Evolution and Human Behavior 44 (2023) 144–146

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.elsevier.com/locate/ens

Creative ideas, but elementary mistakes: A reply to Brandner and colleagues to promote best practices in error management theory research

David M.G. Lewis [a,b,*], Laith Al-Shawaf [c,d], Ayten Yesim Semchenko [e], Kortnee C. Evans [a]

[a] Discipline of Psychology, Murdoch University, Australia
[b] Centre for Healthy Ageing, Health Futures Institute, Murdoch University, Australia
[c] Department of Psychology, University of Colorado Colorado Springs, Australia
[d] Lyda Hill Institute for Human Resilience, University of Colorado Colorado Springs, Australia
[e] Faculty of Science, Charles University, Czechia

ARTICLE INFO

Keywords
Error management theory
Signal detection theory
Sensitivity
Bias
Cognitive biases
Sexual overperception bias

ABSTRACT

In their commentary on our paper, Brandner et al. commit an elementary statistical mistake that leads to entirely erroneous conclusions. When this statistical error is corrected, the effects described in our original paper appear exactly as reported. In principle, we could end our reply there. However, ending our reply there would be a lost opportunity for promoting best practices in Error Management Theory (EMT) research. The commentators make several other missteps that present the opportunity to draw attention to important principles in EMT research and offer clarifications that we hope assist in the operationalization, design, and interpretation of EMT-inspired studies in the future. We discuss these points and provide several EMT research scenarios to help to illustrate a key principle. We hope this reply highlights some of the key elements of best practices in EMT research and sheds light on pitfalls that researchers must make sure to avoid.

> *If the interaction term is not statistically significant, it should be removed from the model and the analysis rerun without the interaction term. Failure to remove an interaction term that was not statistically significant also can lead to an incorrect conclusion (*Engqvist, 2005*).*
> - Beck & Bliwise, 2014, p. 371

> *To conclude this brief review, which can be extracted from almost any statistical textbook […] nonsignificant […] interaction terms must be removed before re-running the final analysis.*
> - Engqvist, 2005, pp. 968–969

In their commentary on our paper, Brandner, Brase, and Young (2022) make an elementary statistical mistake that leads them to entirely erroneous conclusions. In their reanalysis, they failed to remove the non-significant interaction before interpreting the lower-order effects. As captured by the epigraphs above, you cannot do this.

That's the crux of it: the commentators' erroneous conclusions are caused entirely by this fundamental statistical error. Appropriately removing the non-significant higher-order interaction leads to the exact results we reported in our original article (Lewis, Al-Shawaf, Semchenko, & Evans, 2022).

Specifically, when the non-significant three-way interaction between participant sex, target attractiveness, and uncertainty is correctly removed, there are two significant two-way interactions: (1) between the effect of uncertainty and participant sex, which reflects the sex-differentiated biases in response to uncertainty that we originally reported, and which were our primary research findings; and (2) between uncertainty and target attractiveness, which was our secondary research finding. In other words, appropriately removing the non-significant higher-order interaction yields precisely the results we reported in our original article. The interested reader is welcome to consult these analyses, including the commentators' analyses followed by the correct analyses, on the Open Science Framework: https://osf.io/dskp4/?view_only=d9e70dabc844416bbc06105d2a901905.

In principle, we could end our reply here.

Ending our reply there, however, would be a lost opportunity to promote best practices in Error Management Theory (EMT) research.

* Corresponding author at: Discipline of Psychology, Murdoch University, 90 South Street, Murdoch WA 6150, Australia.
E-mail address: davidlewis@utexas.edu (D.M.G. Lewis).

https://doi.org/10.1016/j.evolhumbehav.2023.01.009
Received 12 December 2022; Accepted 2 January 2023
Available online 20 January 2023
1090-5138/

Brandner et al. make many other missteps in their commentary. These present an opportunity to draw attention to important principles in EMT research, identify pitfalls that researchers should avoid, and offer clarifications that we hope assist in the operationalization, design, and interpretation of EMT-inspired studies in the future.

1. Principle 1: The *Average* does not equal the *Truth*

Brandner and colleagues encourage EMT research to substitute a *sample average* where an *objective baseline true value* is needed. This recommendation is flawed, and following it would automatically invalidate the entire analytical approach: although not made explicit, the procedure they describe would inherently treat a participant's judgment as *wrong* if it differs from *the average of other people's judgments*. The two scenarios below illustrate the invalidity of this operationalization of error (and any analyses based on it).

Scenario 1: Imagine you are a researcher studying people's perceptions of the height of a vertical surface such as a cliff. You conduct a study in which you take participants to the top of a cliff with a known height: 50 ft. You have each participant estimate the cliff's height. One participant, Taylor, estimates that the cliff is 50 ft tall. The average of all other participants' estimates is 54 ft. Is Taylor's estimate *wrong* – that is, did Taylor underperceive the height of the cliff – because it differed from the *average of other people's perceptions*?

Of course not. It would be preposterous to substitute the average of other people's perceptions for the *objective true state of the world.* However, this is precisely what Brandner and colleagues recommend.

In Scenario 1 (the cliff), there is an *objective true value* that can be known. In such situations, that objective true value must always be the value against which people's judgments are compared to test for an under- or over-perception bias.

However, a great deal of EMT research is interested in people's judgments about phenomena for which *there is no demonstrable objective true value*. Scenario 2 describes an example of this, and illustrates why using a sample average as a stand-in for objective baseline truth is still deeply flawed in such scenarios.

Scenario 2: Imagine you are a researcher studying men's perceptions of the most attractive female waist-to-hip ratio (WHR). Fifty percent of your sample is from the United States, and the other 50% are Hadza men. If your data parallel those from previous research, the *average* preferred WHR will be approximately 0.8; American men rank a WHR of 0.7 as most attractive, whereas Hadza men rank a WHR of 0.9 as most attractive (see Marlowe & Wetsman, 2001; see also test 1 in Sugiyama, 2004). If a given participant prefers a WHR that differs from 0.8, should you consider that participant to have made an inferential "error"?

Of course not.

Moreover, if the proportions of US men and Hadza men in the sample change, then the average value will change. An extreme example of this would be if your sample consisted of 99 US men and one Hadza man. In such a case, if the US men and the Hadza man had preferences reflective of their population means, the average preferred WHR would be 0.702. Should you then conclude that the Hadza man's preference is an "error" because it deviates from the sample average? Again, most certainly not. The difference between an individual participant's judgment and the average of a sample of other people's judgments is a wholly invalid operationalization of "error."

To be clear, the problem here is not that the Hadza man's response is being compared to the average response of people from a different culture. **The issue is that it is not valid to operationalize "error" as the difference between a given individual's response and the average response from other people**, regardless of whether those other people are from the same or a different culture. For example, if the lone Hadza man preferred a WHR of 0.95 and thereby had a preference that differed from the average preference in his population, his preference would still not be an "error."

The key point is that it is deeply misguided to treat a sample average as if it is the value against which "error" can be established and assessed. Any EMT or SDT analyses that are based on such an operationalization will be invalid.

The scenarios elaborated above illustrate why this operationalization is unsound. We can also illustrate the invalidity of this operationalization of "error" using a more quotidian example: if you were given a slice of pizza and asked how delicious it was, would your answer be *wrong* if it differed from the *average of other people's responses*? Of course not. That operationalization of "error" is clearly incorrect and logically ungrounded.

Yet it is at the core of the commentators' recommendation (and it is also at the core of Brandner and colleagues' analyses in their 2021 paper in *EHB*; see Brandner, Pohlman, & Brase, 2021). This issue is important because further propagation of this approach would be detrimental for Error Management Theory research, and indeed for all research on cognitive biases. "Sample average" and "objective baseline true value" are not the same thing, and conflating them can invalidate all analyses and conclusions predicated on their conflation.

2. Principle 2: Distinguish between primary, secondary, and tertiary findings

In their commentary, Brandner et al. assert that "dichotomization produced a Type I error in the original analysis." This is incorrect and misleading in multiple ways.[1] First, it fails to identify the true cause of Brandner and colleagues' discrepant results: their fundamental statistical error, which we described at the beginning of this reply. Second, it misrepresents our analyses—without exception, when we entered target attractiveness as a predictor, we entered it as a continuous variable.[2] Third, it misleads the reader by discussing our *tertiary* analyses as if an alternative set of results at the tertiary stage of analysis could somehow undermine the primary findings or secondary findings.

In our study, we conducted primary, secondary, and tertiary analyses. The primary focus of our paper was to test the hypothesized MOAB and FUAB biases. We established the existence of these biases. These were our primary findings: on average, men overperceived women's attractiveness under conditions of uncertainty, whereas women on average *under*perceived men's attractiveness under uncertainty. Next, the secondary goal was to determine whether the magnitudes of these biases varied as a function of target attractiveness. We established that they did: the effect of uncertainty depended on target attractiveness, as indicated by the significant interactions between target attractiveness and uncertainty observed among both male and female participants.

We then conducted *exploratory tertiary analyses* to probe these interactions. The commentators frame the idea that we could have probed the observed interactions in a different manner as if that somehow calls into question the existence of those interactions, or even the existence of the biases. This is extremely misleading. No matter what analytical approach is used to probe the interactions between target attractiveness and uncertainty, and no matter what conclusions are drawn from those tertiary analyses, they have no bearing on the primary findings of (1) men on average overperceiving women's attractiveness under

[1] There were several other instances in which Brandner et al. (2022) inaccurately describe or misrepresent our original paper, including our operationalizations and analyses. In this reply, we do not belabor all those errors. Rather, we focus on the issues that we see as most important for promoting best practices in EMT research.

[2] In our original paper, there were occasions in which we described or presented attractiveness in a dichotomous manner (e.g., Fig. 2). However, this was for the purpose of visualization and the reader's ease of understanding; although we occasionally talked about attractiveness using terms like "attractive" and "not attractive," there was not a single analysis in which we entered attractiveness as a dichotomous predictor.

conditions of uncertainty (the MOAB), and (2) women on average overperceiving men's attractiveness under conditions of uncertainty (the FUAB). They *also* have no bearing on the secondary finding that the effect of uncertainty depended on target attractiveness. These primary and secondary findings stand alone.

3. Principle 3: Demand characteristics should only be invoked as an alternative explanation when they can actually account for study findings

In their commentary, Brandner et al. (2022) allude to demand characteristics as a possible alternative explanation for our findings. This gives us the opportunity to clarify how researchers can best think about demand characteristics as possible alternative explanations. The most important question is *whether or not the proposed demand characteristic can actually account for the observed findings*. The commentators' proposed demand characteristic cannot account for any of the key study findings, so it falls flat as an alternative explanation.

Brandner and colleagues propose the following demand characteristic: "a second iteration of the same face and question demands a different response." In principle, we agree with this possibility. However, this demand characteristic generates only one prediction: participants will change their response when presented with a target for the second time. This prediction fails to account for the study findings in multiple ways. First, this demand characteristic alternative only proposes that participants will change their responses, not that they will change their responses *systematically in one direction*. Second, the demand characteristic fails to account for the fact that men and women systematically changed their responses *in opposite directions*. Third and finally, nothing about the demand characteristic alternative suggests that these sex-differentiated tendencies *should vary as a function of the target's attractiveness*. In short, the demand characteristic alternative collapses entirely when placed under scrutiny.

The key take-home message is this: researchers should ensure that a possible demand characteristic is actually capable of accounting for the observed findings before invoking it as an alternative explanation.

4. Principle 4: Sensitivity is not a relevant evidentiary criterion for assessing EMT hypotheses

In principle, we are excited by Brandner et al.'s idea to integrate Signal Detection Theory and EMT. The problem is that their proposed integration of SDT with EMT is not conceptually or analytically appropriate. The "more general signal detection theory (SDT) hypothesis" of "evolutionary optimality" that they propose in their commentary is not a competing alternative to an EMT hypothesis. The commentators suggest that their "evolutionary optimality hypothesis" proposes that "in situations of relative certainty, the heuristic should favor accurate judgments." We agree with this idea. However, by definition, it pertains to conditions of certainty. That is not what EMT hypotheses are about; EMT hypotheses are explicitly (and exclusively) concerned with how humans make inferences under *uncertainty*. This means that the idea that inferential mechanisms should be accurate under conditions of certainty is not in competition with our EMT-inspired hypotheses (or *any* EMT hypothesis, for that matter).

Moreover, neither (1) sensitivity nor (2) the relative influence of sensitivity (versus bias) are valid evidentiary criteria for assessing EMT hypotheses. The commentators state: "sensitivity to stimuli was more influential on responses than bias […] To put it succinctly, the uncertainty manipulation was not strong enough to bypass the evolved heuristic of maximizing correctness." These statements do not help the reader think clearly about the relationship between sensitivity and bias. Selection could have favored psychological mechanisms that are highly sensitive and make accurate inferences about the world in the majority of instances, *and* still exhibit biases in line with Error Management Theory. Many psychological mechanisms likely are characterized by *both* of the following: (1) they are usually accurate, and (2) in the small subset of instances in which they make errors, those errors are systematically biased in one direction over the other. To say that a mechanism is characterized by an EMT bias is not to say that it is *often wrong*; it is to say that *when* it is wrong, it is systematically wrong in one direction more than the other (e.g., biased in the direction of the less costly error). That is the core of EMT.

In sum, future research should be clear that sensitivity is not a valid criterion for testing EMT hypotheses. EMT hypotheses do not generate predictions about sensitivity at all; they generate predictions exclusively about *bias* (and not about the relative influence of bias compared to sensitivity). Researchers should therefore not use sensitivity as a criterion for assessing EMT hypotheses.

5. Conclusion

For the sake of evolutionary psychology and cognitive psychology, it is important to encourage best practices in Error Management Theory research. For the reasons outlined above, the analyses proposed and/or conducted by Brandner and colleagues in their (2022) commentary and their (2021) paper in *EHB* do not represent best practices. We hope that our reply here has illuminated some of the foundational errors involved in this approach. Our goal has been to point out and discuss these issues in order to facilitate best practices for future EMT studies. In particular, we emphasize that a sample average does not equal an objective baseline true value: treating an individual participant's judgment as *wrong* because it differs from the average of other people's responses is a nonsensical operationalization of error, and should not be used in future EMT research.

With respect to the sex-differentiated biases that we reported in our original paper, by no means do we purport to have final answers. Future research will be needed to answer additional questions about these biases and the cognitive architecture responsible for producing them. Are the biases specific to over- and underperceptions of physical attractiveness, or do they pertain to other dimensions of mate value as well? What is the information-processing architecture of the algorithms responsible for producing them?

Our emphasis here is that these and other questions must be tackled through research that applies best practices. We hope this reply has illuminated some of the key elements of best practices in EMT research, as well as invalid approaches and pitfalls that researchers must make sure to avoid.

Declaration of Competing Interest

None.

References

Beck, C. W., & Bliwise, N. G. (2014). Interactions are critical. *CBE Life Sciences Education, 13*(3), 371–372. https://doi.org/10.1187/cbe.14-05-0086

Brandner, J. L., Brase, G. L., & Young, M. E. (2022). Size, scale, and design matter: Commentary on Lewis, Al-Shawaf, Semchenko, and Evans, (2022). *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2022.09.007

Brandner, J. L., Pohlman, J., & Brase, G. L. (2021). On hits and being hit on: Error management theory, signal detection theory, and the male sexual overperception bias. *Evolution and Human Behavior, 42*(4), 331–342. https://doi.org/10.1016/j.evolhumbehav.2021.01.002

Engqvist, L. (2005). The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Animal Behaviour, 70*(4), 967–971. https://doi.org/10.1016/j.anbehav.2005.01.016

Lewis, D. M. G., Al-Shawaf, L., Semchenko, A. Y., & Evans, K. C. (2022). Error management theory and biased first impressions: How do people perceive potential mates under conditions of uncertainty? *Evolution and Human Behavior, 43*(2), 87–96. https://doi.org/10.1016/j.evolhumbehav.2021.10.001

Marlowe, F., & Wetsman, A. (2001). Preferred waist-to-hip ratio and ecology. *Personality and Individual Differences, 30*(3), 481–489. https://doi.org/10.1016/S0191-8869(00)00039-8

Sugiyama, L. S. (2004). Is beauty in the context-sensitive adaptations of the beholder? Shiwiar use of waist-to-hip ratio in assessments of female mate value. *Evolution and Human Behavior, 25*(1), 51–62. https://doi.org/10.1016/S1090-5138(03)00083-7