Incredible: Is (Almost) All Web Content Trustworthy? Analysis of Psychological Factors Related to Website Credibility Evaluation*

Maria Rafalak
Polish-Japanese Institute of
Information Technology
(PJIIT), Warsaw, Poland
maria.rafalak@pjwstk.edu.pl

Katarzyna Abramczuk
University of Warsaw
(UW)Warsaw, Poland
k.abramczuk@uw.edu.pl

Adam Wierzbicki
Polish-Japanese Institute of
Information Technology
(PJIIT), Warsaw, Poland
adam.wierzbicki@pjwstk.edu.pl

ABSTRACT

This paper describes the results of a study conducted in February 2013 on Amazon Mechanical Turk aimed at identifying various determinants of credibility evaluations. 2046 adult participants evaluated credibility of websites with diversified trustworthiness reference index. We concentrated on psychological factors that lead to the characteristic positive bias observed in many working social feedback systems on the Internet. We have used International Personality Item Pool (IPIP) and measured the following traits: trust, conformity, risk taking, need for cognition and intellect. Results suggest that trustworthiness and risk taking are factors clearly differentiating people with respect to tendency to overestimate, underestimate and judge accordingly websites' credibility. Intuitively people characterized by high general trust tend to be more generous in their credibility evaluations. On the other hand, people who are more willing to take risk, tend to be more critical of the Internet content. The latter indicates that high credibility evaluations are being treated as a default option, and lower ratings require special conditions. Other, more detailed psychological patterns related to websites' credibility evaluations are described in full paper.

Categories and Subject Descriptors

H.5.4 [**Information interfaces and presentation**]: Hypertext/Hypermedia – user issues; J.4 [Social and Behavioral Sciences]: Psychology

General Terms

Performance, Reliability, Experimentation, Human Factors

Keywords

credibility, online behavior, trust, risk-taking, bias

1. INTRODUCTION

Flanangin and Metzger in their article [1] proved that people show a strong tendency to turn mostly to Internet while searching for information. Using more traditional sources like books, newspapers, interpersonal contacts or even television were found to be used much less frequently. But even more important conclusion emerging from this study was that people perceive searching for information as the most important functionality of the Internet.

Searching for information on the Internet is quick, easy, and free. But bearing in mind that the very design of this tool was based on the idea that there should be no or very little control over information flow, one should remember that trustworthiness of many content shared via Internet might be dubious. What is more important, other research show that people usually do not seek for verification of information found on the Internet in other sources [1,2,3,4].

Although there are systems measuring reputation of websites they are not very popular among Internet users. For over 2 billion netizens around the globe only 96 millions use plug-ins like WOT that help judge information credibility[1]. But whether people take reputation of a website into account while evaluating information found on the internet is still a gray area.

Free-flow of unauthorized information combined with blind trust approach can turn out to have dangerous consequences. As stated by Matthew [5], costs of inaccurate judgments based on not credible information may be very high for both business, society and personal life.

1.1 Positive bias

A tendency to overestimate credibility of information is described in literature as *positive bias* or *truth bias* [6,7]. Some researchers claim that such phenomenon in relation to information processing in general (not specific to Internet) might be a result of people's tendency to avoid cognitive effort involved in verifying incoming information [8]. But the exact mechanism of automatic labeling incoming messages as trustworthy instead of untrustworthy remains unclear. It is hypothesized that previous experience might be crucial as the opposite effect named *lie bias* or *Othello error* is mainly observed in people who are often confronted with not credible sources of information like law enforcement personnel [9]. But such tendency is not very common.

Positive bias described above can be observed also in the virtual environment. For example Vassilis Kostakos [10] analyzed functioning of rating mechanisms on different online services containing users' review systems, namely Amazon (online retailer), BooksCrossing (online book community) and Imdb (online movie database).

In his study he proved that the distribution of users' voting behavior used in those systems is left-skewed and depends on users' reputation.

There are also many other studies proving that consumer review systems promote positive bias in evaluations - Weijia et al. [11] studied restaurant review online system, Salganik et al. [12] studied music review systems and these are just few examples representing a wide branch of research. But social influence on people's judgments is a well established psychological phenomenon, so it should not be surprising that such factors influence evaluations also in online environment.

In the previous study conducted on Mechanical Turk platform by the research team of Polish Japanese Institute of Technology (PJIIT) [13] concerning evaluations of Web content credibility we showed strong acquiescence bias, which was present despite eliminating described above social influence factor. In our previous study over 50% of presented websites obtained positive credibility ratings despite balanced selection of credible and not credible websites for the experiment.

In the present study we aimed to determine psychological traits connected with overinflated enthusiasm in websites' credibility estimations. We hoped to find such individual's characteristics that would help differentiate people with specific misjudgment tendencies. This would be a big step towards discovering mechanism(s) preventing people from making wrong decisions based on illusionary convictions. Discovering mechanism(s) responsible for misjudgments in websites' credibility evaluation, both overestimations and underestimations, would also enable prediction of people's online behavior and give clues how to design algorithms that would moderate netizens' ratings in systems measuring websites' credibility.

1.2 Credibility

Studies dedicated to credibility evaluations are very challenging. Firstly, due to subjective nature of examined construct. Secondly, because of wide spectrum of factors potentially influencing final credibility judgment.

The very construct of credibility is most often defined as personal belief that particular piece of information is true and trustworthy [5]. In this paper we decided to acknowledge the abovementioned definition accepting that participants might use different criterion to assess what they consider trustworthy. It is worth noticing that nowhere in this paper we refer to objective construct of truth.

Technical research concerning credibility of online sources so far, focused on creating computer systems that would aid decisions on final credibility judgments [1]. On the other hand, psychological research focused mainly on analyzing cognitive processes connected with credibility evaluation [2,4] including determining heuristics for information evaluations used by netizens [6,7].

We decided to check whether differences in intensity of particular psychological traits influence a tendency to misjudge credibility of websites.

Basing on literature review and scientific intuition we chose: trust, conformity, risk taking, need for cognition and intellect to test their connection with credibility evaluations.

2. PSYCHOLOGICAL TRAITS
2.1 Trust

The most intuitive psychological trait that may influence credibility evaluations is trust.

In the literature there is a distinction between *generalized trust* and *specific trust*. The former is defined as 'an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon'[14]. The latter is connected with interpersonal contacts and can generally be understood as expectations that other people are going to cooperate [15]. In our study we decided to measure general trust level. We hypothesize that *people having higher general trust would give more generous evaluations of websites' credibility comparing to people with lower trust level* (**H1**).

2.2 Conformity

Conformity understood as a dimension of cooperativeness was hypothesized to be conductive to overestimating websites' credibility. *People characterized by high conformity level should avoid conflicts and therefore rarely give negative evaluations* (**H2**).

2.3 Risk taking

Risk taking defined as a tendency to engage in behavior that might be dangerous or socially unacceptable in our opinion is likely to facilitate negative credibility evaluations. If we assume that negative evaluations or criticism is likely to elicit potential conflicts, choosing 'not credible' option may be socially dangerous. Therefore we hypothesize that *people characterized by high level of risk taking will show a tendency to give more negative credibility evaluations comparing to people with low risk taking level* (**H3**).

2.4 Need for cognition

Need for cognition can be defined as a propensity to derive pleasure from demanding cognitive activities.

In accordance with papers claiming that positive bias is a result of *cognitive laziness* [8,9], we hypothesized that *people with higher level of need for cognition trait would be less prone to positive bias as they enjoy cognitive effort which may be connected with verifying information* (**H4**).

2.5 Intellect

Intellect which we define as an estimation of one's own intellectual abilities in our opinion can bolster negative credibility estimations. As this construct might be moderated by self-esteem we hypothesize that *people with higher intellect level will treat themselves as experts in the field and underestimate credibility of evaluated websites* (**H5**).

3. HYPOTHESES

To sum up, we hypothesize that users classified as *over-raters* should present higher levels of *trust* (**H1**) and *conformity* (**H2**) when compared to *under-raters*.

People classified as *under-raters* should present higher levels of *risk taking* (**H3**), *need for cognition* (**H4**) and *intellect* (**H5**) when compared to *over-raters*.

We do not formulate any hypotheses concerning intensity of psychological traits in people giving adequate credibility judgments (*average-raters*) and treat this study as an exploratory research in this matter.

4. EXPERIMENT DESIGN

The experiment was conducted between February and May 2013 via Amazon Mechanical Turk Platform (www.mturk.com).

Participants were recruited from registered users with at least 75% approved assignments and originating from countries except Pakistan, India, Thailand, Russia and China.

Our sample consisted of 803 male and 1243 female participants of mean age 30.

Each participant was rewarded with 0.35$ after first "hit" (completing the task). Every person could participate in more than one round of the experiment, increasing reward up to 0.55$ per hit.

The main task of participants was to evaluate credibility of presented websites on scale from 1 to 5, where higher evaluations represented greater credibility. To do this participants chose one of the answer options:

Rate credibility of this website:

a.　　*completely not credible*

b.　　*mostly not credible*

c.　　 *somewhat credible, although with major doubt*

d.　　*credible, with some doubt*

e.　　 *completely credible*

f.　　*do not know*

Every participant declared how many websites he/she wanted to evaluate. Websites were selected for the experiment according to their domain WOT index (http://www.mywot.com/) which was treated as an external credibility reference. WOT index ranges from 0 to 100 where greater values represent higher credibility. Proportion of low credible, highly credible and moderately credible websites was balanced in website sample. WOT credibility index is based on combination of *wisdom of crowds* approach (system collects votes from vast population of Internet users) and external data-bases search but the details of algorithm calculating final credibility index is hidden from the public. Despite that, WOT as a free and user-friendly software is one of the most popular site-rating tool. This is why we decided to rely on it in our study.

Websites were assigned randomly to participants.

In order to avoid additional uncontrolled manipulation of experiment settings, participants saw neither credibility evaluations given by other users nor website WOT credibility index.

 Additionally, in order to ensure that participants paid sufficient attention to evaluations (in other words to eliminate obtaining careless and random evaluations) the experimental task included some additional tasks like answering one additional question connected with evaluated website's content.

Moreover, as stated by Flanagin and Metzger [1], the way people perceive credibility of information is dependent on whether person attempts to verify it. This is why we asked our participants to supply URL links of websites they used for assessing credibility.

In order to eliminate the effect of particular topics instigating low or high credibility ratings, topics of websites used in our experiment were diversified and formed five broad categories: medicine ($N = 761$), personal finance ($N = 788$), healthy lifestyle ($N = 713$), entertainment ($N = 747$), politics and economy ($N = 148$).

Intensity of psychological traits (trust, conformity, risk-taking, need for cognition and intellect) among the participants were assessed using International Personality Item Pool (http://ipip.ori.org). This is an open-source bank of psychological scales that can be used in scientific research. For our experiment we chose scales describing selected psychological traits that showed greatest reliability Cronbach's alpha coefficients (exact values listed in brackets):

- NEO: A1 for measuring trust (0.82),
- JPI: Cpr for measuring conformity (0.71),
- JPI: Rkt for measuring risk taking (0.78),
- AB5C: V+/V+ vs V-/V- for measuring intellect (0.81),
- CHS for measuring need for cognition (0.84).

In psychometrics reliability coefficient is a metrics reflecting how precise is the applied measure. It takes values from 0 to 1 where the bigger value the lower measurement error.

As presented above all scales used in the experiment have satisfactory reliability so it may be assumed that results obtained in this way are trustworthy.

5. ANALYSIS

In our experiment 3157 different websites were evaluated. The number of participants who took part in the experiment exceeded two thousand ($N = 2046$).

Intensity of psychological traits measured with IPIP scales was estimated using *graded response model* from Item Response Theory (IRT). As in all IRT models, results are expressed in standardized scale ($M = 0$, $SD = 1$).

Descriptive statistics of person's trait level estimations are summarized in Table 1.

In order to check whether people with a tendency to underestimate and overestimate credibility of websites differ in intensity of psychological traits we adopted the following procedure.

Firstly we standardized both WOT ratings for websites used in the experiment and evaluations of credibility given by participants. Secondly, for every pair website-person we calculated the difference between WOT index and credibility ratings. For every person we calculated median from the abovementioned difference. All "do not know" answers were treated as missing data and excluded from analysis.

Finally we classified people with median lower or equal than first quartile of the abovementioned difference distribution (<= -0.5) as *over-raters*, people with median difference higher or equal to third quartile (<= 0.5) as *under-raters*. All other people from the sample were classifies ad *average-raters* .

Next we calculated one way analysis of variance (ANOVA) checking whether those three groups differ in intensity of psychological traits. As all the results turned out to be statistically significant, we computed post-hoc Tukey's test to check how exactly groups differ from one another. Results of this comparisons are illustrated in Table 2 and Figure 1.

5.1 Trust

People classified as *average-raters* obtained higher trust level among all analyzed groups. Additionally *over-raters* obtained higher results in trust scale than *under-raters*. This confirms our first hypothesis (**H1**).

5.2 Conformity

People with tendency to underestimate credibility achieved highest scores in conformity scale. Moderately negative score on conformity scale was achieved by both people who have tendency to overestimate credibility and assess credibility in accordance with WOT ratings. Obtained result does not confirm our hypothesis (**H2**).

5.3 Risk Taking

People showing tendency to overestimate websites' credibility achieved relatively lowest scores in risk taking scale. People showing tendency to underestimate websites' credibility achieved highest results in risk taking scale. This result confirms our research hypothesis (**H3**).

Additionally, people evaluating websites' credibility accordingly to WOT index achieved results similar to people overestimating websites' credibility.

Table 1. Descriptive statistics for estimated psychological traits levels using graded response model

	Min	Q1	Me	Mean	Q3	Max
TRUST	-2.88	-0.60	-0.07	0.05	0.66	2.69
RT	-2.64	-0.29	0.14	0.18	0.71	2.53
NFC	-2.94	-0.45	0.17	0.22	0.89	1.66
INTEL	-3.50	-0.57	0.09	0.05	0.68	1.68
CONF	-2.39	-0.60	0.05	-0.05	0.55	3.50

TRUST – trust; NFC – need for cognition; RT – risk taking; INTEL-intellect; CON – conformity

Table 2. Results of post-hoc Tukey tests conducted for psychological measures to compare under-raters, average-raters and over-raters

	OR–AR	UR–AR	UR–OR
TRUST	-0.223*	-0.549**	-0.326**
NFC	-0.141*	-0.251**	-0.110
RT	-0.398**	0.119*	0.517**
INTEL	- 0.406*	-0.038	0.368**
CON	-0.024	0.518**	0.542*

* p<0.05; ** p<0,001

TRUST – trust; NFC – need for cognition; RT – risk taking; INTEL-intellect; CON – conformity

OR – over-raters; UR – under-raters; AR – average-raters

Mean standardized intensity of psychological traits

TRUST   NFC   INTEL   CONF   RT

0,50
0,40   0,398
0,30   0,285   0,302
0,20   0,144
0,10 0,100 0,101   0,083
0,00   -0,009
-0,10 -0,100 -0,101   -0,144
-0,20
-0,30   -0,285   -0,302
-0,40
-0,50   -0,425   -0,215

OVER-RATERS   AVERAGE-RATERS   UNDER-RATERS
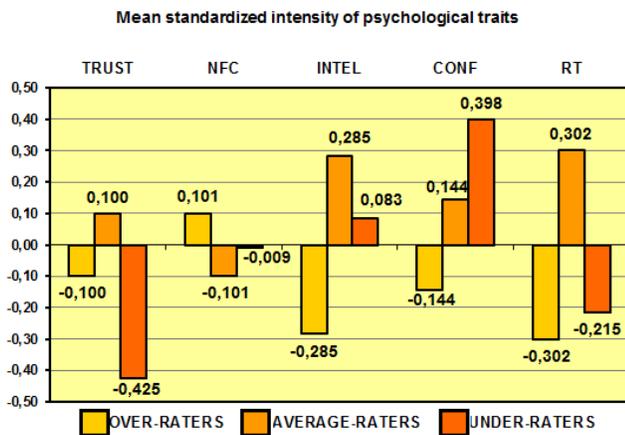
Figure 1. Mean estimates of psychological trait levels for groups

5.4 Need for cognition

People classified as *average-raters* obtained relatively highest results in need for cognition scale. People with a tendency to overestimate websites' credibility obtained moderate result in need for cognition scale. Relatively lowest need for cognition intensity can be assigned to people with a tendency to underestimate websites credibility. Because the difference in need for cognition scale between *over-* and *under-raters* did not reach statistical significance our research hypothesis was not confirmed (**H4**).

5.5 Intellect

The most extreme results in intellect scale were obtained by people who have a tendency to overestimate websites' credibility – they achieved the lowest score among all analyzed groups. People with tendency to underestimate websites' credibility did not differ from those people who give accurate evaluations in terms of intellect level. Obtained result confirms our research hypothesis (**H5**).

5.6 Summary

Hypotheses concerning differences in the intensity of trust (**H1**), risk taking (**H3**) and intellect (**H5**) between users characterized as *under-* and *over- raters* have been confirmed.

Hypotheses concerning differences in the intensity of conformity (**H2**) and need for cognition (**H4**) has not been confirmed. Pattern of obtained results in need for cognition is consistent with our expectations but failed to reach the level of statistical significance.

6. DISCUSSION

Our hypothesis stating that people showing tendency to overestimate websites' credibility will obtain higher results is general trust scale and lower results on scale measuring risk taking than people showing tendency to underestimate websites' credibility has been confirmed.

This result might suggest that people treat evaluating websites as credible to be a *safe option*. Risk in this case may be connected with necessity to protect or reason one's choice when formulating critical judgment.

Additional conclusion stemming from our research might be that moderate positive level of trust may be associated with adequate credibility evaluations, while negative level with erroneous judgments. Negative moderate trust level seems to be connected with credibility overestimation, while negative trust level of greater intensity with credibility underestimation. Such conclusion is plausible but requires further and more detailed research for final confirmation.

Our hypothesis concerning high level of conformity in people having tendency to overestimate websites has not been confirmed. Maybe relatively high results obtained in conformity scale by people classified as *under-raters* can be explained by their tendency to behave accordingly to experimenter's expectations. Those people might think that they should be critical in experiment in which they are asked to evaluate credibility. So conformity connected with willingness to cooperate, in our experimented was redirected towards the experimenter not towards websites' creators. Such explanation is plausible but requires further empirical confirmation.

Hypothesis concerning the level of need for cognition has not been confirmed as in our study this was not differentiating people with credibility misjudgment tendencies. But an interesting result has been observed - people obtaining highest scores in need for cognition were giving accurate credibility judgments. This might suggest that cognitive laziness connected with verifying information may lead both to credibility overestimations or underestimations. The exact mechanism is yet to be determined.

Our last hypothesis concerning level of intellect has not been confirmed. Scales used for measuring intellect based on self-report of ones' cognitive abilities, so obtained results might be heavily distorted by self-esteem or other similar psychological factors. In future it would probably be recommended to use aptitude tests for measuring intellect instead of questionnaire methods.

Interestingly *average-raters* showed highest levels of risk-taking, trust and intellect among all analyzed groups.

Psychological traits that clearly differentiate all three analyzed groups are trust and risk taking.

6.1 Psychological profiles of experiment participants

People showing tendency to underestimate websites' credibility can be characterized by relatively high negative level of trust and low need for cognition but relatively high positive level of conformity and risk taking.

People showing tendency to underestimate websites' credibility can be characterized by moderate but negative level of trust and conformity , moderate positive level of need for cognition, strong negative level of intellect and risk taking.

People giving adequate credibility judgments can be characterized by relatively high positive level of need for

cognition, moderate positive level of trust and risk taking and negative moderate level of conformity.

7. GENERAL CONCLUSIONS AND FUTURE WORK

Level of general trust and risk taking are good measures to include in further research dedicated to evaluating websites' credibility as they clearly differentiate three groups of people: those having tendency to overestimate, underestimate or judge accordingly the credibility of presented websites.

Because of discovered relationship between credibility evaluations and risk taking level we are currently designing further study introducing gamification mechanisms, which as we hypothesize, are likely to moderate positive bias in credibility evaluations.

We also suggest to study further the effect of conformity and intellect as the results obtained in those scales clearly differentiate extreme groups: people having tendency to overestimate and underestimate websites' credibility.

Our final conclusion from this study is that the usage of need for cognition scale in research connected with evaluating websites' credibility is questionable, as our findings show that results obtained in this scale do not differentiate people having tendency to overestimate and underestimate websites' credibility.

8. POTENTIAL APPLICATIONS

There already have been some attempts to built automated systems that could bolster credibility evaluations based on information source technical characteristics like visual design or twitter profile history [16][17]. Results of our study inscribe at least partially in this branch of potential research applications. Credibility assessment is by definition a very subjective matter. Its final estimate is a product of interaction between source/message technical features and receiver's individual characteristics. Our research suggests which psychological traits may need to be controlled while designing such complex algorithmic solutions. Definitely future research is needed to improve our knowledge on the subject as definite patterns of those interactions are yet to be determined.

Polish-Japanese Institute of Information Technology is currently preparing an automated system that is supposed to support credibility evaluation of Web content (www.reconcile.pl). The system, similarly to myWOT is partially based on users' evaluations (wisdom of crowds approach). Therefore we are looking forward to introducing a mechanism that would be able to predict people's misjudgment tendencies and correct their votes in the system accordingly.

Newest scientific literature already presents proof that predicting personality traits from online behavior (namely facebook likes) is possible [18]. Perhaps predicting other psychological traits from online behavior will also be obtainable in the near future and support (or substitute) using psychological questionnaires to predict behavior.

9. REFERENCES

[1] Flanagin A.J., Metzger M., 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly*, *77*,3, 515-540.

[2] Sparks J.R., Rapp D. N.,2011. Readers' Reliance on Source Credibility in the Service of comprehension. *Journal of Experimental Psychology, 37*, 1, 230–247.

[3] Lussen T., 2013. *Trust in Online information*. Doctoral Thesis. Universiteit Twente.

[4] Metzger M.J., 2007. Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research. *Journal of the American Society for Information Science and Technology, 58,* 13, 2078-2091.

[5] Jensen M.J., Lowry P.B., Jenkins J., 2011.Effects of Automated and Participative Decision Support in Computer-Aided Credibility Assessment. *Journal of Management Information Systems*, *28*,1, 201-234.

[6] Bond, C.F., and DePaulo, B.M., 2008. Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*, 4 , 477–492.

[7] Burgoon, J.K., Blair, J.P., Strom, R.E., 2008. Cognitive biases and nonverbal cue availability in detecting deception. Human Communication Research, 34, 4, 572–599.

[8] DePaulo, B., and Kashy, D.,1998. Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology, 74*, 1, 63–79.

[9] Ekman, P. Telling Lies. New York: Norton, 1985.

[10] Kostakos V., 2009. Is the *crowd's* wisdom biased? A quantitative assessment of three online communities. *Computational Science and Engineering, 4*.

[11] Weijia D., Jin G., Lee, J., Luca. M., 2012. Optimal Aggregation of Consumer Ratings: An Application to Yelp.com., *Harvard Business School Working Paper*, 13, 42.

[12] Salganik M.,J., Dodds, P.S., Watts D.J., 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science, 311*, 5762, 854-856.

[13] Kakol M, Jankowski-Lorek M., Abramczuk K.Catasta M. , Wierzbicki A., 2013. On the Subjectivity and Bias of Web Content Credibility Evaluations, *WWW2013 Research Papers*, 1131-1136.

[14] Rotter, J. B,. 1967. A New Scale For the Measurement of Interpersonal Trust, *Journal of Personality, 35,* 651–665.

[15] Pruitt, D. G., Kimmel M. J.,1977. Twenty Years of Experimental Gaming. Critique,Synthesis and Suggestions For the Future, *Annual Review of Psychology, 28*, 363–392.

[16] Gupta, A., Kumaraguru, P. 2012. Credibility ranking of tweets during high impact events. *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*.

[17] Castillo, C., Mendoza, M., Poblete, B., 2011. *Information Credibility on Twitter. Proceedings of WWW2011.*

[18] Kosinski, M., Stillwell, D.J., Graepel, T, 2013. Private traits and attributes are predictable from digital records of human behavior, Proceedings of the National Academy of Sciences (PNAS).