Ecological Validity in Studies of Security and Human Behaviour

Andrew Patrick, Ph.D.

Freelance Scientist, R&D Consultant

Adjunct Research Professor
Psychology & Computer Science, Carleton University

andrew@andrewpatrick.ca
http://andrewpatrick.ca

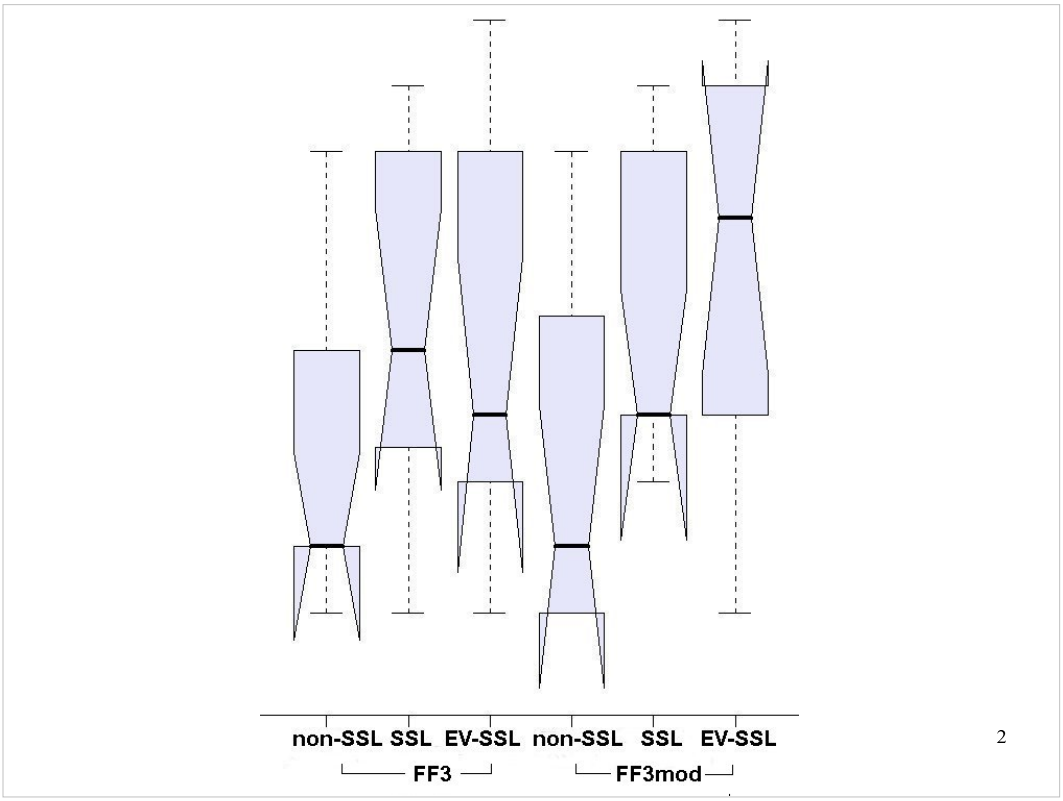
1

Abstract:

It is becoming increasingly clear that studies of the effectiveness of information security solutions must take into account the human factor -- the behaviour of the users of the systems. Conducting research on human behaviour is hard, however, and it is often difficult to witness authentic behaviour in a laboratory environment. Ecological validity refers to the extent to which the results of a test or experiment can be applied to the real-life of the people being studied. Using a series of case studies from research on security-related behaviours, Dr. Patrick will lead a discussion about the nature of validity in research, the issues surrounding ecological validity, and research techniques that can be used to increase the validity of security studies.

Biography:

Dr. Andrew Patrick is a Freelance Scientist and R & D Consultant. He is also an Adjunct Research Professor of Psychology and Computer Science at Carleton University. Until recently, Dr. Patrick led the Information Security Group at the National Research Council of Canada (NRC) where he conducted research on the human factors of security systems and tools for privacy protection. Dr. Patrick has also worked at Nortel and the Communications Research Centre (CRC). Dr. Patrick holds a Ph.D. in Cognitive Psychology from the University of Western Ontario. WWW Site: www.andrewpatrick.ca



Dependent Variable (DV) is willingness to do e-commerce transactions on the website.

Independent Variables (IV) are type of SSL certificate for plain and modified Firefox browsers

Findings are that only with the modified Firefox is there a benefit of EV certificates

Case: EV-SSL Indicators

Purpose: test new SSL interfaces for EV certificates

Subjects: N=28; age=18-29; male=16, female=12;
24 undergraduates, 2 comp. sci. students

Materials:
• plain or modified Firefox 3 browser; specially crafted self-served web sites

Context: university research lab.

Procedure:
• eye tracking
• locate 3 requested items on web site and record the prices

Dependent Variables: rating of willingness to make a purchase (10-point scale)

3

* the same size is small and limited to young university students
* the testing environment was artificial (university lab)
* eye-tracking is not a normal experience
* the task did not involve actual purchases
* the scale may not reflect actual willingness to purchase on these web sites

J. Sobey, R. Biddle, P.C. van Oorschot, A.S. Patrick. Exploring User Reactions to Browser Cues for Extended Validation Certificates. ESORICS 2008 - European Symposium on Research in Computer Security (to appear). October 6-8, 2008, Malaga, Spain.

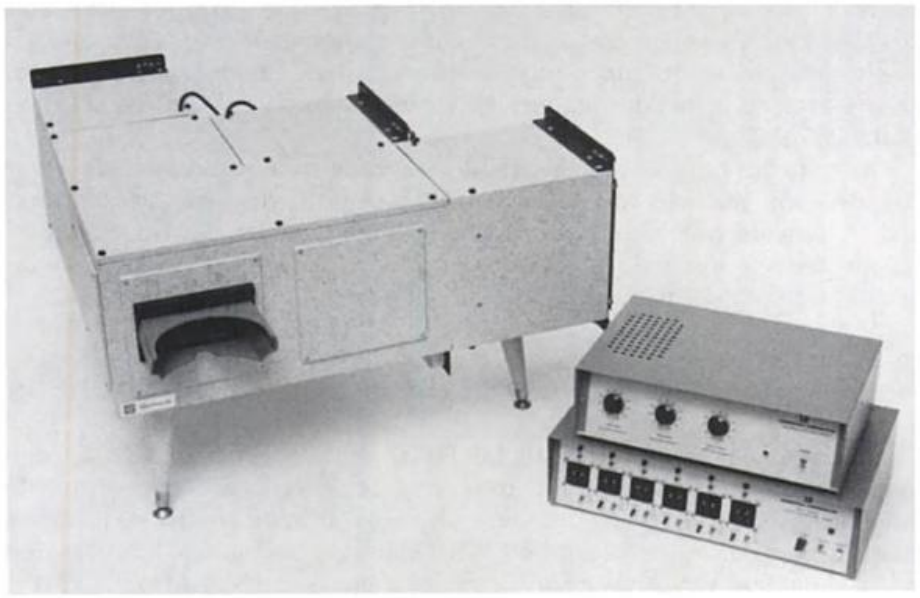
validity

the degree to which a test supports conclusions drawn from the results

the extent to which results are representative of the real world

What are these people doing? What can the psychologist learn by having the woman look into a box?

Figure 3. A modern three-channel mirror tachistoscope, showing the cabinet, lamp driver (smaller component), and timer units. (Photograph courtesy of Ralph Gerbrands Company.)



This is a tachistoscope – a device for very rapid display of visual materials.

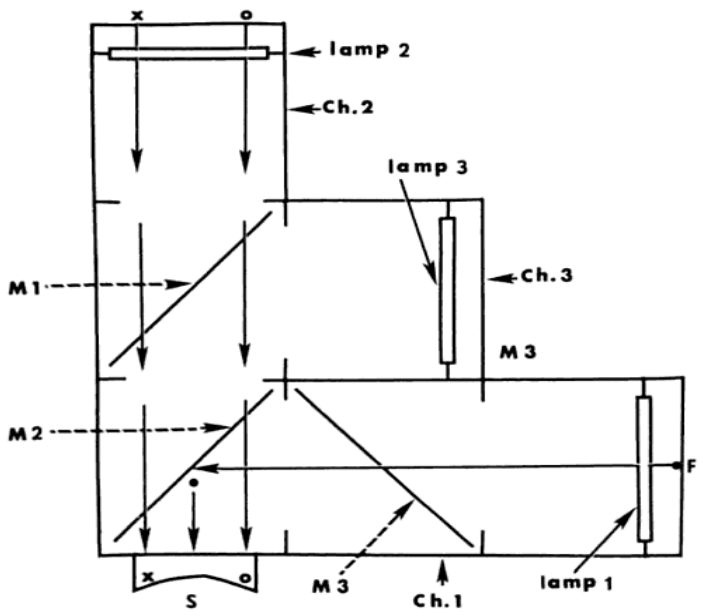
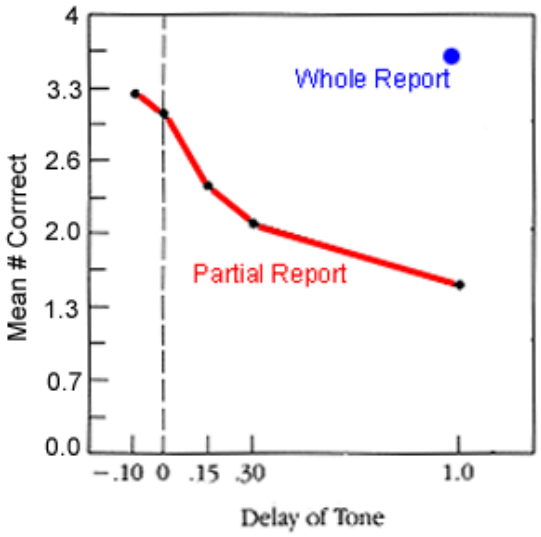


Figure 4. Interior arrangement of lamps and mirrors in a three-channel mirror tachistoscope. Abbreviations are explained in the text.

7

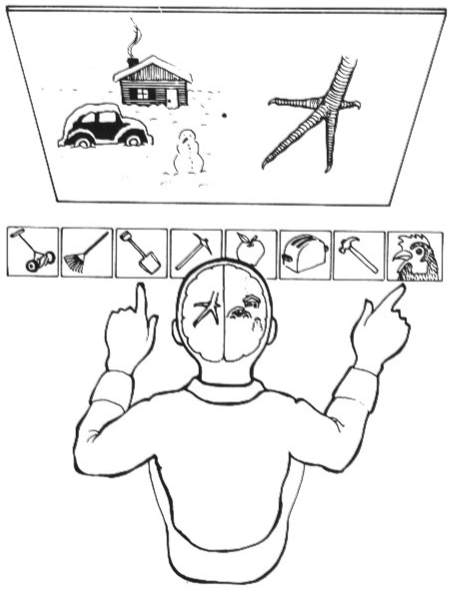
The cabinet contains half-silvered mirrors and special high-speed bulbs. By lighting different arms of the chamber, the subject can be shown different displays.

Often a fixation point is followed by the stimulus of interest. This might be later followed by a mask.



8

T-scope experiments were used to support the notion of an iconic memory that can hold visual information, but only for a short period (e.g., less than 1 sec).



9

It has also been used to study brain function. If the material is presented quickly enough, subjects (Ss) will not be able to move their eyes so each image will go to different visual fields, and thus different hemispheres of the brain. This has been useful for studying normal people, and those who have experienced brain surgery to separate the hemispheres.

face validity

construct validity

predictive validity

external validity

10

There are different kinds of validity:

The test is said to have **face** validity if it "looks like" it is going to measure what it is supposed to measure

construct validity refers to whether a scale measures or correlates with a theorized psychological construct

predictive validity is the extent to which a score on a scale or test predicts scores on some criterion measure

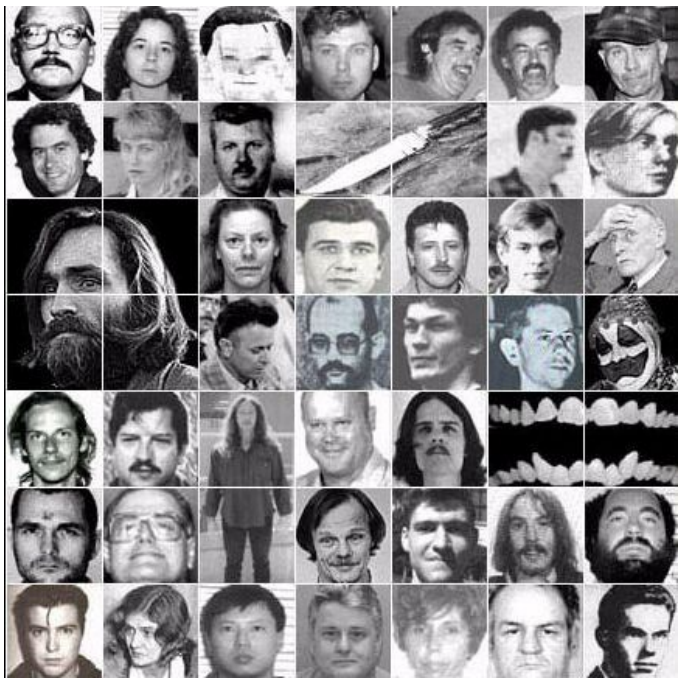
findings are said to possess **external** validity if they may be generalized from the unique and idiosyncratic settings, procedures and participants to other populations and conditions

Definitions from wikipedia.

11

Polygraph testing (lie detectors) have been questioned on all these aspects of validity.

It may have face validity, but it is lacking in construct, predictive, and external validity.



12

Criminal profiling is another example of a something that has general support and some face validity, but little predictive validity.

In the mid-nineties, the British Home Office analyzed a hundred and eighty-four crimes, to see how many times profiles led to the arrest of a criminal. The profile worked in five of those cases. That's just 2.7 per cent, which makes sense if you consider the position of the detective on the receiving end of a profiler's list of conjectures.

Malcolm Gladwell (2007-11-12). "Dept. of Criminology: "Dangerous Minds: Criminal profiling made easy."". The New Yorker (New York). http://www.newyorker.com/reporting/2007/11/12/071112fa_fact_gladwell. Retrieved on 2008-01-04.

“ecological validity refers to the extent to which the environment experienced by the subjects in a scientific investigation has the properties it is supposed or assumed to have by the experimenter” (Bronfenbrenner, 1977).

13

Ecological validity refers to the extent to which the results of a test or experiment can be applied to the real-life of the participants under study.

An experiment is said to be ecologically valid if it represents what real people would do in the real world.

An experiment is invalid if it examines only a **special sub-set** of people (e.g., university students) in an **artificial environment** (e.g., a psychology lab).

ecological validity

vs.

ethics

Ecological validity is often in conflict with considerations of ethics. For tests to be truly valid, it may be necessary to use unethical procedures or measurements. This is discussed later...

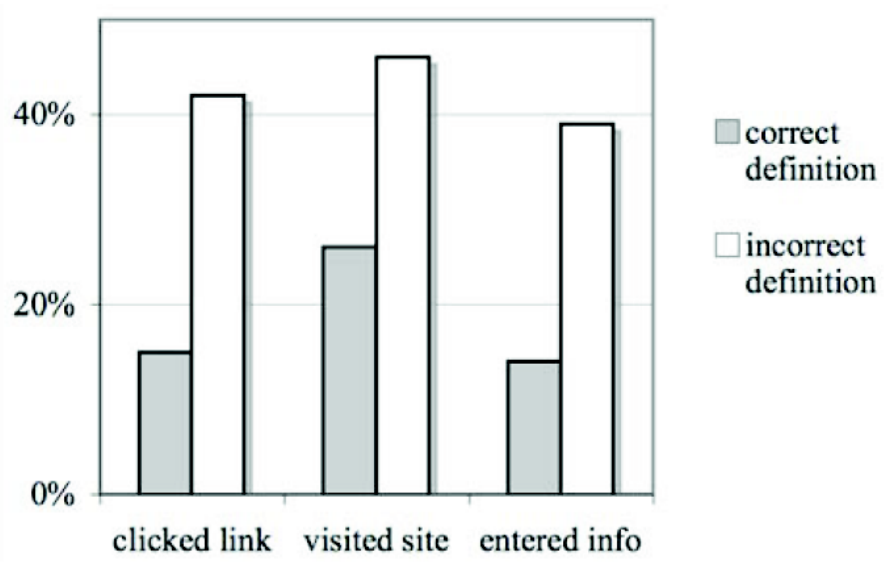


Figure 1. Behavior of participants who gave correct and incorrect definitions of phishing

DV is percent of people who fell for a phishing attack.

IVs are the type of failing behavior and prior knowledge about phishing.

Findings are that education about phishing can help people avoid phishing attacks.

Case: Phishing Survey

Purpose: test susceptibility to phishing

Subjects: N=232; age=?; faculty, staff, students; participants in "security summit", some "participated to make amends for violating computing policy such as exceeding allotted bandwidth"

Materials:
- **images of emails in in-box; contains URL link; 3 legitimate and 2 phishing**

Context: online survey; email role-playing

Procedure:
- **5 email messages**
- **seven response options: reply, phone, click link, etc.**

Dependent Variables:
- **% of people falling for phishing attacks**

16

* participants only come from university setting, and some have questionable motivations
* not real emails but images
* role-playing, not real behavior
* simulated responses to the emails

Downs, J. S., Holbrook, M., and Cranor, L. F. 2007. Behavioral response to phishing risk. In Proceedings of the Anti-Phishing Working Groups 2nd Annual Ecrime Researchers Summit (Pittsburgh, Pennsylvania, October 04 - 05, 2007). eCrime '07, vol. 269. ACM, New York, NY, 37-44.

representativeness

generalizability

17

Validity concerns two main concepts.

Is the group, materials, procedure representative of the population of interest?

Can the results be generalized beyond the specific people, materials, procedures?

Challenges to Ecological Validity

Conducting studies that have strong ecological validity can be challenging for a number of reasons...

reliability:

the consistency of a set of measurements

in experiments: the extent to which the measurements of a test remain consistent over repeated tests

measured through test-retest correlations, and internal consistency (split-half)

You cannot have validity without having reliability.

Context:

- **person**
- **place**
- **time**
- **setting**
- **other people**

...

Context can be very important for validity. In human behaviour, context can affect behaviour. It is often difficult to generalize from one context to another because peoples' behaviour is so affected by it.

21

placebo effect: sham intervention leads to perceived improvements

learning effects

By doing the experiment, the participants learn something and this modifies their behaviour from what it would have been without the study

Also, Ss may begin an experiment with different experiences and learning.

motivation, novelty

Ss may be more/less motivated during an experiment than they normally would be.

Experiments are also novel experiences, often involving novel materials and procedures, and this can affect behaviour.

Hawthorne Effect ?

Hawthorne effect: subjects improve an aspect of their behavior being experimentally measured simply in response to the fact that they are being studied (perhaps over-estimated)

demand characteristics

demand characteristics: an experimental artifact where participants form an interpretation of the experiment's purpose and unconsciously change their behavior accordingly

the weakness of informing participants that they are taking part in a psychology experiment and yet expecting them to act normally

Clever Hans (in German, der Kluge Hans) was a horse that was claimed to have been able to perform arithmetic and other intellectual tasks.

After formal investigation in 1907, psychologist Oskar Pfungst demonstrated that the horse was not actually performing these mental tasks, but was watching the reaction of his human observers. Pfungst discovered this artifact in the research methodology, wherein the horse was responding directly to involuntary cues in the body language of the human trainer, who had the faculties to solve each problem. The trainer was entirely unaware that he was providing such cues.[1]

http://en.wikipedia.org/wiki/Clever_Hans

Perceived authority can have a profound effect on Ss' behaviour.

The classic example is Milgram's experiments where Ss were made to believe they were shocking another person in the name of a learning experiment. Ss were surprisingly willing to shock another person.

This experiment was recently successfully replicated.

task focus

Ss often exhibit extreme task focus. They may take the tasks given to them during an experiment very seriously.

This is a problem if we expect them to notice things outside the task, such as security messages.

materials effects

The materials used during experiments can change behaviour. Researchers often used meaningless or neutral materials, but these are not representative of what Ss see in their daily lives.

nature of the task, behaviour, or response

The nature of the task, behaviour or response are also important. Experiments often involve unnatural tasks or unusual behaviours that are not normally done in everyday life.

	Group 1 Role playing		Group 2 Sec. primed		Group 3 Pers. accnt.		Groups 1 ∪ 2		Total	
Sent password	18	100%	17	100%	23	92%	35	100%	58	97%
Didn't login	0	0%	0	0%	2	8%	0	0%	2	3%
Total	18		17		25		35		60	

Table 3. Participant responses to the removal of site-authentication images.

DV is percent of people who entered banking information into a false web site when a security indicator is missing.

IVs are type of bank information: real, role-played, or security primed.

Findings were that the security indicator is not effective in preventing people from falling for phishing.

Case: SiteKey

Purpose: test effectiveness of SiteKey authentication mechanism

Subjects: N=67; all customers of a single bank; age=18-25; 91% univ. students; up to 8 people refused to participate because of privacy concerns

Materials:
- **5 online banking tasks: e.g., lookup account balance**
- **real or role-playing credentials**

Context: university research lab.

Procedure:
- **different attack clues: remove https; remove site images; present warning page**

Dependent Variables:
- **% of people entering banking credentials**

32

See

http://www.andrewpatrick.ca/essays/commentary-on-research-on-new-security-indicators

for a critique of this study.

Schechter, S. E., Dhamija, R., Ozment, A., and Fischer, I. 2007. The Emperor's New Security Indicators. In Proceedings of the 2007 IEEE Symposium on Security and Privacy (May 20 - 23, 2007). SP. IEEE Computer Society, Washington, DC, 51-65. DOI= http://dx.doi.org/10.1109/SP.2007.35

Constraints:

- **ethical**
- **methodological**
- **technical**

...

33

Constaints on Obtaining Ecological Validity

ethical

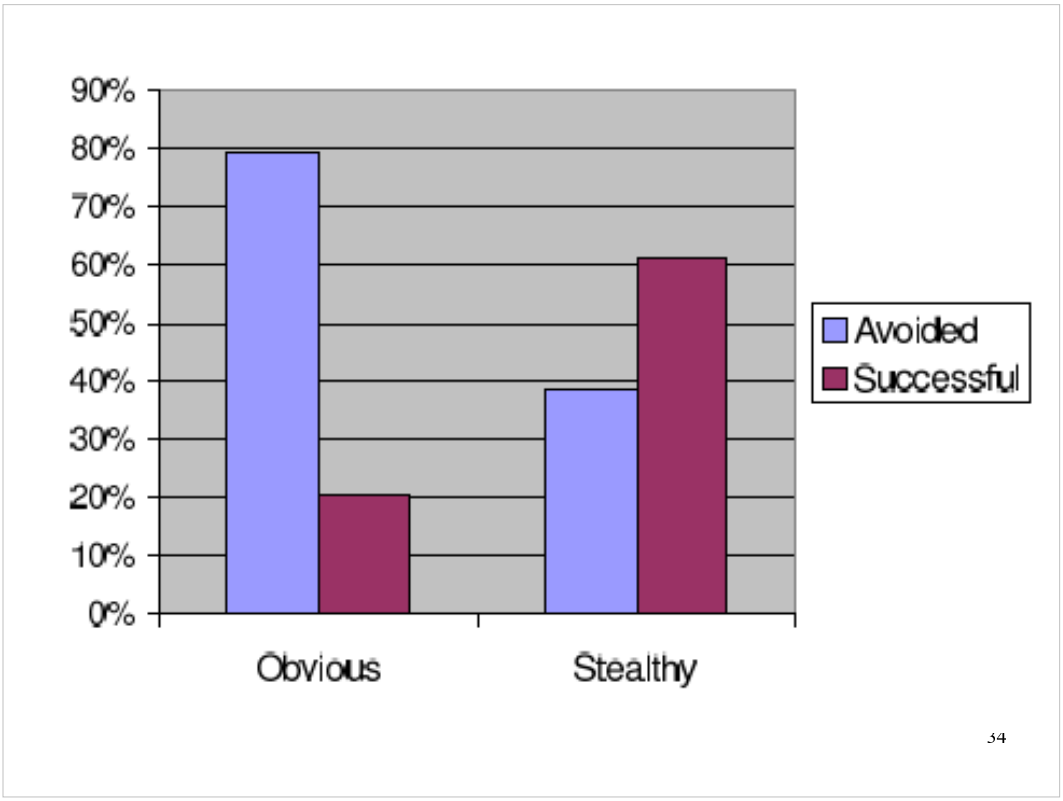
* IRBs, ERBs
* waiving informed consent and using deception
* online research new to most ERBs
* debriefing after deception causing more harm than good (online phishing)

methodological

* phishing: surveys, closed-lab experiments, real-life experiments (Finn & Jakobbson)

technical

* robustness of prototypes



DV is % of people who fell for man-in-the-middle attacks on an out-of-band transaction authentication system using SMS messages

IV is the type of attack

Findings were that authentication mechanism may help with obvious attacks, but not with stealthy ones

Case: Transaction Authorization

Purpose: test out-of-band transaction authorization using SMS

Subjects: N=92; age=?; 89% univ. students and staff

Materials:
- **simulated bank; email to simulate SMS messages**

Context: "online banking security experiment"

Procedure:
- **10 "virtual" transactions**
- **some transactions had bank account number changed by 1 or 5 digits (out of 8)**

Dependent Variables:
- **% of people detecting altered transactions**

35

* unrepresentative sample
* simulated bank, transactions, SMS messages
* no real risks
* manipulation of transactions may not be representatvie of real attacks

Mohammed AlZomai, Bander AlFayyadh, Audun Jøsang & Adrian McCullagh. An Experimental Investigation of the Usability of Transaction Authorization in Online Bank Security Systems. Proc. 6th Australasian Information Security Conference (AISC 2008), Wollongong, Australia, 2008

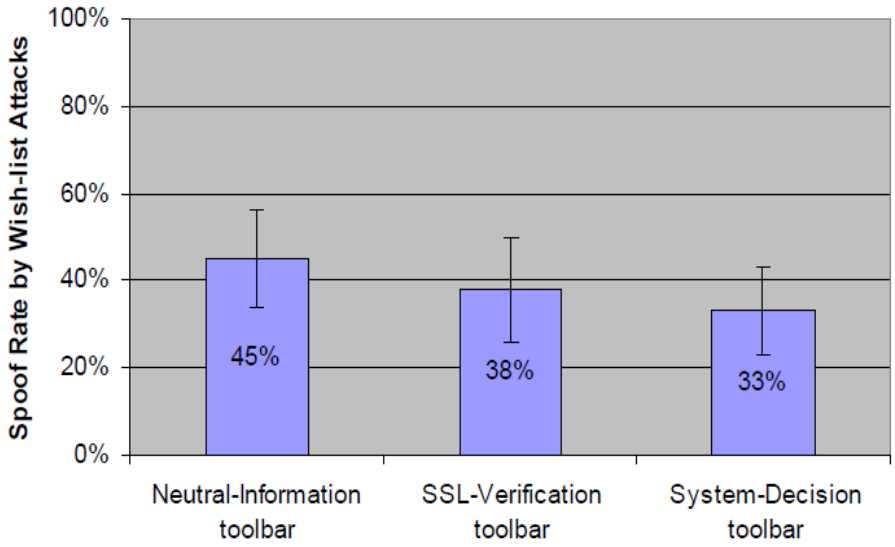


Figure 5. Spoof rates with different toolbars

DV is percent of people who fell for phishing attacks.

IV is the type of toolbar used to present authentication information in the web browser.

Findings were that none of the toolbars are very effective and preventing attacks.

Case: Phishing Toolbars

Purpose: test the effectiveness of anti-phishing toolbars

Subjects: N=30; age=18-50; 66% univ. students

Materials:
• simulated Internet Explorer; dummy accounts at legitimate e-commerce sites; 20 phishing email messages

Context: in-lab; "online banking security experiment", Ss told about fake web sites, encouraged to detect fraud

Procedure:
• 20 email messages
• proceed or "report fraud"

Dependent Variables:
• % of people falling for phishing attacks

* non-representative sample
* artificial test environment
* Ss primed to look for fraud

Wu, M., Miller, R. C., and Garfinkel, S. L. 2006. Do security toolbars actually prevent phishing attacks?. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. CHI '06. ACM, New York, NY, 601-610.

improving ecological validity...

There are some things we can do to improve ecological validity...

real dependent measures

39

It is possible to have Ss do real purchases:

study of buying sex toys: study of the impact of privacy policies on purchasing decisions. Ss paid $45 and then asked to shop for two items worth $15 each (keeping the difference). Items were batteries or the “Pocket Rocket Jr. vibrating sex toy”. Final payment made only after Ss confirmed items had been shipped.

drop-out rates: 2 dropped out, 6 opted out of sex toy, 1 did no purchases. final N=48

recruitment from general population: flyers and Craigslist

J. Tsai, S. Egelman, L. Cranor, and A. Acquisti. The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. Paper presented at the Workshop on the Economics of Information Security, June 7-8, 2007, Pittsburgh, PA.

random, representative sampling

Sampling is important not only for the participants, but also for the materials. Different web sites, email messages, etc. must be tested in order to have generalizability.

experimental controls

* counterbalancing

* reducing nuisance variables

Wizard of Oz Technique

Simulating the behavior of a not-yet-built system so that Ss can have a naturalistic experience.

http://en.wikipedia.org/wiki/Wizard_of_Oz_experiment

replication

repeating experiments, perhaps with different contexts

cover story

camouflaging the purpose of the study. subtle deception

deception

There are degrees of deception.

	Successful	Targeted	Percentage	95% C.I.
Control	15	94	16%	(9–23)%
Social	349	487	72%	(68–76)%

DV is number of people who fell for phishing attack.

IV is whether the phishing messages appeared to come from a friend or a stranger.

Findings were that phishing attacks can be much more effective if they appear to come from a friend.

Case: Social Phishing

Purpose: test susceptibility to phishing attacks, especially when they appear to come from friends

Subjects: N=581; age=18-24; univ. students; users of social network service

Materials:
- **phishing email from friend or stranger**

Context: students in the wild

Procedure:
- **send phishing emails; record and check login credentials**

Dependent Variables:
- **% of people falling for phishing attacks**

47

Jagatic, Johnson, Jakobsson & Menezer, Social phishing. Comm. ACM., 50, 2007.

Researchers at Indiana University were allowed to skip informed consent and use deception.

Friendship information harvested from social network services (e.g., Facebook).

At end of study, all participants told of the experiment. Led to some controversy, mostly from people not realizing how easily friend information could be gathered, and how easy it is to spoof the send of email messages.

Researchers in a second study argued successfully that debriefing would cause more harm than good, because phished people would never know.

ethical principles

vs.

implementation

48

policy statements from a variety of organizations

actual decisions from ethics review boards

informed consent

scientific value

beneficence

confidentiality

49

These are the four main components of ethical research with human Ss.

Article 3.8 The research ethics board (REB) may approve a research proposal and may waive the requirement to obtain informed consent, provided that the REB finds and documents that:

1. The research involves no more than minimal risk to the participants;
2. The waiver is unlikely to adversely affect the well-being and welfare of the participants;
3. The research could not practicably be carried out without the waiver;
4. Whenever possible and appropriate, the participants will be provided with additional pertinent information after participation; and
5. The waived consent does not involve a therapeutic intervention.

50

Draft second edition of Tri-Council Policy Statement

Deception is allowed in research, as long as it is justified. We should be educating our review boards about the necessity for deception.

minimal risk

51

definition of minimal risk

“the probability and magnitude of possible harms implied by participation in the research is no greater than those encountered by the participant in those aspects of his or her everyday life that relate to the research”

Tri-Council Draft Second Edition

research not requiring ethical review

52

Draft second edition of Tri-Council Policy Statement

Article 2.2 Research that relies exclusively on publicly available information does not require research ethics board review. This includes research on living individuals and research on organizations such as governments or corporations, so long as the research is based entirely on material to which the public has access.

Article 2.5 Research involving observation of people in public places that does not allow for the identification of the individuals in research material and that is not staged by the researchers does not require research ethics board review.

Reading list:

Vinson, N.G., Singer, J.A., & Patrick, A.S. (2009). Ethical and privacy issues in HCI research. Manuscript under review. Available from Andrew Patrick.

Patrick, A.S. (2007). Commentary on research on new security indicators. Available at http://andrewpatrick.ca

Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. 2007. Social phishing. Commun. ACM 50, 10 (Oct. 2007), 94-100.

Jakobsson, M. and Ratkiewicz, J. 2006. Designing ethical phishing experiments: a study of (ROT13) rOnl query features. WWW '06. ACM, New York, NY, 513-522.

Jakobsson, M., Johnson, N., and Finn, P. 2008. Why and how to perform fraud experiments. IEEE Security and Privacy 6, 2 (Mar. 2008), 66-68.