of Vignettes

We use the labeled Twitter and survey data of examples for scams and official messages and keep SMS examples that were submitted either as a picture or copy-pasted text (n=1,836). From the focus group discussions, we learned that some scams are near-universally recognized while others are much harder to spot. To generate variation in the measure of SIA, we create a database of ambiguous messages. For this, we use the certainty rating of coders who indicate their confidence in their classification on a scale from 1 (very confident) to 5 (not at all confident). In addition, we examine whether the two coders agree in their classification. We then build a dataset with ambiguous messages for which the two coders disagree and at least one coder is not fully confident. We also include messages for which the average confidence rating is smaller than three.

C.3 Qualitative Data Collection

To complement the social media data, we interviewed six stakeholders in (digital) financial services in Kenya, including the Captial Markets Authority, Credit Information Sharing of Kenya, and Financial Sector Deepening. We additionally conducted five focus groups with a diverse group of individuals who are not active on social media. Results from the qualitative data collection on scam perceptions are in line with the scam examples we find online.

Stakeholders highlight that awareness in the population is mixed and postulate that it varies by the frequency of DFS use. The youth and the elderly, women, and rural populations are perceived to be especially vulnerable. Both interviewees and focus group participants

recommend continuous education of consumers, mass sensitization on phone scams, and stronger law enforcement. Stakeholders further recommend stronger collaboration among all parties involved, standardization and consistency in communication by providers, and the implementation of technological innovations such as biometric identification.

C.4 Scam-related Keywords

Table D1: Scam Keywords

Keyword	Description (if necessary)
Scam	
Scams	
Phone Scam	
Financial Scam	
Scammer	
Scammers	
Fraud	
Fraudster	
Digital Security	
Financial crime	
Financial fraud	
Digital Fraud	
Payment fraud	
Report fraud	
Scam call	
Scam calls	
Scam message	
Scam messages	
Fraud risk	
Tuwaanike	A campaign by Safaricom aimed at exposing fraudsters and equipping the public with information to raise the alarm on them and their tactics.
333	Safaricom's official toll-free SMS line to report fraudsters and forward their message and phone numbers for further action.
707	Safaricom's official phone number that will automatically SMS you if someone tries to register your SIM. If you reply no, it won't be registered in your place.
SIM swap	A popular scam where someone experiences their SIM card getting duplicated without their knowledge after they are socially engineered (typically over a phone call) to provide their data e.g. ID number and some more sensitive info.
"Kaa Chonjo"	A campaign by the Kenya Bankers Association aimed at equipping people with the fact that they should keep their ATM PIN a secret and stay alert at ATMs for fraudsters.
M-Pesa PIN	
Hakikisha	A tool by Safaricom to authenticate the identity of the person you're about to send money to (sometimes used to show that the fraudster uses a fake name other than the one on their ID card).
Kamiti	
Bank Account Takeover Fraud	
Debit Card Fraud	
Credit Card Fraud	
Yahoo boys	This is in reference to Nigeria scammers
Internet fraud	
mail fraud	
pyramid schemes	
identity theft	
bank card fraud	
ponzi schemes	
advance fee scam	
swift/rtgs message fraud	
forgery	
manifest fraud	This occurs when shipping agents illegally alter manifests prior to uploading them to the Customs Manifest Management System (MMS), thereby setting the stage for false declarations.
invoice fraud	
tax fraud	
phishing	
scamming	
wash wash	

Notes: Keywords were determined based on discussions with Kenyan DFS experts.

D Vignette-Level Analysis

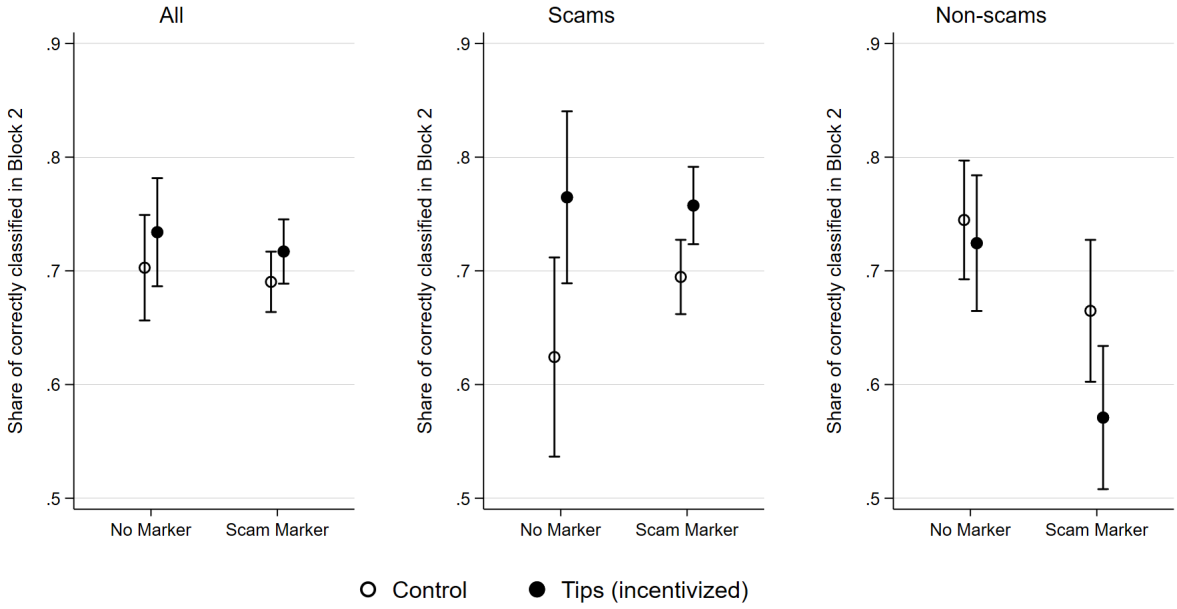
To assess treatment effects and the effects of scam markers at the vignette level, we estimate

$$\begin{aligned} y_{im} = \alpha_0 &+ \alpha_1 Tips_i^U + \alpha_2 Tips_i^I + \alpha_3 Incentives_i \\ &+ \alpha_4 Tips^U * Block2_{im} + \alpha_5 Tips_i^I * Block2_{im} \\ &+ \alpha_6 Block2_m + \alpha_7 ScamMarker_m + \alpha_8 Block2 * ScamMarker_m \\ &+ \alpha_9 Tips^U * ScamMarker_{im} + \alpha_{10} Tips^I * ScamMarker_{im} \\ &+ \alpha_{11} Tips^U * ScamMarker * Block2_{im} \\ &+ \alpha_{12} Tips^I * ScamMarker * Block2_{im} \\ &+ \alpha_{13} Incentives * Block2_{im} + X_i'\gamma + Other_i\delta + \epsilon_{im} \end{aligned} \tag{D1}$$

, where y_{im} denotes whether individual i has classified the message m correctly. The tips treatments are only in effect when the message is shown in block 2, hence we interact the treatment indicators with an indicator for block 2. We also include the interaction of Incentives and block 2 to allow for more flexibility. As the order of the blocks is randomized on the respondent level, this indicator varies on the individual and message level. X_i is a set of individual characteristics for respondent i. These include gender, age, income, and education level. $Other_i$ captures additional controls, such as the order of the two SIA blocks. We cluster standard errors on the individual level.

Figure D1 plots the average marginal effects obtained from our estimates for control and the tips (incentivized) treatment in block 2.

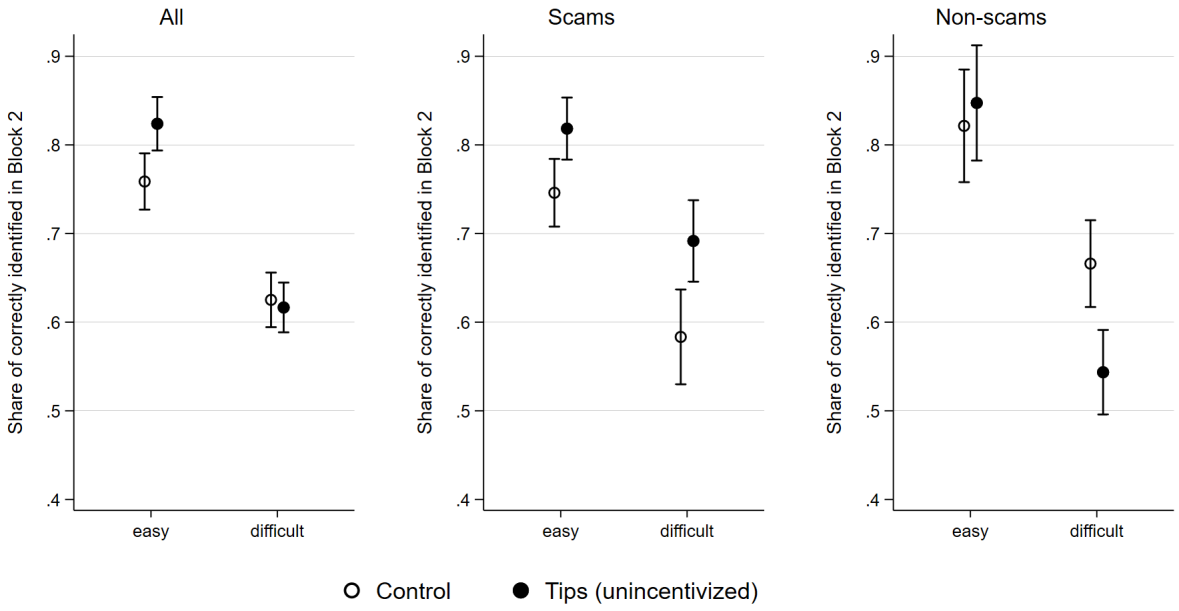
Figure D1: Vignette-level effects by whether the message contains a scam marker



Notes: Figures plot the average marginal effects from triple-differences estimation with 95% confidence intervals based on standard errors clustered at the respondent level (see also Appendix D). Scam Marker is an indicator for whether the message contains at least one of the scam markers the tips warn about. The left panel contains all vignettes, the center panel focuses on scams, and the right panel on non scams. For ease of exposition, only the control and the Tips (unincentivized) treatment are displayed. The econometric specification contains the full set of interactions and individual-level controls.

For the analysis of vignette difficulty, we replace the scam marker indicator in Equation D1 with an indicator for above median difficulty (based on how often a given message was classified incorrectly in block 1). We analyze treatment effects on the vignette level in block 2.[14] Figure D2 summarizes the results.

Figure D2: Vignette-level effects by difficulty of the vignette



Notes: Figures plot the average marginal effects from triple-differences estimation with 95% confidence intervals based on standard errors clustered at the respondent level (see also Appendix D). Vignettes are coded as difficult if they are misclassified more often than the median vignette in block 1. The left panel contains all vignettes, the center panel focuses on scams, and the right panel on non scams. For ease of exposition, only the control and the Tips (unincentivized) treatment are displayed. The econometric specification contains the full set of interactions and individual-level controls.

[14] A limitation of this analysis is that difficulty is not evenly distributed across the scam and non-scam vignettes: 3 out of 4 non-scam vignettes are difficult based on block 1 classifications, whereas 3 out of 8 scam vignettes are difficult.

E Discussion of Deviations from the Pre-Analysis Plan

The data collection proceeded as planned and there were no changes to the pre-registered experimental design. In a few instances, we deviate from the pre-analysis plan in the analysis, mostly for expositional clarity.

First, we now use a regression model in which all treatments enter as dummy variables rather than interacting the two treatment indicators. This allows us to omit the coefficient of the incentive treatment in the main tables while keeping the interpretability of coefficients.

Second, we use the share of correctly identified scam and non-scam messages rather than the number of incorrectly identified scam and non-scam messages (referred to as 'not cautious enough' and 'overly cautious' in the pre-analysis plan). This makes coefficients comparable across these two variables, changes the sign of the coefficients and makes the description of results easier to follow. For consistency, we also analyze SIA as the share of correctly identified messages, rather than the absolute number.

Third, we omit 'recent use of DFS' from the analysis as 95.5% indicate having used DFS, such that we do not have meaningful variation in this variable. We similarly exclude 'knowing a scam victim' as 84.4% state knowing someone who has been a victim of a scam. Fourth, we omit the LATE estimation for brevity. The estimation shows no significant results.

Finally, we note an error in the pre-analysis plan where we stated "We will control for attention (binary indicator for passing the attention check) in all our main analyses and will conduct robustness checks in which we include inattentive respondents." The first part of the sentence is what we do; the second part should read "exclude". Results for this exclusion are shown in Table B1.

The pre-analysis plan and the survey instrument can be accessed on the AEA registry after publication and are available upon request.