Commentary on Eskreis-Winkler and Fishbach (2019): A Tendency to Answer Consistently Can Generate Apparent Failures to Learn From Failure

Stephen A. Spiller

UCLA Anderson School of Management

August 5, 2024

Accepted for publication at *Psychological Science*

Author Note

Stephen A. Spiller https://orcid.org/0000-0001-6951-6046

Data, materials, and code are available at https://researchbox.org/2603.

Correspondence concerning this article should be addressed to Stephen A. Spiller, UCLA Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA, 90095. Email: stephen.spiller@anderson.ucla.edu

Abstract

Recent research proposes failure undermines learning: people learn less from failure than from success because failure is ego-threatening and causes people to tune out. I argue evidence from the core paradigm (the Script Task) is not sufficient to support that claim. When people do not learn from test feedback, they may give internally-consistent answers on a subsequent test. The Script Task's scoring guidelines mark consistent answers as correct following success but incorrect following failure. As a result, differences in performance between conditions may result from equivalent learning combined with consistent responding when people do not learn. A descriptive mathematical model shows lower performance is insufficient to conclude less learning. An experiment demonstrates a retroactive manipulation without feedback replicates the effect. Because the effect of failure on performance is confounded with consistency, unless consistency is ruled out, the Script Task is not diagnostic regarding whether people learn less from failure.

Statement of Relevance

Prior research has proposed that people learn less from failure than from success because the threat from failure feedback leads them to tune out. This provocative finding is of general interest to researchers across sub-fields of psychology as well as the general public. It has been well-cited, has served as the basis for practitioner-oriented publications, and provides a paradigm (the Script Task) that multiple independent research teams use to understand failures to learn from failure. In the Script Task, learning is operationalized as test performance. But the scoring guidelines used to assess performance in the Script Task mean a plausible alternative explanation is equally compatible with the data. The effect of failure on performance is confounded with people's tendency to give consistent responses across multiple tests. Uniform learning coupled with a uniform tendency to respond consistently when people do not learn can generate an apparent failure to learn from failure.

Research Transparency Statement

General Disclosures

Conflicts of interest: I have no conflicts of interest to disclose.

Funding: This research was supported by funding from the UCLA Anderson School of Management.

Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article.

Ethics: This research was certified exempt from the relevant IRB.

Computational reproducibility: The author is applying for a Computational Reproducibility Badge which will be awarded pending checks by the STAR Team.

Experiment Disclosures

Preregistration: The hypotheses, methods, and primary analysis plan for Experiment, Posttest, S1, and S2 were preregistered (see https://researchbox.org/2603) on 2024-02-17, 2024-06-24, 2024-06-24, and 2024-04-05, respectively, prior to data collection which began later the same day for each experiment.

Materials: All study materials are publicly available (https://researchbox.org/2603).

Data: All primary data are publicly available (https://researchbox.org/2603).

Analysis scripts: All analysis scripts are publicly available (https://researchbox.org/2603).

Recent research suggests people learn less from failure than from success because the threat from failure causes them to tune out (Eskreis-Winkler & Fishbach 2019). Subsequent research has replicated the effect using the core paradigm (the Script Task; Eskreis-Winkler et al. 2024; Keith et al. 2022; Gok & Fyfe 2022). I show the test of the effect of failure is perfectly confounded with the test of participants' tendency to respond consistently.

The Script Task

The Script Task from Eskreis-Winkler and Fishbach's (2019) Study 2a exemplifies the core paradigm; see Table. Participants were randomly assigned to experience success or failure. An initial Round 1 quiz provided an opportunity to learn through feedback. Participants answered "Which of the following characters in an ancient script represents an animal?" by selecting ᚴ or ᛗ. Regardless of their answer, success participants were notified "You answered this question correct!" and failure participants were notified "You answered this question incorrect!" Participants then answered two more questions, regarding person (ᛇ, ᚼ) and bird (ᛗ, ᛉ), and received the same condition-specific feedback after each.

In Round 2, participants answered "Which of the following characters represents a non-living, stationary object?" three times, once for each of the three Round 1 symbol pairs: (ᚴ, ᛗ), (ᛇ, ᚼ), and (ᛗ, ᛉ). The correct Round 2 answers were the complements of the correct Round 1 answers. For success participants, whatever the participant selected (e.g., ᛉ for bird) was deemed correct in Round 1, so the other symbol (i.e., ᛗ) was correct in Round 2. For failure participants, whatever the participant selected (e.g., ᛉ for bird) was deemed incorrect in Round 1, so that same symbol (i.e., ᛉ) was correct in Round 2. Learning was operationalized as Round 2 performance and was approximately 20 percentage points higher after success than failure.

Table

Script Task With Example Correct Answers, Consistent Responses, and Scores, by Condition

		Success Condition	Failure Condition
Round 1	Question 1	Which of the following characters in an ancient script represents an animal? 𐤀 or 𐤌	
	Potential guess	𐤀	𐤀
	Feedback	Correct	Incorrect
	Implied correct answer	Animal = 𐤀	Animal = 𐤌
	Question 2	Which of the following characters in an ancient script represents a person? 𐤋 or 𐤇	
	Potential guess	𐤇	𐤇
	Feedback	Correct	Incorrect
	Implied correct answer	Person = 𐤇	Person = 𐤋
	Question 3	Which of the following characters in an ancient script represents a bird? 𐤍 or 𐤓	
	Potential guess	𐤓	𐤓
	Feedback	Correct	Incorrect
	Implied correct answer	Bird = 𐤓	Bird = 𐤍
Round 2	Question 1	Which of the following characters represents a non-living, stationary object? 𐤀 or 𐤌	
	Implied correct answer	Animal = 𐤀, so Object = 𐤌	Animal = 𐤌, so Object = 𐤀
	Response if learn symbol for animal	𐤌	𐤀
	Response if guess consistently	𐤌	𐤌
	Question 2	Which of the following characters represents a non-living, stationary object? 𐤋 or 𐤇	
	Implied correct answer	Person = 𐤇, so Object = 𐤋	Person = 𐤋, so Object = 𐤇
	Response if learn symbol for person	𐤋	𐤇
	Response if guess consistently	𐤋	𐤋
	Question 3	Which of the following characters represents a non-living, stationary object? 𐤍 or 𐤓	
	Implied correct answer	Bird = 𐤓, so Object = 𐤍	Bird = 𐤍, so Object = 𐤓
	Response if learn symbol for bird	𐤍	𐤓
	Response if guess consistently	𐤍	𐤍
Score	If everyone learns each symbol	100%	100%
	If everyone guesses consistently	100%	0%
	If half learn and half guess consistently	100%	50%

Note. Round 1 guesses depict modal choices in experiment. Guessing the other option would lead to the same feedback, so both correct answers and consistent responses would be reversed.

Differences in Performance Are Confounded With Consistency

The Round 2 scorecard depends on Round 1 responses and condition. Regardless of whether participants learn, consistent responses (e.g., ᛉ is a bird, ᛗ is an inanimate object) are deemed correct after success but incorrect after failure. This positively confounds performance with consistency for success participants and negatively confounds performance with consistency for failure participants. The test of failure's effect on performance is thus perfectly confounded with, and exactly equivalent to, the test of greater-than-chance consistency (Abelson 1995; Brauer & Judd 2000; Shaffer 1977). If people learn equally from success and failure, equal tendencies to respond consistently when they do not learn will generate apparent failures to learn from failure. This is depicted in the Table.

Why Would People Answer Consistently?

When participants learn in Round 1, they answer correctly in Round 2. But not everyone learns everything. Performance averaged near 75%. If people guessed randomly when they did not learn, the probability of learning was 50%.[1] But random guessing is not the only response strategy when someone has not learned. Someone who has not learned (and so cannot truly know the correct answer in this task) may systematically guess instead. Absent learning, why might participants respond consistently? Prior beliefs and measurement present two possibilities.[2]

First, consider belief-induced consistency. Participants may rely on stable, preexisting beliefs to generate answers across rounds. Features that make one symbol a better representation of an animate being (e.g., physical resemblance, sound-shape mapping, or convention) may make it a worse representation of an inanimate object. This can lead to consistent responses.

[1] Half of random guesses are correct, half are incorrect. The 25% of answers that are incorrect represent incorrect-guessing, so another 25% represent correct-guessing. The remaining 50% represent learning.
[2] Other processes can also generate consistency (e.g., alternating responses). Consistent responses generate the confound, no matter the cause(s).

Second, consider measurement-induced consistency. Taking tests can induce consistency. Beliefs that are initially independent may shift to align with one another through deliberation (e.g., Holyoak & Simon 1999). Alternatively, people may recruit their Round 1 responses for consideration when answering Round 2 (e.g., Feldman & Lynch 1988). In either case, responding in Round 1 induces a consistent response in Round 2.

Belief-induced consistency depends on preexisting associations and in principle could be addressed by selecting stimuli for which no individual has any tendency to give complementary answers. Measurement-induced consistency may still arise even with such stimuli. Any type of consistency when people do not learn results in the confound: consistent responding generates better performance following success than failure.

Model and Evidence

Consistent responding in the Script Task leads to lower performance following failure. Next, I present a descriptive mathematical model to more-precisely specify the concern. Given the arguments above that participants likely respond consistently when they do not learn, I then examine whether participants respond consistently to the Script Task questions when they cannot learn. I retroactively assign condition after an adapted Script Task with no feedback (and therefore no learning) yet find an apparent effect on performance. In an extension, I retroactively reassign condition labels in the original studies' datasets and replicate the same apparent effect.

A Descriptive Mathematical Model of Performance in the Script Task

Model

After each Round 1 answer, participants receive feedback. Based on that feedback, there is some probability that they learn the meaning of the symbol matching the Round 1 concept.[3]

[3] Drawing on the American Psychological Association's (n.d.) definition, to *learn* in this context means to gain new knowledge from experience. This involves attending to relevant information and integrating it with what is already

Call the probability of learning from feedback, averaged across success and failure, λ. Call the additional probability of learning from success, beyond learning on average, δ. The probability of learning from success is then $\lambda + \delta$ and the probability of learning from failure is $\lambda - \delta$. If people learn equally from either, $\delta = 0$ and the probability of learning after any feedback is λ.

If participants learn the implied meaning of the symbol matching the Round 1 concept, then they answer the corresponding Round 2 question correctly.[4] Even if they do not learn the implied meaning of the target symbol, people may still answer the corresponding Round 2 question systematically (e.g., by guessing systematically rather than randomly). Call the probability of giving an internally-consistent answer in Round 2, conditional on not learning the implied meaning of the target symbol, ρ.

Recall that consistent answers are scored as correct following success but incorrect following failure. The probability that people answer correctly in Round 2 after success, P(correct|success), is: (probability learned) + (probability did not learn) × (probability respond consistently conditional on having not learned) = $(\lambda + \delta) + \big(1 - (\lambda + \delta)\big)\rho$. The probability that people answer correctly in Round 2 after failure, P(correct|failure), is: (probability learned) + (probability did not learn) × (probability *do not* respond consistently conditional on having not learned) = $(\lambda - \delta) + \big(1 - (\lambda - \delta)\big)(1 - \rho)$.

---

known. The target to be learned here is the implied meaning of the symbol matching the Round 1 concept (e.g., if one indicates ᛉ for bird, failure feedback implies ᛗ represents bird). This can be determined by attending to the question, attending to one's answer, attending to the feedback, and integrating them together. Using this knowledge, participants can then answer Round 2 via a process of elimination. Some participants might systematically answer correctly without having learned (e.g., through systematic guessing or preexisting incidentally-accurate beliefs), but learning provides a sound basis on which to do so knowledgeably (i.e., the feedback-implied meaning). Other than leading to correct answers, the model is agnostic regarding the psychological process and consequences of learning.

[4] This assumption can be relaxed by redefining λ and δ to refer to the joint probability of both learning and using that knowledge, rather than just learning. Both the original paper's proposal and the current model assume the probability of applying knowledge deduced from feedback is the same across conditions. For the current model, this assumption is merely for simplicity and is not a necessary condition. Differential application of what was learned could provide another possible interpretation of the results (e.g., "I learned ᛗ means bird in Round 1, but everything else I know indicates ᛉ means bird. I trust that broader knowledge base more, so ᛗ must mean inanimate object.")

Results

The difference between performance in the success condition and performance in the failure condition is then given by $(2\rho - 1)(1 - \lambda) + \delta$. If $\delta = 0$ and people learn equally well from success or failure, the difference is $(2\rho - 1)(1 - \lambda)$. There are four key related results.

1. Any Result Can Be Represented by Multiple Parameter Configurations

With three parameters determining performance and only two conditions, the parameters are not uniquely identified: any pattern of results has multiple interpretations. For example, success performance of 85% and failure performance of 65% is consistent with the original explanation: greater learning from success than failure ($\lambda = 0.5$, $\delta = 0.2$) and no systematic consistency ($\rho = 0.5$). But it is also consistent with equal learning from success or failure ($\lambda = 0.5$, $\delta = 0$) and high consistency ($\rho = 0.7$).[5]

2. Reduced Performance Does Not Imply Reduced Learning

Following from Result 1, observing that performance after failure is lower than performance after success is not sufficient to conclude that there is less learning after failure than there is after success (i.e., that $\delta > 0$). It only enables that conclusion if one assumes or can prove that there is no consistency conditional on not learning (i.e., that $\rho \leq 0.5$).

3. Consistent Responding Can Masquerade as a Difference in Learning

With equal learning from success or failure, performance after failure is lower than performance after success if people answer consistently when they do not learn ($\rho > 0.5$). This could plausibly account for the original effect. For example, for $\rho = 0.7$, $\lambda = 0.5$, and $\delta = 0$, average performance after success is 85% and average performance after failure is 65%.

[5] Table S1 in the supplemental materials estimates two sets of parameters for each study. One set assumes no systematic consistency in the absence of learning and freely estimates λ and δ. The other set assumes no differential learning from success or failure and freely estimates λ and ρ.

4. Randomly Reassigning Condition Labels Does Not Change the Estimated Effect

Recall the success scorecard is the complement of the failure scorecard. Suppose that before calculating performance, every observation has its condition label flipped: people who received failure feedback are labeled "success" and people who received success feedback are labeled "failure." As a result, "success" observations (i.e., people who received failure feedback) would be scored according to the success scorecard. Their new scores would be the complement of their true scores: rather than $P(correct|failure)$, they would be calculated as $1 - P(correct|failure)$. Similarly, "failure" observations (i.e., people who received success feedback) would be scored according to the failure scorecard. Their new scores would be the complement of their true scores: rather than $P(correct|success)$, they would be calculated as $1 - P(correct|success)$.

Perhaps counterintuitively, the difference between scores in the group labeled "success" (which received failure feedback) and scores in the group labeled "failure" (which received success feedback) is again $(2\rho - 1)(1 - \lambda) + \delta$. This is the same value, with the same sign, as the difference using correct condition labels.

Given equal cell sizes, any shuffling of condition labels in the raw response data will necessarily result in two subsamples, each balanced between success and failure. In one, observations are scored by the correct scorecard; in the other, observations are scored by the wrong scorecard. For both, the difference in means is $(2\rho - 1)(1 - \lambda) + \delta$. The overall difference in means will be a weighted average of those two differences, so the same difference holds for any shuffling of condition labels.[6] When analyzing results of the Script Task, whether the scorecard used for analysis matches the one implied by the feedback manipulation does not

[6] In a sample with similar but not equal cell sizes, the difference will be similar but not necessarily equal.

affect the results. Any allocation of condition labels to the raw response data results in the same effect, despite the fact that randomly-shuffled condition labels cannot affect learning.[7]

I next test these implications using a version of the Script Task that precludes learning the correct answer.

Experiment

Method

Participants

I aimed to recruit 400 participants from Amazon Mechanical Turk using CloudResearch's approved participant pool (Hauser et al. 2023; Litman et al. 2017). The dataset included 401 complete observations (225 men, 165 women, 7 non-binary or third gender, 4 preferred not to say; after excluding one implausible response, M_{age} = 43.66, SD_{age} = 13.14). 6 participants were missing a response to at least one quiz question, leaving 395 participants for analysis. Attrition and alternate exclusion rules are detailed in the supplemental materials.

Design

This experiment was adapted from the original paper's Study 2a (described above and represented in the Table). There were three changes in addition to the larger sample. First, participants received no feedback, making the participant experience indistinguishable between conditions and precluding participants from learning the correct answer; instructions were adjusted accordingly. Second, success vs. failure condition was assigned retroactively at the end of the experiment, after all measures were collected. Together, these changes made it impossible for condition to affect behavior. Third, answers were not incentivized; instructions were adjusted accordingly.

[7] Related consequences can occur whenever one differentially transforms data across conditions. Whether such consequences are cause for concern depends on the research question.

Results

I calculated consistency (proportion of complementary responses) and performance (proportion of correct responses) and regressed each on a contrast-coded variable for retroactive condition label (1 = success, -1 = failure).

Consistency Analysis

Participants' answers were internally consistent across Rounds 1 and 2, as indicated by the intercept (M = 88%, SD = 23%; b = 0.878, se = 0.012; vs. 50%: $t(393) = 32.65$, $p < .001$, Cohen's $d = 1.64$). As anticipated given retroactive random assignment of condition, consistency neither substantively nor significantly varied by condition (success: M = 86%, SD = 24%; failure: M = 89%, SD = 22%; b = -0.016, se = 0.012; $t(393) = -1.41$, p =.158, Cohen's $d = 0.14$). The null hypothesis of no difference between conditions must be true, as random assignment came after both rounds.

Performance Analysis

The intercept reveals average performance did not differ from chance (M = 48%, SD = 44%; b = 0.484, se = 0.012; vs. 50%: $t(393) = -1.41$, $p = .158$, Cohen's $d = 0.04$). Recall consistency and performance are positively confounded following success but negatively confounded following failure. As a result, because consistency was high in both conditions, performance was substantially and significantly higher in the success condition than in the failure condition (success: M = 86%, SD = 24%; failure: M = 11%, SD = 22%; b = 0.378, se = 0.012; $t(393) = 32.65$, $p < .001$, Cohen's $d = 3.29$). Analyses of consistency and performance are precisely equivalent. Because the answer key is flipped across conditions, the test of the intercept against chance for consistency is equivalent to the test of the effect of condition on performance,

and the test of the effect of condition on consistency is equivalent to the test of the intercept against chance for performance (Abelson 1995; Brauer & Judd 2000; Shaffer 1977).[8]

As indicated by the model, if condition labels are flipped in the raw response data and scores calculated anew using the scorecards matching the new labels, rather than finding a reversed effect, we instead reproduce the same signed difference between conditions (success: M = 89%, SD = 22%; failure: M = 14%, SD = 24%; b = 0.378, se = 0.012; $t(393) = 32.65$, $p < .001$, Cohen's $d = 3.29$). In expectation, any assignment of condition will generate an equivalent raw difference.

Extensions

Posttest Assessing Types of Consistency. The results above are compatible with belief-induced consistency, measurement-induced consistency, or both. A posttest ($N = 403$) indicates both may contribute, though possibly differentially across stimuli. The posttest replicated the experiment with a key change: half of the sample faced the standard order (i.e., animate version of each question in Round 1, inanimate version of each question in Round 2); the other half faced the other order (i.e., inanimate version in Round 1, animate version in Round 2).[9] Full results are reported in the supplemental materials.

In each order, more than half of participants gave consistent responses to each version of each question (e.g., ᛉ for bird and ᛗ for inanimate object, or vice versa; ps < .001). Supporting a role for belief-induced consistency for question 3, in Round 1 participants tended to select ᛉ for bird and ᛗ for inanimate object (ps < .001). There was no such evidence for question 1 or 2. Supporting a role for measurement-induced consistency for questions 2 and 3, the inanimate

[8] Apparent differences in Cohen's d are due to use of total standard deviation when calculating the one-sample Cohen's d for levels but pooled standard deviation when calculating the two-sample Cohen's d for differences.
[9] Thank you to the AE for suggesting this design.

choice shares elicited in Round 2 differed from those elicited in Round 1 (e.g., the choice share for whether ᛗ or ᛉ represents an inanimate object differed when it followed vs. preceded the question of whether ᛗ or ᛉ represents a bird; *p*s < .001). There was no such evidence for question 1.

Though question 1 answers were internally-consistent, neither test of type of consistency was significant. This illustrates the implications of heterogeneity. If half of the population believes ᚹ represents animal and ᛗ represents an inanimate object and half believes the opposite, the null hypothesis of equal choice shares for each test would be true, despite the presence of belief-induced consistency and the possibility of measurement-induced consistency.

The Effect of Failure for Belief-Induced Inconsistency. Two experiments in the supplement tested the effect of success vs. failure feedback when participants tended to give repeated responses across rounds rather than consistent responses across rounds. Experiment S1 used stimuli selected to induce repeated responding (i.e., systematic inconsistency). As predicted by the model, the effect reversed, revealing apparent failure to learn from success. Experiment S2 manipulated belief-induced consistency, replicating a failure to learn from failure when the stimuli induced consistency and a failure to learn from success when the stimuli induced inconsistency. The reversal of the effect depending on the stimuli is explainable by consistency but not by tuning out. Unlike in the experiment above, Experiments S1 and S2 enabled learning by providing feedback, demonstrating that consistency still matters when people can learn.

Reanalysis of Original Studies. Using data from each Script Task study from the original paper, I reversed condition labels and recalculated performance according to the new scorecard. As the model indicates and the experiment finds, the signed difference in means remains the same; see Table S7 in the supplement. Relabeling conditions implies using the wrong

scorecard. As a result, all correct answers are scored as incorrect and all incorrect answers are scored as correct, thereby reversing the difference. Because the difference between groups is reversed again due to relabeling, the original difference (now twice reversed) reappears. Shuffling labels is similarly ineffectual; see Table S8 in the supplement. If a researcher had access to raw question responses but not condition labels, any retroactive assignment of condition labels would generate the same apparent effect. This is because the difference in performance is confounded with the level of consistency.

The supplement details how a related set of concerns can account for each of the results reported in the original paper.

Discussion

Any tendency toward consistency will induce an apparent effect of failure on performance in the Script Task. Prior theory suggests people are likely to respond consistently. The experiment indicates that when they receive no feedback and cannot learn, participants do respond consistently. The scoring guidelines mean reversing or shuffling condition labels reproduces the original effect. Together, these results offer a plausible alternative explanation for apparent failures to learn from failure. Observing that failure reduces performance in the Script Task is insufficient to conclude that failure reduces learning. Determining failure's effect on learning requires making strong assumptions, ruling out any role of consistency, or using a different paradigm.

References

Abelson, R. P. (1995). *Statistics as Principled Argument*. Erlbaum.

American Psychological Association. (n.d.). Learning. In *APA dictionary of psychology*. Retrieved July 6, 2024, from https://dictionary.apa.org/learning

Brauer, M., & Judd, C. M. (2000). Defining variables in relationship to other variables: When interactions suddenly turn out to be main effects. *Journal of Experimental Social Psychology*, *36*(4), 410-423.

Eskreis-Winkler, L., & Fishbach, A. (2019). Not learning from failure—The greatest failure of all. *Psychological Science*, *30*(12), 1733-1744.

Eskreis-Winkler, L., Woolley, K., Erensoy, E., & Kim, M. (2024). The exaggerated benefits of failure. *Journal of Experimental Psychology: General*.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*(3), 421-435.

Gok, S., & Fyfe, E. R. (2022). Learning from failure with self vs task focused feedback. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, *55*(8), 3953-3964.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*(1), 3-31.

Keith, N., Horvath, D., Klamar, A., & Frese, M. (2022). Failure to learn from failure is mitigated by loss-framing and corrective feedback: A replication and test of the boundary

conditions of the tune-out effect. *Journal of Experimental Psychology: General*, *151*(8), e19-e25.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433-442.

Shaffer, J. P. (1977). Reorganization of variables in analysis of variance and multidimensional contingency tables. *Psychological Bulletin*, *84*(2), 220-228.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493-504.