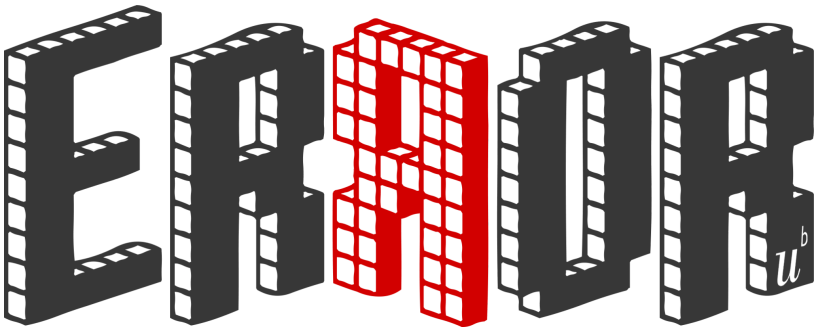


ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

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

reviewed by

Joop Adema, Ifo Institute

Jun 23, 2024

*For the sections below, indicate whether you discovered any errors using the dropdown menu. Describe the **errors** you discovered, the **methods** that you used to find them, and the **amount of time** you invested in the search. Refer to specific files to allow verification of your review. For the assessment below, make sure to check if authors have provided **supplementary analyses** as these may clear concerns arising from the (interpretation of) the primary analyses. If you have written code yourself for the review, please attach it to the report. Please indicate the version of software/packages you used to run the original code and/or your own code.*

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.

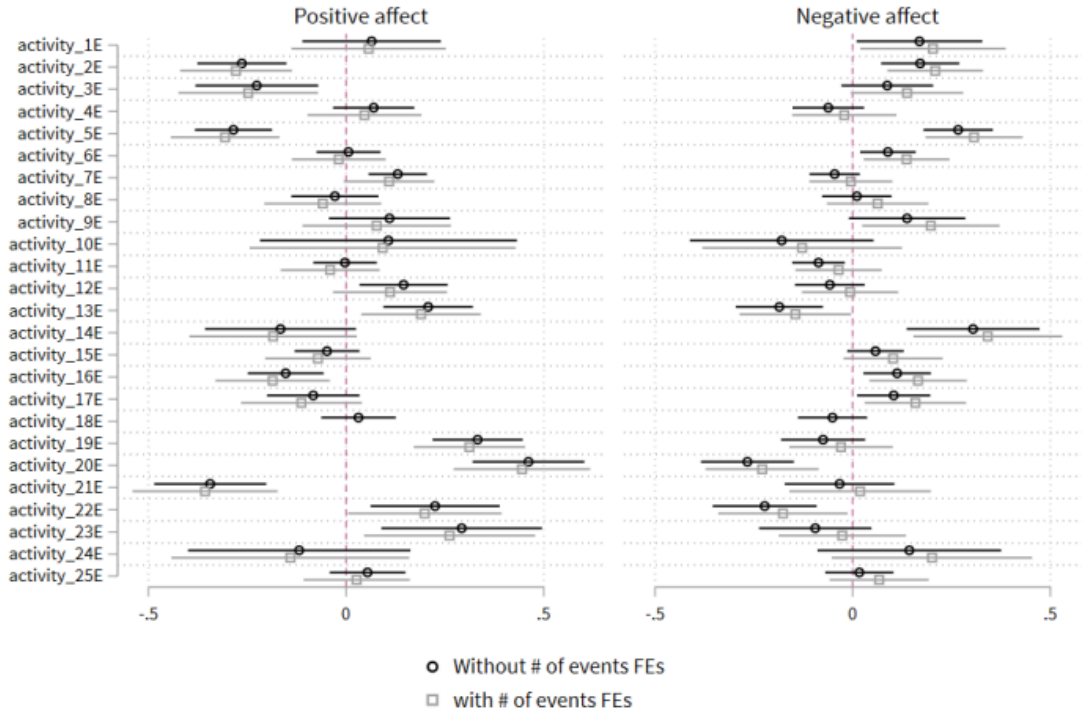
Time spent: 30 minutes

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