Supplemental information for

Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings

doi: 10.1093/jcmc/zmz012

R. Kelly Garrett[1] and Shannon Poulsen

Ohio State University School of Communication

Table of Contents

[1] Corresponding author: garrett.258@osu.edu

Appendix S1: Study 1 sample

Exclusion criteria

Recruiting from opt-in online panels allows researchers to access a much more diverse pool of potential participants than would otherwise be possible, but it also has important limitations. A non-trivial number of individuals recruited in this way do not make a good faith effort to participate. They ignore directions, answer questions without reading, enter nonsensical responses, interrupt the study to pursue other interests, etc., all of which can introduce error. For these reasons, we worked to identify and exclude such cases prior to undertaking analyses. We do this using a variety of strategies. Some are obvious: we exclude participants who choose the same response for every item in a scale that includes reverse-coded items. For example, a participant who describes a source as "extremely trustworthy" and "unbiased" but also as "sensational" and "not at all sincere" is very unlikely to be reading questions carefully. Others are more nuanced: although participants are likely to choose neutral (midpoint) responses to some items, if they do so across dozens of questions tapping three or more different concepts, it is unlikely that they are attending carefully to the study. We also exclude participants who spent more than 2 hours on either wave of the study because disruptions of the task make it more difficult for participants to follow relevant instructions. The second wave of the study also included three open-ended questions. We exclude individuals who skipped or provided nonsense answers to these items (e.g., "cool beans dude").

Demographics

Participants were between 19 and 85 years old (*M* = 49.07, *SD* = 16.62). The modal level of educational attainment was "some college but no degree" (33.5%), followed closely by those holding an associates or bachelor's degree (31.7%), and those having only completed a high school degree or less (24.3%). Participants were fairly evenly divided by political ideology (34.9% liberal and 33.9% conservative), and party (43.6% Democrats and 29.9% Republican). The sample was disproportionately White (88.5%), and Blacks were underrepresented (7.3%). The largest bias in the sample was associated with gender: almost three-quarters of study participants (73.4%) were women.

Appendix S2: Study 1 instructions

Fact-checker flagging (*n* = 50)

Facebook is looking for ways to fight the spread of misleading information on its service. The company has developed a new feature intended to help users recognize questionable information that appears on their newsfeeds so that they can make well informed decisions about the information they read and share.

In this study, we are asking for your help testing this new feature.

With the new feature, Facebook users can "flag" articles that contain misleading or inaccurate information. If several users flag the post, it will be sent to a pool of 3rd party fact-checking organizations. Fact checkers will review the claims made in the article, and if at least two organizations conclude the article contains misleading or inaccurate information, the post will be flagged. When a story that has been flagged by fact checkers shows up on your newsfeed, you'll see a warning attached to the post.

Peer flagging (*n* = 56)

Facebook is looking for ways to fight the spread of misleading information on its service. The company has developed a new feature intended to help users recognize questionable information that appears on their newsfeeds so that they can make well informed decisions about the information they read and share.

In this study, we are asking for your help testing this new feature.

With the new feature, Facebook users can "flag" articles that contain misleading or inaccurate information. When a story that has been flagged by other Facebook users like you shows up on your newsfeed, you'll see a warning attached to the post.

Self-identified humor (*n* = 53)

Facebook is looking for ways to fight the spread of misleading information on its service. The company has developed a new feature intended to help users recognize questionable information that appears on their newsfeeds so that they can make well informed decisions about the information they read and share.

In this study, we are asking for your help testing this new feature.

With the new feature, Facebook has created a list of websites that describe themselves as providing potentially deceptive information, including satire, parody, hoaxes, etc. When a story hosted on one of these websites shows up on your newsfeed, you'll see a warning attached to the post.

Control (*n* = 59)

Facebook is looking for ways to fight the spread of misleading information on its service. The company is developing new features intended to help users recognize questionable information that appears on their newsfeeds so that they can make well informed decisions about the information they read and share.
In this study, we are asking for your help evaluating current Facebook posts.

Appendix S3: Study 2 sample

Exclusion criteria

Observing the flag was a critical part of this study. The study included instructions describing the importance of the flag, and flagged messages were displayed multiple times. Participants who, despite this, did not recall ever seeing the flag were excluded (*n* = 99). We also excluded individuals who selected mutually exclusive responses on two separate scales (*n* = 16). The questionnaire included one open-ended item, and we excluded nonsense responses to this item. (E.g., "He was my first time for you to be a great time for you to be a great time for you to be a great time.", *n* = 50). One participant skipped every question except an attention check.

To help ensure that the effects observed here are not the product of post-treatment bias (Montgomery, Nyhan, & Torres, 2018) we reran the analyses without exclusions based on flag recall, straightlining, or nonsense answers. The coefficients magnitude and direction were the same in every analysis (see Tables S3a and S4a).

References

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science, 62*(3), 760-775. doi:10.1111/ajps.12357

Demographics

Participants were between 19 and 90 years old (*M* = 46.28, *SD* = 16.28). The modal level of educational attainment was holding an associates or bachelor's degree (38.5%), followed by those having only completed a high school degree or less (25.0%), and "some college but no degree" (22.8%). Participants were fairly evenly divided by political ideology (32.3% liberal and 36.7% conservative) and party (41.0% Democrats and 32.2% Republican). In terms of race, the sample was predominantly White (77.5%), but Blacks (12.1%) and Asians (6.7%) were also included. The sample included comparable numbers of men and women (52.9% female).

Appendix S4: Study 2 instructions

Facebook is looking for ways to fight the spread of misleading information on its service. The company has developed a new feature intended to help users recognize questionable information that appears on their newsfeeds so that they can make well informed decisions about the information they read and share.

In this study, we are asking for your help testing this new feature.

With the new feature, Facebook can “flag” articles that contain misleading or inaccurate information. When a story that has been flagged shows up on your newsfeed, you’ll see a warning attached to the post. There will be a red warning symbol followed by a brief message explaining why the message has been flagged. The warning symbol looks like this:

⚠

Table S1. Study 1 random-effects models estimating flag type influence on message perceptions

	Acceptance of falsehood		Sharing intention		Source credibility	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
Peer-generated flag [a]	-.520, .508	-.431, .726	-.579, .659	-.711, .597	-.427, .466	-.618, .382
Fact-checker flag [a]	-.749, .309	-.623, .577	-.374, .902	-.247, 1.100	-.480, .440	-.385, .648
Self-identified humor flag [a]	**-1.185, -.142**	-.861, .341	**-1.263, -.006**	-1.115, .223	-.889, .018	-.826, .207
Inaccurate issue beliefs, t-1	—	**1.505, 2.649**	—	**.296, 1.058**	—	**.474, 1.356**
Peer X inaccurate	—	-1.300, .340	—	-.348, .751	—	-.346, .921
Fact-checker X inaccurate	—	-1.513, .144	—	-.989, .095	—	-1.082, .186
Self-ID X inaccurate	—	**-2.040, -.392**	—	-1.067, .034	—	-1.061, .211
Constant	**3.447, 4.164**	**2.595, 3.396**	**2.755, 3.620**	**2.469, 3.378**	**2.926, 3.549**	**2.534, 3.227**
Variance components						
Random intercept	.497, 1.392	.543, 1.283	2.090, 3.181	2.099, 3.179	.736, 1.336	.719, 1.272
Residual	1.894, 2.758	1.424, 2.075	.480, .699	.421, .613	.830, 1.209	.714, 1.040
Number of observations	436	436	436	436	436	436
Number of participants	218	218	218	218	218	218
Likelihood ratio test	$\bar{\chi}^2(1) = 16.10$ *p* <.001	$\bar{\chi}^2(1) = 24.31$ *p* <.001	$\bar{\chi}^2(1) = 239.49$ *p* <.001	$\bar{\chi}^2(1) = 257.25$ *p* <.001	$\bar{\chi}^2(1) = 61.99$ *p* <.001	$\bar{\chi}^2(1) = 70.24$ *p* <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category is the no flag (control) condition.

Table S1a. Study 1 message perception models with continuous belief accuracy measures

	Acceptance of falsehood		Sharing intention		Source credibility	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
Peer-generated flag [a]	-.520, .508	-1.050, .671	-.579, .659	-1.280, .303	-.427, .466	-.1.174, .260
Fact-checker flag [a]	-.749, .309	-.285, 1.356	-.374, .902	-.191, 1.370	-.480, .440	-.261, 1.115
Self-identified humor flag [a]	**-1.185, -.142**	-.725, 1.039	**-1.263, -.006**	-1.215, .404	-.889, .018	-.942, .530
Issue belief inaccuracy, t-1 [b]	—	**.423, .668**	—	**.056, .225**	—	**.155, .350**
Peer X inaccuracy [b]	—	-.168, .199	—	-.004, .246	—	-.043, .248
Fact-checker X inaccuracy [b]	—	**-.037, -.027**	—	-.202, .032	—	-.255, .020
Self-ID X inaccuracy [b]	—	**-.425, -.056**	—	-.193, .061	—	-.223, .071
Constant	**3.447, 4.164**	**1.115, 2.251**	**2.755, 3.620**	**2.101, 3.181**	**2.926, 3.549**	**1.778, 2.734**
Variance components						
Random intercept	.497, 1.392	.478, 1.126	2.090, 3.181	2.082, 3.150	.736, 1.336	.661, 1.179
Residual	1.894, 2.758	1.232, 1.798	.480, .699	.404, .589	.830, 1.209	.674,.984
Number of observations	436	436	436	436	436	436
Number of participants	218	218	218	218	218	218
Likelihood ratio test	$\bar{\chi}^2(1) = 16.10$	$\bar{\chi}^2(1) = 24.47$	$\bar{\chi}^2(1) = 239.49$	$\bar{\chi}^2(1) = 260.18$	$\bar{\chi}^2(1) = 61.99$	$\bar{\chi}^2(1) = 67.23$
	p <.001	*p* <.001	p <.001	*p* <.001	p <.001	*p* <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category is no flag (control) condition. b. Continuous measure of pre-test belief accuracy

Table S2. Study 1 random-effects models estimating flag influence on flagging system perceptions

	Reactance		Value of flagging	
	Main effect	Interaction	Main effect	Interaction
Peer-generated flag [a]	-.117, 1.091	-.493, .797	-.904, .163	-.635, .537
Fact-checker flag [a]	-.144, 1.098	-.367, .959	-.539, .559	-.336, .868
Inaccurate issue beliefs, t-1	—	-.507, .249	—	-.165, .619
Peer X inaccurate	—	**.263, 1.330**	—	**-1.302, -.196**
Fact-checker X inaccurate	—	-.123, .930	—	**-1.112, -.018**
Constant	**2.526, 3.392**	**2.554, 3.486**	**4.702, 5.468**	**4.554, 5.402**
Variance components				
Random intercept	1.831, 2.984	1.822, 2.963	1.349, 2.255	1.357, 2.258
Residual	.399, .620	.370, .574	.444, .689	.416, .646
Number of observations	318	318	318	318
Number of participants	159	159	159	159
Likelihood ratio test	$\chi^2(1) = 181.07$	$\chi^2(1) = 187.47$	$\chi^2(1) = 136.66$	$\chi^2(1) = 143.44$
	p <.001	p <.001	p <.001	p <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category is self-identified humor flag.

Table S2a. Study 1 flagging-system perceptions models with continuous belief accuracy measures

	Reactance		Value of flagging	
	Main effect	Interaction	Main effect	Interaction
Peer-generated flag [a]	-.117, 1.091	-.979, .628	-.904, .163	-.596, .954
Fact-checker flag [a]	-.144, 1.098	-.572, 1.014	-.539, .559	-.457, 1.062
Issue belief inaccuracy, t-1 [b]	—	-.092, .091	—	-.078, .114
Peer X inaccuracy [b]	—	**.034, .289**	—	-.267, .001
Fact-checker X inaccuracy [b]	—	-.056, .185	—	-.199, .054
Constant	**2.526, 3.392**	**2.380, 3.540**	**4.702, 5.468**	**4.447, 5.567**
Variance components				
Random intercept	1.831, 2.984	1.817, 2.956	1.349, 2.255	1.367, 2.275
Residual	.399, .620	.370, .574	.444, .689	.422, .655
Number of observations	318	318	318	318
Number of participants	159	159	159	159
Likelihood ratio test	$\bar{\chi}^2(1) = 181.07$ p <.001	$\bar{\chi}^2(1) = 185.67$ p <.001	$\bar{\chi}^2(1) = 136.66$ p <.001	$\bar{\chi}^2(1) = 142.69$ p <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs not do not contain zero are in **bold**. a. Reference category is self-identified humor flag. b. Continuous measure of pre-test belief accuracy

Table S3. Study 2 random-effects models estimating flag type influence on message perceptions

	Acceptance of falsehood		Sharing intention		Source credibility	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
Story, self-identified (StorySID)[a]	**-.806, -.125**	**-.878, -.152**	-.713, .121	-.756, .110	**-.700, -.079**	**-.692, -.033**
Story, Facebook (StoryFB)[a]	-.574, .126	-.608, .139	-.678, .179	-.757, .249	-.302, .336	-.398, .286
Site, self-identified (SiteSID) [a]	-.582, .111	-.563, .170	-.636, .211	-.705, .177	-.506, .125	-.558, .114
Site, Facebook (SiteFB) [a]	-.616, .085	-.570, .164	-.507, .350	-.601, .287	-.472, .167	-.511, .165
Inaccurate issue beliefs, t-1	—	**.802, 1.469**	—	**.025, .556**	—	**.109, .574**
StorySID X inaccurate	—	-.085, .895	—	-.241, .542	—	-.341, .343
StoryFB X inaccurate	—	-.383, .611	—	-.250, .549	—	-.133, .565
SiteSID X inaccurate	—	-.471, .493	—	-.215, .545	—	-.215, .451
SiteFB X inaccurate	—	-.446, .536	—	-.103, .669	—	-.212, .466
Constant	**3.555, 4.033**	**3.064, 3.577**	**2.713, 3.297**	**2.577, 3.190**	**3.029, 3.465**	**2.871, 3.339**
Variance components						
Random intercept	1.093, 1.560	.858, 1.244	2.354, 3.018	2.245, 2.885	1.231, 1.602	1.180, 1.536
Residual	1.175, 1.471	1.005, 1.260	.487, .609	.470, .590	.403, .505	.382, .479
Number of observations	1,220	1,220	1,220	1,220	1,220	1,220
Number of participants	610	610	610	610	610	610
Likelihood ratio test	$\chi^2(1) = 174.10$ p <.001	$\chi^2(1) = 154.53$ p <.001	$\chi^2(1) = 713.49$ p <.001	$\chi^2(1) = 678.74$ p <.001	$\chi^2(1) = 518.69$ p <.001	$\chi^2(1) = 516.04$ p <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category is no flag (control) condition.

Table S3a. Study 2 message perceptions models, without exclusions

	Acceptance of falsehood		Sharing intention		Source credibility	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
Story, self-identified (StorySID)[a]	**-.665, -.008**	-.682, .009	-.637, .175	-.687, .154	**-.666, -.049**	**-.652, -.004**
Story, Facebook (StoryFB)[a]	-.452, .200	-.454, .243	-.557, .251	-.618, .223	-.250, .364	-.352, .298
Site, self-identified (SiteSID) [a]	-.539, .110	-.517, .174	-.622, .181	-.682, .152	-.469, .141	-.507, .137
Site, Facebook (SiteFB) [a]	-.551, .102	-.553, .136	-.525, .282	-.615, .220	-.419, .194	-.442, .203
Inaccurate issue beliefs, t-1	—	**.850, 1.486**	—	**.047, .537**	—	**.158, .594**
StorySID X inaccurate	—	-.161, .749	—	-.178, .527	—	-.303, .324
StoryFB X inaccurate	—	-.382, .518	—	-.211, .491	—	-.063, .561
SiteSID X inaccurate	—	-.443, .435	—	-.203, .472	—	-.215, .387
SiteFB X inaccurate	—	-.280, .607	—	-.086, .593	—	-.217, .388
Constant	**3.641, 4.110**	**3.107, 3.612**	**2.903, 3.483**	**2.761, 3.367**	**3.180, 3.465621**	**3.000, 3.469**
Variance components						
Random intercept	1.248, 1.688	.960, 1.320	2.662, 3.307	2.761, 3.367	1.467, 1.841	1.372, 1.725
Residual	1.181, 1.441	1.017, 1.243	.461, .562	.447, .546	.392, .479	.373, .456
Number of observations	1,550	1,550	1,550	1,550	1,550	1,550
Number of participants	775	775	775	775	775	775
Likelihood ratio test	$\bar{\chi}^2(1) = 251.74$ p <.001	$\bar{\chi}^2(1) = 215.57$ p <.001	$\bar{\chi}^2(1) = 1010.47$ p <.001	$\bar{\chi}^2(1) = 949.06$ p <.001	$\bar{\chi}^2(1) = 762.78$ p <.001	$\bar{\chi}^2(1) = 730.57$ p <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category is no flag (control) condition.

Table S4. Study 2 random-effects models estimating flag influence on flagging system perceptions

	Reactance		Value of flagging	
	Main effect	Interaction	Main effect	Interaction
Story, Facebook (StoryFB) [a]	-.139, .602	-.176, .630	-.539, .157	-.682, .059
Site, self-identified (SiteSID) [a]	-.135, .598	-.114, .676	-.622, .066	-.725, .005
Site, Facebook (SiteFB) [a]	-.170, .570	-.158, .637	-.644, .053	**-.760, -.026**
Inaccurate issue beliefs, t-1	—	-.112, .505	—	**-.566, -.070**
StoryFB X inaccurate	—	-.458, .430	—	-.004, .710
SiteSID X inaccurate	—	-.578, .273	—	-.088, .593
SiteFB X inaccurate	—	-.554, .312	—	-.055, .638
Constant	**3.097, 3.606**	**3.011, 3.560**	**5.016, 5.495**	**5.109, 5.615**
Variance components				
Random intercept	1.577, 2.133	1.564, 2.118	1.484, 1.974	1.481, 1.968
Residual	.579, .747	.579, .747	.353, .455	.349, .450
Number of observations	955	955	956	956
Number of participants	478	478	478	478
Likelihood ratio test	$\bar{\chi}^2(1) = 372.76$ p <.001	$\bar{\chi}^2(1) = 368.14$ p <.001	$\bar{\chi}^2(1) = 511.10$ p <.001	$\bar{\chi}^2(1) = 512.64$ p <.001

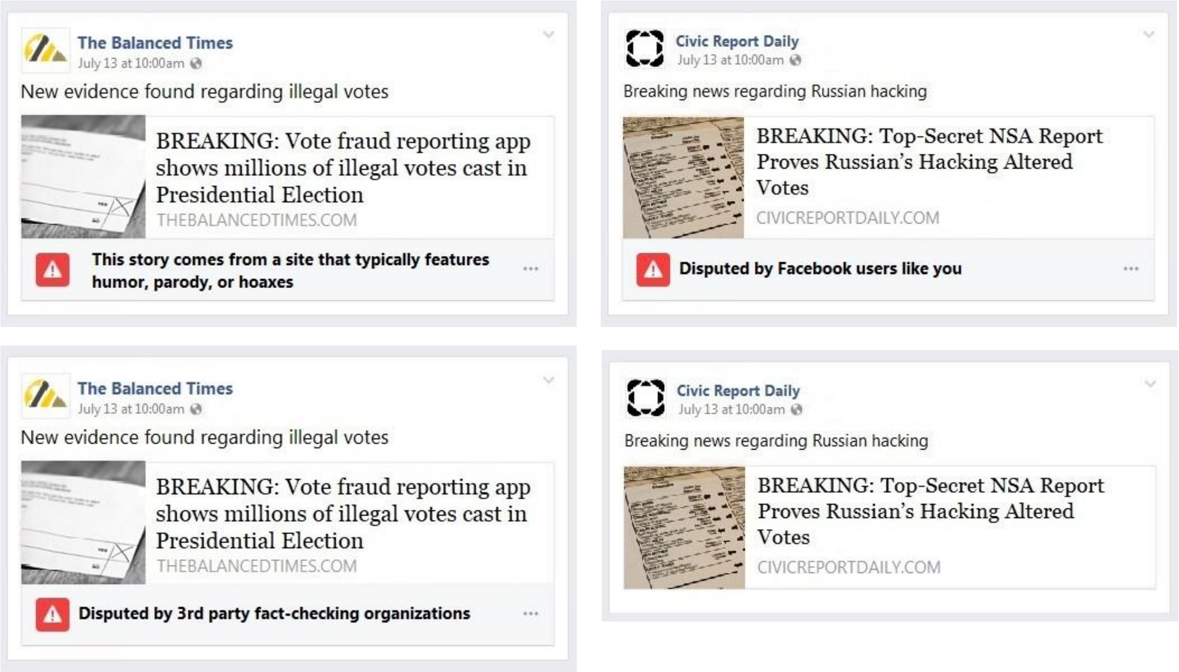
Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category: Story, self-identified

Table S4a. Study 2 flagging system perceptions models, without exclusions

	Reactance		Value of flagging	
	Main effect	Interaction	Main effect	Interaction
Story, Facebook (StoryFB) [a]	-.123, .552	-.118, .611	-.577, .020	-.625, .015
Site, self-identified (SiteSID) [a]	-.143, .527	-.091, .631	-.611, .017	**-.637, -.003**
Site, Facebook (SiteFB) [a]	-.132, .542	-.162, .560	**-.651, -.053**	**-.714, -.080**
Inaccurate issue beliefs, t-1	—	**.024, .572**	—	-.435, .003
StoryFB X inaccurate	—	-.500, .270	—	-.216, .399
SiteSID X inaccurate	—	-.603, .141	—	-.252, .342
SiteFB X inaccurate	—	-.377, .371	—	-.166, .432
Constant	**3.234, 3.714**	**3.112, 3.625**	**4.967, 5.392**	**5.109, 5.615**
Variance components				
Random intercept	1.738, 2.251	1.704, 2.211	1.419, 1.820	1.481, 1.968
Residual	.562, .702	.561, .700	.346, .431	.349, .450
Number of observations	1255	1255	1256	1256
Number of participants	628	628	628	628
Likelihood ratio test	$\bar{\chi}^2(1) = 538.55$ p <.001	$\bar{\chi}^2(1) = 520.24$ p <.001	$\bar{\chi}^2(1) = 659.23$ p <.001	$\bar{\chi}^2(1) = 512.64$ p <.001

Notes. Cells show 95% Confidence Intervals (CI) for random-effects model (nested by respondent ID). Coefficients with CIs that do not contain zero are in **bold**. a. Reference category: Story, self-identified

Figure S1. Study 1 visual presentation of misinformation and flags



The Balanced Times
July 13 at 10:00am
New evidence found regarding illegal votes

BREAKING: Vote fraud reporting app shows millions of illegal votes cast in Presidential Election
THEBALANCEDTIMES.COM

This story comes from a site that typically features humor, parody, or hoaxes

Civic Report Daily
July 13 at 10:00am
Breaking news regarding Russian hacking

BREAKING: Top-Secret NSA Report Proves Russian's Hacking Altered Votes
CIVICREPORTDAILY.COM

Disputed by Facebook users like you

The Balanced Times
July 13 at 10:00am
New evidence found regarding illegal votes

BREAKING: Vote fraud reporting app shows millions of illegal votes cast in Presidential Election
THEBALANCEDTIMES.COM

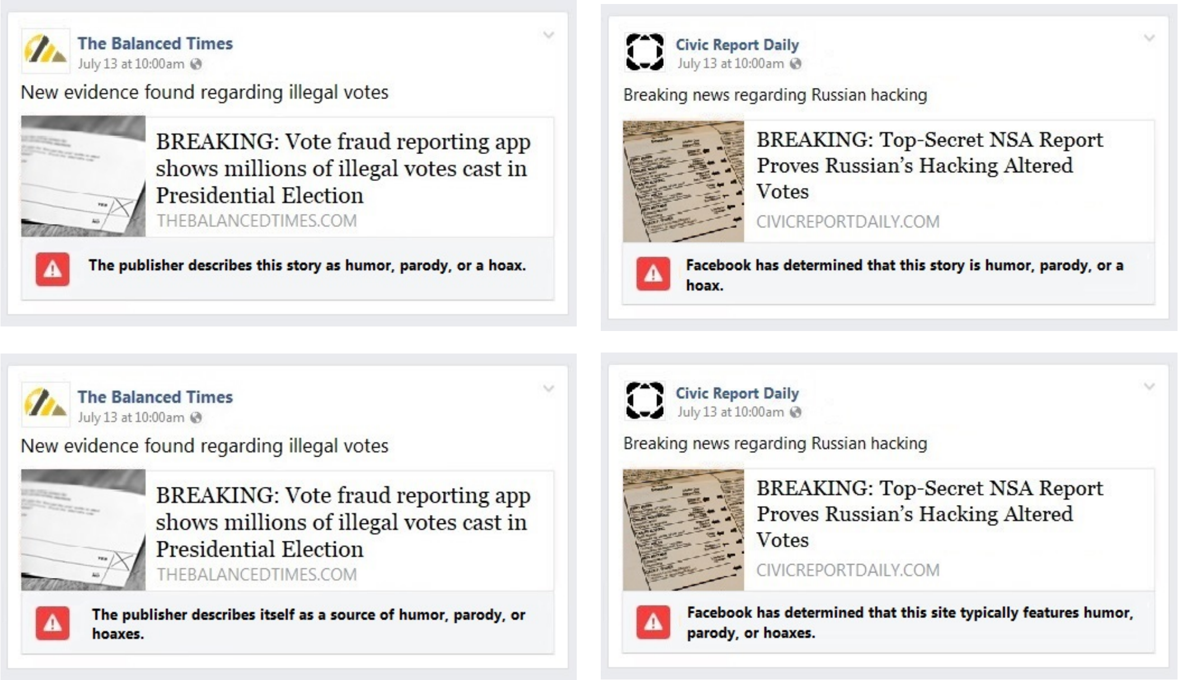
Disputed by 3rd party fact-checking organizations

Civic Report Daily
July 13 at 10:00am
Breaking news regarding Russian hacking

BREAKING: Top-Secret NSA Report Proves Russian's Hacking Altered Votes
CIVICREPORTDAILY.COM

Note. All flags, messages, and sources are shown, but there were eight combinations in all: 2 (messages: Vote fraud in *The Balanced Times* or Russian hacking in *Civic Report Daily*) X 4 (flag type: self-identified humor flag, fact-checker flag, peer-generated flag, no flag)

Figure S2. Study 2 visual presentation of misinformation and flags



The Balanced Times
July 13 at 10:00am

New evidence found regarding illegal votes

BREAKING: Vote fraud reporting app shows millions of illegal votes cast in Presidential Election
THEBALANCEDTIMES.COM

The publisher describes this story as humor, parody, or a hoax.

Civic Report Daily
July 13 at 10:00am

Breaking news regarding Russian hacking

BREAKING: Top-Secret NSA Report Proves Russian's Hacking Altered Votes
CIVICREPORTDAILY.COM

Facebook has determined that this story is humor, parody, or a hoax.

The Balanced Times
July 13 at 10:00am

New evidence found regarding illegal votes

BREAKING: Vote fraud reporting app shows millions of illegal votes cast in Presidential Election
THEBALANCEDTIMES.COM

The publisher describes itself as a source of humor, parody, or hoaxes.

Civic Report Daily
July 13 at 10:00am

Breaking news regarding Russian hacking

BREAKING: Top-Secret NSA Report Proves Russian's Hacking Altered Votes
CIVICREPORTDAILY.COM

Facebook has determined that this site typically features humor, parody, or hoaxes.

Note. All flags, messages, and sources are shown, but there were eight combinations in all: 2 (messages: Vote fraud in *The Balanced Times* or Russian hacking in *Civic Report Daily*) X 4 (humor flag type: story self-identified flag, story Facebook flag, site self-identified flag, site Facebook flag, no flag). No flag (control) condition the same as in Study 1.