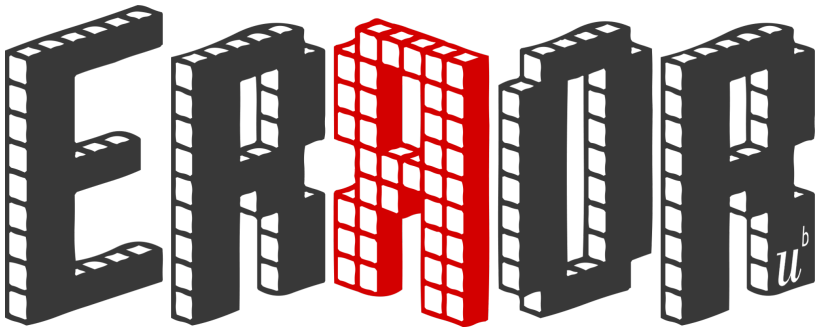


ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

ERROR REPORT

Lades, L. K., Laffan, K., Daly, M., & Delaney, L. (2020). Daily emotional well-being during the COVID-19 pandemic. *British Journal of Health Psychology, 25*(4), 902-911. https://doi.org/10.1111/bjhp.12450

DECISION:

Moderate errors that do not affect the core conclusions

Reviewer: **Joop Adema**, Ludwig Maximilian University of Munich
Author response: **Leonhard Lades**, University of Stirling
Recommender: **Ian Hussey**, University of Bern

Cite as: ERROR (2024) Error review of Lades et al. (2020). Version 1. https://doi.org/10.31234/osf.io/e3c7p

All ERROR reports can be found at error.reviews & osf.io/fpw4r

DECISION & ЯECOMMENDATION

Lades, Laffan, Daly, & Delaney (2020) "Daily emotional well-being during the COVID-19 pandemic" was determined to contains **Moderate Errors that do not affect the core conclusions**. That is, errors that have the benefit of being detectable thanks to the presence and sharing of research materials, whose scope is moderate given that a large proportion of the reported results are affected (likely to a small degree), but which do not affect the article's core conclusions. A determination of Moderate Errors involves the recommendation that the authors **seek a correction**.

Following the ERROR project's emergent guidelines, the recommendations associated with a Moderate Errors decision are as follows:

- The report, author response, and recommendation have been posted on the ERROR website (error.reviews/reviews/lades-2020) and as a preprint on PsyArXiv (osf.io/preprints/psyarxiv/e3c7p). Their associated materials have been posted to OSF (see osf.io/sdz5y).
- The authors are asked to recognize these errors in future discussions of the article.
- The authors are recommended to pursue a correction notice with the journal. Specifically, with regard to:
    - The text on p. 905 states that multiple testing corrections were used. However, corrections were in fact only applied to results in the supplementary materials and not any results reported in the article itself. 11 of 42 (26.2%) statistically significant results reported in the article were not significant after multiple testing corrections. Although, these do not alter the main conclusions. The use of multiple testing corrections should either (a) more clearly state that these corrections were applied only to the results reported in the supplementary materials and not any results reported in the article, or (b) all results reported in text should be those with the multiple testing corrections applied.
    - The primary analyses reported in the article did not acknowledge dependencies in the data (i.e., each participant completed the survey multiple times). The authors agreed with the reviewer that they should have employed clustered Standard Errors to acknowledge these dependencies. Applying clustered Standard Errors changes the statistical significance of one result (positive affect ~ schooling children), although this does not change the core conclusions. Corrections should be made to the Standard Errors reported in text, the Confidence Intervals in all panels of Figure 1 and Figure S1, and the results reported in Table S2. The "†"

significance bar should be removed from the positive affect ~ schooling children result in Table S2.

- All analyses and summary statistics should use corrected data that use unique participant IDs. A small number of duplicate participant IDs were found by the reviewer, which the authors argue are distinct participants but are not treated as such in the reported analysis. Correcting this error may produce additional changes in the results of the regressions reported in the text, tables, and plots. These magnitude of any such changes is currently unknown but likely to be very small.

RECOMMENDER'S ЯEPORT

Ian Hussey

I'd like to thank Joop Adema (reviewer) and Prof Leonhard Lades, Prof Kate Laffan, Dr Michael Daly, and Prof Liam Delaney (authors) for participating in this error review. The article, Lades et al. (2020), studied daily emotional well-being during the COVID-19 lockdowns, a subject that was of great interest and concern to both scientists and the public during that time. As such, this work was conducted under more difficult data collection and time pressure conditions than usual. Lades et al. (2020) is highly cited – just under 600 times at time of writing, and is among each of the authors' most highly cited articles. These factors made Lades et al. (2020) an excellent candidate for error review: research that is important in terms of the societal relevance of its substantive conclusions, but also in terms of its citations and therefore influence on subsequent work.

The substance and style of the reviewer report and authors' response are an example of what we hope to see more of in academic research: acceptance of the possibility that errors occur; inspection and useful discourse about potential errors that is well-documented and verifiable; and acknowledgement and suitable correction when errors are found. I would like to applaud the authors and reviewer for their productive exchange – for example where the authors note in their response that "Joop had additional excellent comments that he did not label 'error' but that will influence the way we conduct future analyses."

ERROR recommendations are public documents whose function is to (1) communicate the presence or absence of any errors detected, (2) consider their severity, and (3) provide discussion of how similar errors elsewhere might be prevented or detected. Materials for all error reports can be found at https://osf.io/fpw4r/.

Summary of errors detected & how they could be prevented in future

The original article assessed daily activities and affect on a day during the COVID-19 lockdowns, examined associations between activities and affect, and made recommendations for activities to improve affect.

I asked the reviewer to focus his review on implementation of the analysis and reporting of results, as well as the causal assumptions and claims.

The issues and errors raised relate to: First, the assumption of independence. Participants were assessed multiple times in one day, and the non-independence of these data points is important to acknowledge in the statistical model. While the authors did present plots in their supplementary materials of the results of random effects model that likely did acknowledge these dependencies, the primary results of the fixed effects models reported in the text and used for the conclusions did not. The authors agreed that the clustered standard errors should have been employed. Future research could avoid this error by ensuring that dependencies in longitudinal data are acknowledged in the statistical model, such as by using clustered standard errors or (perhaps preferably) employing random effects models for the primary analyses.

Second, the alignment of the results reported analyses and the reported results, and the potential for lack of clarity when these do not clearly and fully align. The original article states that "To manage the risk of finding false associations in our multiple testing approach, we used the Benjamini Hochberg method, identifying significant associations at a false discovery rate of 0.05 (see Supplementary Materials)." (Lades et al., 2020, p. 905). To the authors' credit, this statement is literally true; their supplementary materials do provide additional information; and the major conclusions of the article are not changed by applying these corrections vs. not. However, there is also a risk that it gives the impression to the reader that the results reported in the article employ these corrections, when they were employed only in the supplementary materials, and how the impact of employing them vs. not is not discussed in the article (this review found that roughly one-in-four statistically significant results did not remain significant after multiple testing corrections). Future research could avoid this issue by being explicit and precise about what statistical approaches are applied to what results, in order to prevent confusion or misinterpretation for readers. Additionally, future research could enhance the utility of robustness tests by being explicit in the main article about how results change between analytic choices.

Discussion of individual issues raised

Statistical model should have acknowledged dependencies in the data

The reviewer raised the point that multiple observations were collected from each individual but these dependencies among the data were not modelled (e.g., using clustered standard errors in the fixed effects model). The reviewer's adjusted model changed one result from significant to non-significant, changing the conclusion for the influence of 'Schooling Children' on positive affect. The authors agreed with this principle that the results reported in the original article should have acknowledged this non-independence in the data.

The reviewer notes the use of clustered standard errors increases the standard errors for many or most of the activities (e.g., typically by .01 from what I can see). While this has a relatively minor impact on the significance of the effects (1 of 92 effects, 1.1%), they nonetheless have a broad impact on the reported estimates (13 standard errors reported on pages 905-906 may require correction, as well as all panels of Figure 1, Figure S2, and Table S2). I would therefore characterize this a moderate error, not on the basis of their individual severity but because of collectively they change a large number of results reported in the article.

Multiple testing adjustments

Separately, an exchange between the reviewer and authors highlights an ambiguity in the original article. The original article states "To manage the risk of finding

false associations in our multiple testing approach, we used the Benjamini Hochberg method, identifying significant associations at a false discovery rate of 0.05 (see Supplementary Materials)." (Lades et al., 2020, p. 905). The supplementary materials do indeed indicate which effects are no longer significant after multiple testing corrections. However, the main article does not explicate that these multiple testing corrections are applied only in the results presented in the supplementary materials, nor is any summary of the impact of these corrections provided. Of the 92 activity-affect associations examined, 42 (45.7%) were reported to be statistically significant without corrections, whereas 31 (33.7%) were significant after Benjamini Hochberg corrections. That is, roughly in four (26.2%) of the statistically significant results did not survive multiple testing corrections.

At the same time, it is worth caveating this by acknowledging that most of the affected effects are not central to the verbal conclusions in the discussion. I also recognize that the authors report that this information about how these corrections change the results was provided to the prepublication peer reviewers. However, in aggregate, the original article's text states the risk of false positives was managed when it in effect was not in a substantive or explicit way. In effect most readers would see only the uncorrected estimates and p values. This represents a moderate error.

In correcting the results in light of the above, I invite the authors to more fully discuss their multiple testing corrections method, including making explicit what the family of tests is, and to align the results reported in text with the corrections method.

Duplicate data

The reviewer raised the possibility that 6 of the 604 participants in the analytic sample (roughly 1%) may have had duplicate data in the dataset as they had duplicate participant codes. The author note that these duplicate IDs do have different demographic data and therefore likely represent distinct participants. Nonetheless, by sharing unique identifiers, this will distort both the clustered standard errors in fixed effects analyses and the standard errors for random effects analyses. I agree with the authors that the impact of this error is likely to be quite small, but it does nonetheless represent a minor error.

Causal assumptions and language

The reviewer made some minor comments about causality and causal language that I think are useful for readers to consider. The reviewer is clear that, especially in the context of a piece of work that was produced under time pressure during the COVID-19 pandemic, he finds no error in the article's approach to causality. While the use of causal language and synonyms is a matter of long debate in the literature, I am personally swayed by the argument that causal assumptions (in terms of modelling, theory, or policy

recommendations) are useful to explicate even (or especially) when the research cannot provide strong evidence of causality (e.g., Hernán, 2018: 10.2105/AJPH.2018.304337). My reading is that the Lades et al. (2020) is interested in the causal impact of activities on well-being and not their mere association. For example, the article states: "These findings highlight activities that may play a protective role in relation to well-being during the pandemic" (Lades et al., 2020, p. 902); and "The current study also highlights the value … to inform actions that may promote well-being and enhance the sustainability of self-isolation measures." (Lades et al., 2020, p. 909). I.e., if it were the case that affect influences smartphone screen time the article's recommendations that screen time be limited would not hold. It may have been useful to state in the limitations sections that the substantive conclusions rely on a variety of causal assumptions, e.g., the direction of causality, there being no within-subject/time-varying confounders, etc. Of course, a variety of opinions exist on the use of causal language vs. the explication of causal assumptions, and the purpose of ERROR reviews is not to dictate a single answer to this matter of on-going debate. As such, this does **not** represent an error in the original article, but it is worthy of consideration given that these causal assumptions influence or provide important context the conclusions and recommendations.

Unresolved issues

It is reasonable to expect that ERROR reviews will leave some questions unresolved. It is useful to acknowledge the potential for such issues so that ERROR recommendations do not artificially convey that they represent the final word on issues of error detection and correction for a given article.

In this case, the precise impact of correcting the analysis to include clustered standard errors is likely to increase a yet unknown number of the standard errors reported in the article and its supplementary materials. No other issues appear to be unresolved.

I sincerely thank the reviewer and especially the authors again for their openness to scrutiny and their error-acceptance.

Ian Hussey
Chief Recommender for ERЯOR

ЯEVIEW

Joop Adema

I. METHODS, MEASUREMENT, AND DESIGN

1. Design No errors found

Are there errors in the conceptual design of the study? E.g., flawed randomisation technique

I do not have concerns about the design. The approach builds on Kahneman et al., (2004), with some minor differences and different items relevant to the Covid-19 pandemic. The individual FE approach alleviates some concern of people selecting into activity on traits that are related to levels of positive/negative affect. A limitation of using a single day is that if the weather was very good on the day of study, this could have inflated the positive affect for outdoor activities. Using multiple days could have averaged out that error.

Time spent: 1.5 hours

2. Measurement No errors found

Are there any measures, techniques, or devices that were incorrectly applied or inappropriate for the specific task described in the paper?

The design of the DRM is carefully executed. The empirical analysis has been executed as described in the paper. The methods used are appropriate and intuitive.

Time spent: 1 hour

3. Preregistration Consistency Not applicable

The study was not pre-registered, which is understandable given the timeliness of the article during the early days of the Covid-19 pandemic.

Time spent: 0 minutes

4. Sampling No errors found

Is there an error in the sampling strategy? Is the power analysis reproducible? Does the model used for the power analysis match the model in the substantive analyses? Were separate power analyses conducted for all primary analyses?

Based on my assessment of the code, the sampling strategy seems fine. In total 1494 individuals started the survey, 604 finished and are used for subsequent analysis. I think the number of starting individuals could have been reported in the main text explicitly.

Time spent: 30 minutes

5. Other Aspects Related to Methods and Measures Not applicable

Time spent: 0 minutes

II. DATA, CODE, AND STATISTICAL ANALYSES

1. Code Functionality No errors found

Does the provided code run without the need to make any adjustments and without errors? If not, what steps were needed to get it to run (if it was eventually possible)?

File locations had to be modified in more than one place. Name of .csv file with demographics had a slightly different name. The cleaning code creates the dataset for analysis without further problems.

Time spent: 1 hour

2. Computational Reproducibility of Reported Statistics No errors found

Is there a clear traceability of reported stats to code? Does the code output match what's reported in the paper? Are all reported statistics findable within the analysis code?

Figure 1 and Table 1's results and in-text descriptive statistics can be fully traced back in the code and output.

Time spent: 30 minutes

3. Data Processing Errors Indeterminable

Are there substantive errors during the preparation or cleaning of data (e.g. duplication of rows during a merge) prior to substantive analyses and hypothesis tests?

In a few places the code could be documented a bit better, e.g. when there are some ad-hoc modifications in three duplicate participant identifiers. Three individual identifiers are duplicate, which are subsequently changed. However, the duplicates have exactly the same covariate values: are these truly unique observations? I have no good way of

judging whether this is the case, but it seems that those might have arisen by mistake. The authors do this in their cleaning code:

```
quietly by ParticipantID:  gen dup2 = cond(_N==1,0,_n)
replace ParticipantID = 11187 if ParticipantID == 1187 & dup == 1
replace ParticipantID = 111187 if ParticipantID == 1187 & dup == 2
replace ParticipantID = 12174 if ParticipantID == 2174 & dup == 1
replace ParticipantID = 112174 if ParticipantID == 2174 & dup == 2
replace ParticipantID = 12429 if ParticipantID == 2429 & dup == 1
replace ParticipantID = 112429 if ParticipantID == 2429 & dup == 2
```

It could be that those accidentally got assigned the same identifier by the survey program because of sharing demographic characteristics. This concerns only three observations and has limited influence on the results (see log file). The observations in the final data:

	Partic~D	sex	age	region	county
104.	12429	Female	43	Dublin	Dublin
185.	112429	Female	43	Dublin	Dublin
547.	12174	Female	42	Munster	Cork
705.	112174	Female	42	Munster	Cork
1267.	111187	Female	64	ROL	Westmeath
1566.	11187	Female	64	ROL	Westmeath

Although I do not think this constitutes an error, a short clarification by the authors would be helpful to confirm that this is so.

Time spent: 1 hour

4. Model Misspecification Errors found

Are there any consequential issues with the assumptions or the form of a statistical model (e.g., overfitting, wrong distribution assumption) used to describe data?

A simple FE regression as reported in the paper fits the purpose. However, I have a comment regarding how the model should be interpreted (see section III.1). Another comment relates to the estimated standard errors. As the unit of observation is the episode level, there are multiple observations per individual. The estimations do not make any standard error corrections based on this, which I deem incorrect. First of all, the presence of heteroskedasticity generally overstates precision. Secondly, two episodes of the same person are not statistically independent observations if the aim is to

generalize the study's findings to the population. Hence, I instead cluster on the respondent level by using vce(cluster i.id) in (see the attached .do-file). For panel (a) of Figure I, this change renders 1 point estimate insignificant at the 5% level for the positive affect outcome (the coefficient for positive affect on activity 14 "schooling children", from p=0.045 to p=0.088).

```
.     xtreg zposaff activity_1E activity_2E activity_3E activity_4E activity_5E activit
> E activity_7E activity_8E activity_9E activity_10E activity_11E activity_12E activity
> E activity_14E activity_15E activity_16E activity_17E activity_18E activity_19E activ
> _20E activity_21E activity_22E activity_23E activity_24E activity_25E , fe vce(cluste
> d)
Fixed-effects (within) regression               Number of obs      =      2,795
Group variable: id                              Number of groups   =        604
R-squared:                                      Obs per group:
     Within  = 0.1392                                         min =          1
     Between = 0.0721                                         avg =        4.6
     Overall = 0.0938                                         max =          5
                                                F(25,603)          =      13.55
corr(u_i, Xb) = 0.0427                          Prob > F           =     0.0000
                                  (Std. err. adjusted for 604 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     zposaff | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 activity_1E |   .0644427   .0892007     0.72   0.470    -.1107391    .2396245
 activity_2E |  -.2636997   .0574432    -4.59   0.000    -.3765126   -.1508867
 activity_3E |  -.2260305   .0791084    -2.86   0.004     -.381392    -.070669
 activity_4E |   .0697021   .0522935     1.33   0.183    -.0329973    .1724016
 activity_5E |  -.2850966    .049701    -5.74   0.000    -.3827048   -.1874884
 activity_6E |   .0059284   .0412242     0.14   0.886     -.075032    .0868889
 activity_7E |    .130559   .0375584     3.48   0.001     .0567979    .2043201
 activity_8E |  -.0284077   .0561513    -0.51   0.613    -.1386836    .0818682
 activity_9E |   .1096412   .0780238     1.41   0.160    -.0435901    .2628726
activity_10E |    .107059   .1654796     0.65   0.518    -.2179274    .4320454
activity_11E |  -.0028682   .0408611    -0.07   0.944    -.0831156    .0773792
activity_12E |   .1453435   .0567575     2.56   0.011     .0338771     .25681
activity_13E |   .2075181   .0576087     3.60   0.000      .09438    .3206561
activity_14E |  -.1660266   .0972319    -1.71   0.088    -.3569809    .0249276
activity_15E |  -.0484102   .0417767    -1.16   0.247    -.1304556    .0336352
activity_16E |  -.1529218   .0488546    -3.13   0.002    -.2488676   -.0569759
activity_17E |  -.0831891   .0595831    -1.40   0.163    -.2002047    .0338265
activity_18E |   .0312307   .0479653     0.65   0.515    -.0629686     .12543
activity_19E |   .3326839   .0581064     5.73   0.000     .2185684    .4467993
activity_20E |   .4613417   .0720425     6.40   0.000      .319857    .6028264
activity_21E |  -.3438801   .0721986    -4.76   0.000    -.4856713   -.2020889
activity_22E |    .225055   .0831367     2.71   0.007     .0617823    .3883276
activity_23E |   .2923236   .1034707     2.83   0.005     .0891168    .4955305
activity_24E |   -.118815   .1432598    -0.83   0.407    -.4001637    .1625337
activity_25E |   .0541849   .0487641     1.11   0.267    -.0415832     .149953
       _cons |  -.0012622   .0298024    -0.04   0.966    -.0597914    .0572669
-------------+----------------------------------------------------------------
     sigma_u |  .77873264
     sigma_e |  .63318961
         rho |  .60199767   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Time spent: 30 minutes

5. Erroneous/Impossible/Inconsistent Statistical Reporting Errors found

Are there inconsistencies between test statistics, degrees of freedom, and p-values? Are there implausible degrees of freedom between compared SEM models? Are there point estimates outside the confidence interval bounds?

The paper suggests it performs the FDR-approach by Benjamini-Hochberg (1995). However, it is not discussed in the results section of the paper. Moreover, these corrections are not actually implemented in the code. Hence, it seems that this procedure is not executed. If I try to do it myself, on the "positive affect" outcome in Figure 1A, coming to the conclusion one initially significant coefficient (14E) should be rejected. This is the coefficient on schooling children, so this is important for the conclusions of the paper. In the screenshot below, I order all activities on p-value and calculate the Benjamini-Hochberg condition bottom-up. In text, the screenshot spells out the condition for activity 14. This suggests that all hypothesis tests related to activities from 14 upwards should not be rejected.

zposaff	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
activity_11E	-.0028682	.0424219	-0.07	0.9461	-.0860601	.0803237
activity_6E	.0059284	.0432975	0.14	0.8911	-.0789805	.0908373
activity_8E	-.0284077	.0522646	-0.54	0.5868	-.1309018	.0740864
activity_18E	.0312307	.048219	0.65	0.5173	-.0633297	.1257911
activity_1E	.0644427	.0903101	0.71	0.4756	-.1126607	.2415462
activity_10E	.107059	.1262581	0.85	0.3966	-.1405408	.3546588
activity_24E	-.118815	.1250733	-0.95	0.3422	-.3640913	.1264613
activity_25E	.0541849	.0531148	1.02	0.3078	-.0499764	.1583462
activity_9E	.1096412	.0985909	1.11	0.2662	-.0837015	.302984
activity_15E	-.0484102	.0409233	-1.18	0.2370	-.1286632	.0318427
activity_4E	.0697021	.0532971	1.31	0.1911	-.0348166	.1742209
activity_17E	-.0831891	.0569442	-1.46	0.1442	-.1948601	.0284819
activity_14E	-.1660266	.0827889	-2.01	0.0450	-.3283805	-.0036727
activity_12E	.1453435	.0612772	2.37	0.0178	.0251752	.2655118
activity_22E	.225055	.0882163	2.55	0.0108	.0520575	.3980525
activity_16E	-.1529218	.0512196	-2.99	0.0029	-.2533665	-.052477
activity_7E	.130559	.0392073	3.33	0.0009	.053671	.2074469
activity_23E	.2923236	.0866137	3.38	0.0008	.1224691	.4621782
activity_3E	-.2260305	.0664486	-3.40	0.0007	-.3563402	-.0957208
activity_13E	.2075181	.0567703	3.66	0.0003	.0961881	.318848
activity_5E	-.2850966	.0499376	-5.71	0.0000	-.3830272	-.187166
activity_21E	-.3438801	.0698313	-4.92	0.0000	-.4808234	-.2069368
activity_2E	-.2636997	.0505982	-5.21	0.0000	-.3629258	-.1644735
activity_19E	.3326839	.0599506	5.55	0.0000	.2151172	.4502506
activity_20E	.4613417	.0670665	6.88	0.0000	.3298202	.5928632
_cons	-.0012622	.0303765	-0.04	0.9669	-.0608324	.058308
sigma_u	.77873264					

For Activity 14E, p=0.045>(i/m)q*=(13/24)*0.05= 0.027, so that hypotheses should be rejected, despite p<0.05

For a discussion on the procedure and the calculation in the screenshot above, I refer the reader to the original paper, available here: https://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini%20and%20Y%20FDR.pdf

Time spent: 2 hours

6. Other Aspects Related to Data or Code No errors found

Activities are not labelled in the dataset, making checking the code and analysis and relating the results back to those reported in the article somewhat more error prone.

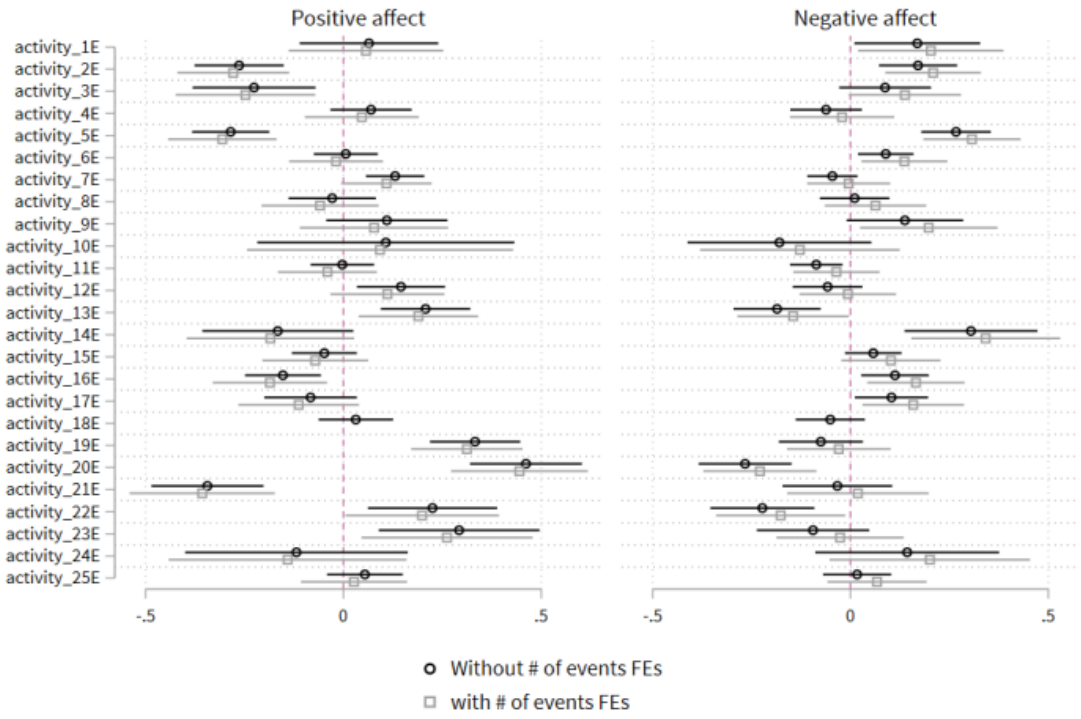
Time spent: 30 minutes

III. CLAIMS, PRESENTATION, AND INTERPRETATION

1. Interpretation Issues No errors found

Throughout the entire paper, is there an incorrect substantive interpretation of data or statistical tests, causal inference issues, etc.?

My main concern about interpretation of the results concerns the following: the paper estimates a linear probability model regressing affect scores on dummies for whether an activity has been pursued during an episode (a partition of one's day). Hence, the estimates conflate the effect of doing a specific activity and the marginal effect of doing one more activity during an episode. This is not acknowledged by the paper, and I think this should be made clear when interpreting the results. It seems more natural to estimate the effect on positive/negative affect of a given an activity vis-a-vis another. Hence, as an alternative, one activity can be chosen as the reference category, after fully saturating the model using dummies for the number of activities. Using resting/relaxing as a reference category, I still find comparable results, but the interpretation is clearer. In the modified coefficient plot of figure 1a, the coefficient for sleeping (activity 18) is not shown, but is not a mutually exclusive category, so it can't be interpreted as a pure reference category:



A similar critique holds for panels c and d of Figure I. As this is usual in the literature, I do not flag this as an error, but think it deserves to be mentioned.

Time spent: 1.5 hours

2. Overclaiming Generalisability No errors found

Does the paper overclaim the generalisability of the findings with regards to stimuli, situations, populations, etc.? Is there hyping or overselling of the importance or relevance of findings?

Altogether, the paper well discusses the limitations of the DRM approach and that it concerns one single day. However, the title references to "Daily", which at least gave me the impression that the task was performed for at least several days, every day. This could have led to a design where daily idiosyncrasies could be cancelled out, such as weather patterns affecting the experience of outdoor activities. Moreover, in the statement of contribution it states "previous studies…and the dynamics of daily experience have been neglected". After reading this statement, I expected that the study would e.g. look at whether the positive affect in episode t spilled over to episode t+1 (the dynamics). However, this was not the case.

Time spent: 1 hours

3. Citation Accuracy No errors found

Are there misrepresentations of substantive claims by cited sources? Inaccurate direct quotes? Incorrectly cited or interpreted estimates? Citations of retracted papers?

I have not found any considerable mistakes in the citations. White and Dolan (2009) should maybe have been credited for already including an "outdoor" item in a DRM in earlier work.
Benjamini-Hochberg (1995) should probably have been cited, as the method is said to be used (see section II.5).

Time spent: 1 hour

4. Other Aspects Related to Interpretation No errors found

Of course, this study does not fully answer the causal question whether those activities cause different levels of affect. Although it does not establish causality, the paper contributes to a pressing policy question during a crisis. The authors carefully deal with this, I believe sufficiently so that the paper is not wrongly interpreted as causal. Nevertheless, through inclusion of individual fixed effects the paper does tackle one potential endogeneity concern: higher affect individuals are more likely to take up certain activities, such as outdoor ones. An implicit "side effect" of this is that only instantaneous within-individual correlations between activities and affect in the same episode are accounted for. Hence, through introduction of fixed effects spillover effects across episodes (anticipatory or delayed) on affect are not accounted for. Moreover, a remaining reverse causality-concern could be still present in the fixed-effects design: during higher affect episodes individuals are more likely to take up certain activities.

The paper focuses on the affect by activity, but not on the shift in activities that took place during the Covid-19 pandemic. Although it is clear homeschooling increased, for other activities it is less clear. For example, curfews have a negative effect on outdoor activities, but WFH may have increased those. A discussion on this would have been welcome.

Time spent: 30 minutes

AUTHOR ЯESPONSE

Leonhard Lades

We would like to thank Joop Adema for reviewing our paper very thoroughly. We would also like to thank the ERROR Team and their funders for this very valuable initiative. Thank you!

Here is our very short summary of Joop's three critical comments:

1. There might be three duplicates in the sense that six observations might come from three participants.
2. We should have clustered the standard errors in our regressions on the respondent level (by using the Stata command "vce(cluster i.id)").
3. We did not properly communicate in the paper how we used the Benjamini–Hochberg (1995) adjustments and we made an error as well.

We respond to these three comments below. However, we would like to highlight that Joop had additional excellent comments that he did not label "error" but that will influence the way we conduct future analyses.

The first instance where Joop did not select "No errors found" relates to three potentially duplicate participants (leading to six observations). These three participants had the same original participant ID and the same basic demographics (gender, age, social class [ABC1F50+,C2DEF50-], and county of residence). We did not elicit these basic demographics ourselves but relied on the survey provider's database which had records of these variables. We did elicit other stable variables (e.g. education or occupation) and the participants with the same original participant IDs do differ in these self-reported variables. Hence, we believe it was the correct decision to assume that these are 6 different survey participants and that the same Participant ID was assigned to some of them by mistake as Joop conjectured. In the end, as Joop clarifies, this has limited influence on the results of the paper.

The second critical comment suggests that the standard errors in our estimates are too small because we did not correct for heteroskedasticity and the fact that we have multiple observations coming from the same person. Joop suggests that we should have used standard errors clustered on the respondent level (using the Stata command "vce(cluster i.id)"). We agree. We appreciate Joop carrying out and presenting the analysis with the clustered standard errors and are glad to see that the existing results hold in all but one case (further discussed below).

Joop's third critical comment refers to the Benjamini–Hochberg (1995) adjustments and that it is not obvious that we made these adjustments. We have two responses. First, we included Benjamini–Hochberg adjustments and included this in the supplementary

tables to the original paper. The Supplementary Information published on the journal website does include Table S2 in which we indicate which coefficients stay significant at $p<0.05$ after Benjamini-Hochberg adjustment. The Supplementary Information also includes a short paragraph describing the Benjamini–Hochberg (1995) method. However, our second response is that we made a mistake and Joop found it (which we are grateful for). Indeed, the coefficient for the association between positive affect and schooling children should not include the †-sign indicating that it is still significant after the adjustments at the 0.05 level. Below we added a table which highlights in red background the coefficient indicating the association between schooling children and positive affect.

Joop's second and third critical comments show that in our data, the positive affect when "schooling_children" is not significantly lower compared to the average positive affect. However, the very strong positive association between "schooling_children" and negative effect is still one of the most noteworthy findings of the paper. Re-reading the paper, we do not think that any of the conclusions we draw throughout the paper regarding schooling children are thus invalidated and are happy to see our other findings hold.

Thanks again to everybody involved and especially Joop Adema.

Leonhard, Kate, Michael, and Liam

[Explanatory note by Recommender] This table shows in red highlight the coefficient that was reported as significant in the original article but is no longer significant when clustered standard errors are used in the model (which the original authors now agree they should have employed) or when Benjamini-Hochberg adjustments are applied (which the authors reported employing in the original article; i.e., the † was included in error). It also shows in red color the coefficients that were reported as significant in the original article but which lose significance when Benjamini-Hochberg adjustments are applied. Note that this information was present in the original article's supplementary materials (i.e., these effects correctly do not have a †); the coloring of the text serves to more clearly highlight the number of affected results. A still unknown number of the Standard Errors reported in the below table may be erroneous (highlighted in yellow highlight), as the appropriate use of clustered standard errors will increase their width. This table, without highlighted and colored text, was presented in the original article's supplementary materials (link).

Table S2: Separate within-person estimates of the relationship between: (A) activities, (B) locations, (C) personal interactions and (D) remote interactions and affect levels.

	Positive affect[a]	Negative affect[b]		Positive affect[a]	Negative affect[b]
	b (SE)	b (SE)		b (SE)	b (SE)
Activities			**Locations (Base = At home)**		
Exercising	0.46***† (0.07)	-0.27***† (0.06)	Outdoors / nature	0.59***† (0.05)	-0.25***† (0.05)
Going for a walk	0.33***† (0.06)	-0.07 (0.05)	Other	0.05 (0.08)	0.04 (0.07)
Gardening	0.29***† (0.09)	-0.09 (0.07)	At other people's home	0.03 (0.15)	0.01 (0.12)
Pursuing a hobby	0.23**† (0.09)	-0.22***† (0.08)	At work	-0.16** (0.07)	0.24***† (0.06)
Taking care of children	0.21***† (0.06)	-0.19***† (0.05)	At a shop	-0.2**† (0.08)	0.09 (0.07)
Socialising	0.15**† (0.06)	-0.06 (0.05)			
Eating	0.13***† (0.04)	-0.05 (0.03)	**In-person interactions**		
Drinking alcoholic beverages	0.11 (0.1)	0.14 (0.08)	Friends	0.34***† (0.1)	-0.14* (0.08)
Pray/worship/meditate	0.11 (0.13)	-0.18* (0.11)	Pets	0.22***† (0.08)	-0.06 (0.07)
Internet	0.07 (0.05)	-0.06 (0.05)	My children	0.13** (0.05)	-0.04 (0.05)
Commuting to work	0.06 (0.09)	0.17** (0.08)	Parents/relatives	-0.03 (0.08)	0 (0.07)
Other	0.05 (0.05)	0.02 (0.05)	Nobody	-0.14** (0.06)	0.11** (0.05)
Resting/relaxing	0.03 (0.05)	-0.05 (0.04)	Other	-0.16** (0.07)	0.05 (0.06)
Doing housework	0.01 (0.04)	0.09** (0.04)	Spouse / significant other	-0.17***† (0.05)	0.09** (0.04)
Preparing food	0 (0.04)	-0.09** (0.04)	Work interactions	-0.41***† (0.08)	0.3***† (0.07)
Drinking	-0.03 (0.05)	0.01 (0.04)			
Watching TV, Netflix or similar	-0.05 (0.04)	0.06 (0.04)	**Over-distance interactions**		
Listening to the radio	-0.08 (0.06)	0.1** (0.05)	Pets	0.2 (0.16)	-0.1 (0.13)
Doing nothing	-0.12 (0.13)	0.14 (0.11)	Other	0.08 (0.08)	-0.13* (0.07)
Using social media	-0.15***† (0.05)	0.11**† (0.04)	My children	0.06 (0.07)	0.08 (0.05)
Schooling children	-0.17**† (0.08)	0.3***† (0.07)	Nobody	0.04 (0.06)	-0.1** (0.05)

Shopping	-0.23***† (0.07)	0.09 (0.06)	Parents/relatives	0.03 (0.06)	0.06 (0.05)
Working/studying	-0.26***† (0.05)	0.17***† (0.04)	Friends	0 (0.06)	-0.01 (0.05)
Informing myself abt. Covid-19	-0.29***† (0.05)	0.27***† (0.04)	Spouse / significant other	-0.08 (0.06)	-0.01 (0.05)
Sleeping	-0.34***† (0.07)	-0.03 (0.06)	Work interactions	-0.27***† (0.06)	0.26***† (0.05)

Standardized relationships are presented. *** p<0.01, ** p<0.05, * p<0.1, † p<0.05 after Benjamini-Hochberg adjustment. All models are based on 2795 observations from 604 individuals.

[a] Positive affect is estimated as the average of the calm and happy affect items.

[b] Negative affect is estimated as the average of overwhelmed, sad, bored, frustrated, lonely, and worried affect items.