American Psychological Association

Journal of Personality and Social Psychology:
Interpersonal Relations and Group Processes

ISSN: 0022-3514

2019, Vol. 117, No. 6, 1139–1144
http://dx.doi.org/10.1037/pspi0000220

REPLY

It Helps to Ask: The Cumulative Benefits of Asking Follow-Up Questions

Michael Yeomans, Alison Wood Brooks,
and Karen Huang
Harvard Business School

Julia Minson
Harvard Kennedy School of Government

Francesca Gino
Harvard Business School

In a recent article published in *Journal of Personality and Social Psychology* (*JPSP*; Huang, Yeomans, Brooks, Minson, & Gino, 2017), we reported the results of 2 experiments involving "getting acquainted" conversations among strangers and an observational field study of heterosexual speed daters. In all 3 studies, we found that asking more questions in conversation, especially follow-up questions (that indicate responsiveness to a partner), increases interpersonal liking of the question asker. Kluger and Malloy (2019) offer a critique of the analyses in Study 3 of our article. Though their response is a positive signal of engaged interest in our research, they made 3 core mistakes in their analyses that render their critique invalid. First, they tested the wrong variables, leading to conclusions that were erroneous. Second, even if they had analyzed the correct variables, some of their analytical choices were not valid for our speed-dating dataset, casting doubt on their conclusions. Third, they misrepresented our original findings, ignoring results in all 3 of our studies that disprove some of their central criticisms. In summary, the conclusions that Kluger and Malloy (2019) drew about Huang et al. (2017)'s findings are incorrect. The original results are reliable and robust: Asking more questions, especially follow-up questions, increases interpersonal liking.

Keywords: rejoinder, follow-up questions, social relations model, communication

In a recent article published in *Journal of Personality and Social Psychology* (*JPSP*; Huang et al., 2017), we reported the results of two experiments involving "getting acquainted" conversations among strangers and an observational field study of heterosexual speed daters. In all three studies, we found that asking more questions in conversation, especially follow-up questions (that indicate responsiveness to a partner), increases interpersonal liking of the question asker.

Kluger and Malloy (2019), henceforth "KM," offer a critique of the analyses in our article. KM's critique does not consider any of the experimental results from Studies 1 and 2 and is based exclusively on their reanalyses of the speed-dating dataset in Study 3 of our article, which they primarily conduct by applying the Social Relations Model (henceforth "SRM") to the dataset.

KM's response to our article is a positive signal of engaged interest in our research on question-asking. This exchange provides a canonical example of open science: KM reanalyzed a dataset we posted publicly on Open Science Framework (OSF) alongside our article in 2017, which was itself constructed from data collected by a different research team (Jurafsky, Ranganath, & McFarland, 2009; Ranganath, Jurafsky, & McFarland, 2009, 2013), who shared it generously with us in 2016. Furthermore, KM helped to connect question-asking to the listening literature, which they cite in their work.

However, KM made several mistakes in conducting their analyses that render their critique invalid. First, they tested the wrong variables in the speed-dating dataset. While they assert that they tested "second dates received" (as in our original article), the results they report actually tested "second dates offered." This was somewhat challenging to diagnose because most of their analyses are not reproducible. They failed to provide code for many analyses, while others were nearly impossible to read or run. Second, even if they had analyzed the correct variables, some of their analytical choices are not valid for the speed-dating dataset, which casts doubt on their conclusions. Third, they misrepresented our original findings, ignor-

Michael Yeomans, Alison Wood Brooks, and Karen Huang, Department of Negotiations, Organizations and Markets, Harvard Business School; Julia Minson, Department of Public Policy, Harvard Kennedy School of Government; Francesca Gino, Department of Negotiations, Organizations and Markets, Harvard Business School.

All data and code for analyses reported here are stored at https://osf.io/8k7rf/.

Correspondence concerning this article should be addressed to Michael Yeomans, Department of Negotiations, Organizations and Markets, Harvard Business School, Soldiers Field, Boston, MA 02163. E-mail: myeomans@hbs.edu

ing results in all three studies of our article that disprove some of their central criticisms about our Study 3.

In this response, we describe these three issues and, with an eye toward improving our research community, provide recommendations that might help other researchers avoid the same mistakes. We recognize that the debate instigated by KM has led to a technical consideration of methods that may be difficult or tedious to follow. For readers interested in the technical details of KM's errors, we explain below. For others, we offer a summary of our response: The conclusions provided by KM are incorrect, and our original findings are reliable and robust.

Error #1: KM Tested the Wrong Variables

There is an error in KM's article that explains why their results were directionally the opposite of ours: KM tested the wrong outcome variables. In their Table 8, they describe the outcome as "Receiving a Second Date Offer," but in their code, they actually tested whether someone *offered* a second date (on line 75 of this file: https://osf.io/zvf7p/). Likewise, in Table 7, they report two outcomes: "Number of dates offered" and "Number of date offers received." However, these column labels are reversed (on lines 36–38 of this file: https://osf.io/3f4k7/). In Appendix, we include detailed explanations of KM's code, line by line, that illustrate exactly how these mistakes made their way into the article. This same error—switching the outcome variable—was committed in different ways across at least three separate files. When the column labels of Table 7 are corrected, it corroborates our earlier finding: asking more questions is associated with receiving more second-date offers.

There are places in KM's code where validation checks could have revealed their coding errors. For example, one of their validation checks returned the result that men received more second-date offers than women, a highly unusual reversal of the typical pattern that women receive more second-date offers than men (lines 76–81 of this file: https://osf.io/dcmvj/). However, the comments in the R script suggest that instead of further exploring the sources of such red flags, KM ignored them. KM do not provide any record (syntax or code) for most of their other analyses, which were conducted in SPSS and BLOCKO. However, we suspect these analyses are also wrong because their data cleaning for SPSS was completed in an R script where the outcome variables were mislabeled.

KM's Analyses Are Not Reproducible

When we tried to reproduce every number in KM's article ourselves, we could do so for some analyses but not others, as the code was missing or incomplete. For example: The code for Table 1 and Tables 3–6 is missing; we found the code to produce Table 2 (that is accurate) and Tables 7–8 (that is incorrect). Failing to provide reproducible analyses does not mean that the analyses are incorrect, per se. However, in this case, because KM's critique was made entirely on analytical grounds, best practices suggest that they should provide reproducible analyses so that the research community can evaluate the validity of their claims.

There are several aspects of their analyses that undermine their reproducibility. KM relied on three different software applications: BLOCKO (Kenny, Kashy, & Cook, 2006), the SPSS SRM package (Ackerman, Kashy, & Corretti, 2015), and R. This mix of tools means that anyone who wants to reproduce their work must have all three software applications, and the transitions between applications are not clear. Sometimes output from BLOCKO and SPSS was copy-pasted into R as block text without a record of its provenance. Furthermore, neither BLOCKO nor SPSS is open source, so the internal code base cannot be inspected or corrected by outsiders.

As KM correctly note (pp. 11–12), both BLOCKO and SPSS have many limitations. BLOCKO cannot handle missing data or groups of more or less than 16 people, does not compute standard errors, and uses a GUI that does not record the analysis choices. The SPSS SRM package often requires arbitrary transformations of the data, cannot handle binary outcomes, and requires many *days* for a single model to converge. These shortcomings require the introduction of R scripts modified from a template written by David A. Kenny, an inventor of the SRM (lines 1–4 of this file: https://osf.io/zvf7p/). The final product is very hard for outsiders to read: Package dependencies are hidden or out-of-date, important variables are repeatedly renamed, and much of the documentation does not accurately describe what is in the code.

In summary, KM's analyses are not reproducible. This made it difficult for KM's work to be scrutinized (by ourselves, by their reviewers, and by the authors themselves), which perhaps helps to explain why their errors went undetected until now.

Error #2: KM Made Unjustified Analytical Choices

KM's second error is a collection of divergent analytical choices they made while applying the SRM to our speed-dating data. KM's critique of our work stems from their belief that we should have used the SRM for our analyses. In general, the SRM descriptively parses variance in correlational dyadic data—assigning variance on any dependent measure (e.g., liking) to each individual in the dyad (separately) and to the unique aspects of how the individuals behave as a dyad (together). This has been a useful conceptual framework in psychology for decades and continues to be a useful framework for researchers to analyze repeated dyadic interactions (Kenny, 1988; Kenny et al., 2006; Kenny & La Voie, 1984; Malloy & Kenny, 1986). However, many of the analytical choices in KM's application of the SRM—and, in this case, using the SRM at all—were not appropriate for our speed-dating dataset.

Poorly Estimated Dyad-Level Terms

Our greatest concern with KM's analysis is the use of dyad-level terms in their SRM models. This dyad effect subtracts the dyad-level average from the individual scores of the dyad members. This means that the remaining variance in the individual scores is because of *differences* in liking within the dyad, and their estimates of question-asking effects test whether questions affect the size of that difference. In other words, this means they treated a speed date where both people equally *liked* each other as equivalent to a speed date where both people equally *hated* each other (because the difference in liking would be zero in both cases).

In addition to being conceptually problematic, these dyad-level terms cannot be estimated well with only two observed outcomes per dyad. In fact, when we ran their SRM code in R, we repeatedly received a warning: "singular fit" (e.g., lines 81–87 and 150–163 of this file). The dyad-level variance is estimated as zero, which is a problem for reasons well explained by the documentation of that package (p. 48 of the lme4 manual; Bates, Mächler, Bolker, & Walker, 2015):

> [T]here are real concerns that (1) singular fits correspond to overfitted models that may have poor power; (2) chances of numerical problems and mis-convergence are higher for singular models (e.g., it may be computationally difficult to compute profile confidence intervals for such models); (3) standard inferential procedures such as Wald statistics and likelihood ratio tests may be inappropriate.

We note that SPSS has similarly ominous documentation about model singularity, though in at least one SPSS syntax file posted by KM, we see that KM manually turned this warning off (on line 115 of this file: https://osf.io/fd8wc/). We are not sure if BLOCKO tests for singularity. In any case, when the dyad term is removed, this warning disappears. If KM chose their model despite receiving these warnings, then they owe it to readers to justify this choice.

Improper Random-Effects Specifications

Even when these dyad-level terms are removed, the specifications of the partner-level terms violate the assumptions of the regression models used by KM. Throughout, they use random effects specifications for person-level and partner-level variables rather than fixed effects specifications. Random effects models make explicit assumptions about the correlation structure of the data (Angrist & Pischke, 2008), which do not hold in our speed-dating dataset.

We have conducted standard empirical specification tests (Hausman, 1978), which are now included in our original OSF repository (https://osf.io/8k7rf/). These tests explicitly reject any model that includes both partner-level random effects and any question-asking term. In our data, question-asking variance is primarily at the trait level. This means that models of our data that contain both a question-asking effect and partner-level random effects will fail the Hausman test. Although the estimated effects of question-asking are similar across both fixed- and random-effect specifications, these results suggest that KM did not adequately account for the error structure in our data.

Problematic Transformations and Exclusions

KM modified our dataset in several other ways that did not follow our original workflow. For example, they selected on the dependent and independent variables by dropping speed dates with zero follow-up questions and dropping people who offered dates to no one or to everyone. There were many transformations of the question counts (the main independent variable). They also did not compute cluster-robust standard errors, nor did they cross-validate or assess their model fit empirically. While some of these choices are more defensible than others, KM should have prioritized replicating our analytical decisions, so that the isolated contribution of each modification to the analysis could be more cleanly assessed.

These issues speak to KM's central argument, which is that our own analyses were inappropriate and that our data "should have been analyzed using the social relations model." However, KM did not conduct any robustness or specification tests to validate that the SRM is appropriate for our data. Instead, they point to two recent articles that have used SRM on speed-dating data (Ackerman et al., 2015; Jauk et al., 2016). A close examination of the literature reveals that the SRM has also been used to analyze for speed-dating data by several other researchers (Back & Kenny, 2010; Eastwick, Finkel, Mochon, & Ariely, 2007; Finkel & Eastwick, 2008; Finkel, Eastwick, & Matthews, 2007). However, these previous articles do not report any empirical validation of the SRM in comparison to other, simpler regression models. The closest we could find was a footnote in Finkel et al. stating that "One important consideration for SRM analyses is that investigators must include more than a single item to assess each construct of interest to separate the relationship effect from error" (Finkel et al., 2007; bottom of p. 153). Notably, our dataset has only a single item for the construct of interest (liking), which violates this guideline.

Having compared the results of a variety of specification tests, we did not choose the SRM. Both the dyad-level terms and the random-effects specifications induced problematic biases in the output of the model, which we avoided with a more conservative approach. Based on these considerations, we disagree with KM's assertion that our data is "properly analyzed using SRM." We encourage future researchers to consider the value of similar validation tests in data sets with similar parameters.

Error #3: KM Misrepresented the Evidence in Our Original Article

In their critique, KM state that "based on (Huang et al.'s [2017]) own results, there is no evidence that question asking is related to the likelihood of being offered a second date." In fact, we do provide evidence of this relationship, and KM ignored many other results in the original article that directly undercut their central claims.

KM Focused on a Boundary Condition

In our analyses of Study 3, we presented four models to test the link between follow-up questions and liking. Among these, we identified only one boundary condition model—with actor- and partner-level controls—that found no effect of follow-up questions on second-date success. Like our own boundary condition models, their SRM model also found no effect of follow-up questions on second-date success. However, all of KM's analyses include these actor- and partner-level controls as well, meaning their main null results can be seen as rephrased versions of the boundary condition model we already reported.

KM Ignored Supportive Evidence

We reported extensive supportive evidence for our conclusions in Study 3 itself, as well as in the experimental results of Studies 1 and 2, which KM ignored. For example, KM focus on the final-stage regressions in Study 3, which regress the outcome (liking) onto question counts. However, KM do not discuss our first-stage natural language processing (NLP) analyses that generated the question counts (the independent variable).

This omission matters because by overlooking the NLP analyses, KM also reject our distinction between different types of questions (follow-up, switch, introductory, and mirror questions).

Two results from our original article indicate that the distinction between question types is real and important. First, independent human coders reliably agreed on the distinctions between question types in our Studies 1 and 2 (with interrater reliability consistently around 90%). Second, we empirically validated our machine learning model by using nested cross-validation to ensure our model could reliably label question types out-of-sample (that produced turn-level labels that were as accurate as those produced by humans). KM discuss neither of these results.

In their analyses, KM combined question types. To justify this choice, they used a factor analysis that does not account for the nuances of data generated from natural language. The quantities in their factor analysis were not measured independently, but instead came from the same model, with a highly correlated error structure by design. In contrast, we applied a machine-learning algorithm that uses regularization to bias the estimated labels toward each other (Friedman, Hastie, & Tibshirani, 2010). This was a conservative approach that incorporated some of the initial modeling uncertainty into our final regression models. These methods were sophisticated, but that complexity was necessary to accommodate the rich data-generating processes of natural conversation.

Conclusion

KM made several errors in their analyses that undermine their conclusions. They tested the wrong variables, did not provide reproducible analyses, made unjustified analytical choices, and misrepresented our findings. In summary, they presented no new data that would lead one to believe that the original analyses and results of our article are anything but true and robust. We hope this discussion can encourage a community of inquiry around the important topic of question-asking in conversation (e.g., Brooks & John, 2018) and best practices in our field. And if any readers have other follow-up questions about our research, please do not hesitate to ask.

References

Ackerman, R. A., Kashy, D. A., & Corretti, C. A. (2015). A tutorial on analyzing data from speed-dating studies with heterosexual dyads. *Personal Relationships, 22,* 92–110. http://dx.doi.org/10.1111/pere.12065

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.2307/j.ctvcm4j72

Back, M. D., & Kenny, D. A. (2010). The Social Relations Model: How to understand dyadic processes. *Social and Personality Psychology Compass, 4,* 855–870. http://dx.doi.org/10.1111/j.1751-9004.2010.00303.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Brooks, A. W., & John, L. K. (2018). The surprising power of questions. *Harvard Business Review, 96,* 60–67.

Eastwick, P. W., Finkel, E. J., Mochon, D., & Ariely, D. (2007). Selective versus unselective romantic desire: Not all reciprocity is created equal. *Psychological Science, 18,* 317–319. http://dx.doi.org/10.1111/j.1467-9280.2007.01897.x

Finkel, E. J., & Eastwick, P. W. (2008). Speed-dating. *Current Directions in Psychological Science, 17,* 193–197. http://dx.doi.org/10.1111/j.1467-8721.2008.00573.x

Finkel, E. J., Eastwick, P. W., & Matthews, J. (2007). Speed-dating as an invaluable tool for studying romantic attraction: A methodological primer. *Personal Relationships, 14,* 149–166. http://dx.doi.org/10.1111/j.1475-6811.2006.00146.x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33,* 1–22. http://dx.doi.org/10.18637/jss.v033.i01

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46,* 1251–1271. http://dx.doi.org/10.2307/1913827

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology, 113,* 430–452. http://dx.doi.org/10.1037/pspi0000097

Jauk, E., Neubauer, A. C., Mairunteregger, T., Pemp, S., Sieber, K. P., & Rauthmann, J. F. (2016). How alluring are dark personalities? The Dark Triad and attractiveness in speed dating. *European Journal of Personality, 30,* 125–138. http://dx.doi.org/10.1002/per.2040

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 638–646). Boulder, CO: Association for Computational Linguistics.

Kenny, D. A. (1988). Interpersonal perception: A social relations analysis. *Journal of Social and Personal Relationships, 5,* 247–261. http://dx.doi.org/10.1177/0265407588800500207

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology, 18,* 141–182. http://dx.doi.org/10.1016/S0065-2601(08)60144-6

Kluger, A., & Malloy, T. (2019). Question-asking as a dyadic behavior. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/pspi0000156

Malloy, T. E., & Kenny, D. A. (1986). The Social Relations Model: An integrative method for personality research. *Journal of Personality, 54,* 199–225. http://dx.doi.org/10.1111/j.1467-6494.1986.tb00393.x

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2009). It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Vol. 1, pp. 334–342). Singapore: Association for Computational Linguistics.

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language, 27,* 89–115. http://dx.doi.org/10.1016/j.csl.2012.01.005

Appendix

Detailed Annotation of KM's Coding Errors

How KM Tested the Wrong Outcome Variables

In several of their analyses, KM tested "second date offers given" rather than "second date offers received" as the primary dependent measure. Because their code is poorly documented, we provide our own annotations of KM's code here to explain how this mistake arose multiple times across two separate scripts. Many of KM's analyses are unscripted, so we cannot evaluate all of their evidence. Still, we use their own submitted materials, and nothing more, to show that they base central claims in their article on erroneous analyses. This appendix is quite technical, so we encourage readers to follow along with the original R scripts that were posted on OSF (at https://osf.io/x69u2/).

To start, let's consider our original analysis script, "study3code.R," in our original OSF https://osf.io/k8e84/. On lines 49–67, we specify our regression model as a function that tests how various partner question-asking variables (e.g., "partner-.Q.per.turn," "partner.follow.Q.per.turn") predict "date.again," which represents the rater's preference to go on a second date with said partner. Though dyadic analyses can be confusing, the meaning and interpretation of these variables were unambiguous in our work and correspond to how we reported them in our article.

KM started their analyses with the same dataset, but ended with different results. How? Here we provide two examples of problems in their workflow that led to their errant results. First, consider their script "SRM for binary outcome (Dating) by David Kenny.R," which generates their Table 8. On line 75, they specify that "date.again" is their outcome, and in lines 146–164, they use it as the outcome in the regressions for Table 8. However, in the Table 8 caption, they describe this variable as "*Receiving a Second Date Offer*." As previously described, "date.again" is the number of second dates offered, not received.

Furthermore, it is clear from lines 127–128 that KM did not use the question-asking variable we provided, "partner.Q.follow.per-.turn," but instead used a new variable, "actor.Q.follow.per.turn." In other words, they tested whether question asking has an effect on the asker's own preference to date again. They made the same mistake on lines 203–224, which generates Figure 1 in their article.

The second example is spread across the files "BLOCKO in R.R" (henceforth "BiR") and "PrepareEqaulGroupAndImputed-Data.R" (henceforth, "PEGaID"). The original data file is loaded in BiR on line 7 by PEGaID, which does many things, some of which are relevant here. On lines 17–18 of PEGaID, key variables are renamed: "self_id" becomes "RaterID," and "other_id" becomes "TargetID." Lines 20–28 create a separate dataset to calculate actor question-asking variables, and line 30 merges this into the original data (here, the key outcome "date.again" becomes "date.again.x"). Up to this point, these new measures correctly identify actor-level asking.

However, there is a critical mistake on lines 36–38 of PEGaID: the original partner question-asking variables are deleted, and the new actor question-asking variables are simply given the names of the partner question-asking variables. Later, in lines 254–261 of PEGaID, they take these incorrect variables and repeat the procedure from 20–28 of PEGaID, generating "actor question asking" variables. However, because the source columns were mislabeled, the "actor question asking" variables are also mislabeled: "actor" should be "partner," and vice versa.

This incorrectly labeled dataset is the basis for the analyses in BiR. On line 12 of BiR, they again rename all the "actor.Q" variables, so that the asker is unknown (e.g., "actor.Q.total" becomes "Q.total"), but as we noted above, these should be called "partner.Q." Then, on line 18, they rename some other important variables: "date.again.x" becomes "DatesOffered," and "date.again.y" becomes "DatesReceived."

The second half of this mistake comes on lines 23–30 of BiR, which calculates an average of each column for each "TargetID." That is, the target-level average of "DatesOffered" calculates the average dates offered to each target, while "DatesReceived" calculates the average dates offered by each target. However, the comment on line 21 of BiR (and the way they use the data later) suggest the authors believe they are calculating actor-level effects rather than target-level effects. In light of both this mistake in BiR and the actor/partner flip in PEGaID, the correct interpretation of the question-asking variables should now be flipped twice—their supposed actor-level average of actor question-asking is, in fact, the target-level average of target-level question-asking. While this double flip means the question variables now happen to be correct, the outcomes were only single-flipped and, thus, are wrong.

Lines 32–74 of BiR calculate gender-level and session-level grand means, and subtract them from the ActorEffects data. No errors are made there, though line 42 renames "DatesOffered" to "DatesOffered.x," and "DatesReceived" is renamed to "DatesReceived.x." A key line in this block is line 43, where the variable "ActorQT" is constructed as a composite of question-asking measures—as we have established, this double-flipped measure correctly captures actor question-asking behavior.

Lines 78–89 of BiR generate the top half of KM's Table 7. The top-level row labels are correct, but the column labels are reversed. Consider the third row, which is generated from lines 88–89 of

(*Appendix continues*)

BiR. Here, KM correlate outcomes “DatesOffered.x” and “DatesReceived.x” with “ActorQT” as their question-asking measure. In the article, they refer to “DatesReceived.x” as “Number of Dates Received,” but from lines 23–30 above, we know that it is aggregated at the target level—the average number of dates received by the partners of a given target is, in fact, that target’s rate of *offering* dates. Likewise with the first column: The average number of dates offered to a given target is, in fact, that target’s rate of *receiving* dates.

In KM’s article, the analysis that most closely resembles the main hypothesis in our original article is column 1, row 3 of Table 7. When the columns are correctly labeled, these numbers support our original hypothesis—that question asking is associated with receiving more second-date offers, $r = .27$, $p = .01$.

How KM Misinterpreted Their Validation Checks

KM’s script also provides a narrow glimpse into the validation checks that they may have conducted while writing their article. One natural validation check in these speed-dating data is to test the effect of gender on dates received. In these data, women indisputably received more second-date offers from men than vice versa. In fact, we relied on this check frequently while writing our own code, as did KM in several places (e.g., line 48 of PEGaID). Though, on line 76 of BiR, KM suggest that they performed this exact validation check and found that the effect was the opposite of what they expected. However, rather than recognize their actual mistake, they instead flipped the gender labels in Table 7. Lines 80–81 of BiR average across all women raters (sex = 0) and produce the results for the second line of the table ($r = .37$ and $r = -.21$) but are described in Table 7 as averaging male raters.

When calculating the bottom half of their Table 7, KM encountered a second opportunity to notice their error. Lines 97–103 of BiR make a corresponding error to lines 23–30 of BiR: Although KM believe they are aggregating at the partner level, they are using the “RaterID” variable, which aggregates at the rater level. Lines 128–129 of BiR flip the outcome variables: “DatesOffered.x” is relabeled “DatesReceived.z,” and “DatesReceived.x” is relabeled “DatesOffered.z.” Line 145 suggests that they conducted the gender validation check again but found that it was not reversed. In fact, this check only passed because they had unreversed the outcome variables.

Received April 22, 2019
Revision received August 21, 2019
Accepted August 27, 2019 ■