Myriam A. Baum | Saarland University
Campus A2 4, Room 1.02 | D-66123 Saarbruecken

Myriam A. Baum
Saarland University
Department of Psychology,
Applied Statistical Modeling |
Department for Teaching and Studies
Campus A4 4, 1.02
D-66123 Saarbruecken

T: +49 (0) 681 302 4792
myriam.baum@uni-saarland.de

Dear Prof. Dr. Brent Donnellan,

thank you very much for the opportunity to revise and resubmit our manuscript, now entitled “Reexamining the relationship between shift work and health behavior: Do fluid intelligence, socio-economic status, and self-control moderate the relation?”. Please find a revised version of the manuscript attached. In this letter, we detail the revision that we undertook in response to the reviews. Our responses to the comments are in bold, text passages from the manuscript are further highlighted in grey. Moreover, in the manuscript, all changes are highlighted in grey. We would like to thank you and the two referees for the extensive and extremely helpful comments. We believe that the revision has strengthened the manuscript and hope that you find our changes to be satisfactory so that the manuscript can now be accepted for publication.

Yours sincerely,

Myriam A. Baum (for all authors)

UNIVERSITÄT DES SAARLANDES

EDITOR COMMENTS

1) I think my first reaction when getting to the analyses was a sense of puzzlement as to why the sample was split by gender for the analyses. Both Reviewers also commented on this issue. We do not see a strong reason to split the sample and conduct separate analyses. I would simply include gender and gender by shift work interactions (as well as gender by other moderators) for all analyses.

As recommended, for the new analyses, we did not split the sample by gender. Since we have included one person per family in our final sample, the total *N* of 4,251 persons (2,242 females and 2,009 males; in some cases from the same family due to separate analyses) in the separate samples for men and women has been reduced to a new total *N* of 2,590 persons (1,421 females and 1,169 males) from different families. We now include gender, shift work x gender, gender x each moderator as well as shift work x gender x each moderator in our analyses. For this reason and due to several other adjustments based on the recommendations below, we revised the results section and the discussion section accordingly. We have now included seven new/revised tables to show the (standardized as well as unstandardized) results in the results section and as supplement material. Results of the analyses can now be found in Table 2, Table 3, Table 4, Table 5, Table S2, Table S3, and Table S4.

In sum, the new results conducted with the whole sample mirror the prior results – with one exception: We did not find any differences regarding subjectively perceived health between shift and day workers anymore (e.g., p.2):

"In accordance with the findings regarding objective health indicators, shift and day workers did not differ in the subjective perception of their health."

Moreover, in the course of the new analyses as part of the revision, we have also corrected a coding error in the SES variable that was discovered during the double check of our data and syntax. In this case, the results also mirror the prior results – with one exception: Regarding smoking and preventive health care, we found main effects of socio-economic status (p. 25):

"Concerning socio-economic status, none of the investigated moderation terms yielded significant results. Nevertheless, adding socio-economic status in addition to shift work as a predictor significantly predicted smoking as well as preventive health care behavior and led to a significant improvement of explained variance (smoking: β = -.21, 95%-CI[-0.25, -0.17], p <.001; preventive health care: β = .11, 95%-CI[0.07, 0.16], p <.001). Furthermore, the corresponding overall models significantly explained variance (smoking: $F(1, 2{,}585) = 36.71$, p <.001, R = .23, $\sqrt{(\Delta R^2)} = .20$; preventive health care: $F(1, 2{,}585) = 42.52$, p <.001, R = .25, $\sqrt{(\Delta R^2)} = .10$)."

2) Reviewer A provides several helpful references for thinking about effect sizes. I think placing more interpretation on the effect sizes and struggling with the issue of what would count as a practically significant or non-negligible effect size in both the Introduction and Discussion will enhance the paper.

We highly appreciated the references provided and found the work by Funder and Ozer (2019) to be very helpful. Hence, we added a section regarding the importance of effect sizes and their interpretation in the introduction (p. 5-6):

"Moreover, the sizes of the effects found so far might further call into question the practical meaningfulness of these findings. According to Funder and Ozer (2019), evaluating effect sizes – not only *p* levels – is crucial when drawing implications. One recommendation of the authors is to use correlation values as benchmarks of "very small" (*r* = .05), "small" (*r* = .10), "medium" (*r* = .20), "large" (*r* = .30), and "very large" (*r* = .40 or greater) effects when interpreting results as well as their meaningfulness in the short and in the long run."

We also revised the discussion substantially, focusing more on effect sizes rather than p values (p. 35):

"Bearing in mind the recommendations by Funder and Ozer (2019) mentioned earlier, our findings do not support our hypotheses that non-standard work schedules are associated with any of the investigated health-related behavior patterns in a meaningful way. Based on effect sizes, shift work was not found to be a meaningful predictor of either smoking, alcohol consumption, or preventive health care. Associations between shift work and health behavior patterns rather seem negligible. Moreover, neither the moderation analyses nor the examination of subjective health showed significant associations. Although fluid intelligence and self-control significantly predicted some of the relevant health-related behavior patterns (e.g., smoking) when looking at the *p* values, effect sizes reveal that these associations are (very) small and with regards to content not meaningful."

3) I agree with Reviewer A who questioned the winsorizing approach. I also objected to the idea of "correcting" variables for age. Why not just include age in the models and even test for age by shift-work interactions? The term correcting seems to imply something that might not be intended. I suspect this is just convention from how some discuss twin-modeling decisions. Nonetheless, I think a different word other than "correct" would be preferable and a more direct approach to accounting for age would be even better.

To address this comment, we re-ran the analyses with the observed data. As mentioned above, our results have not substantially changed and we now report the results based on the observed data. We excluded the analyses as well as the corresponding results obtained with the winsorized data.

Moreover, we followed the suggestion and included age as a covariate in the first step of all hierarchical regression analyses (for example, also see p. 15):

"Age and gender were included as covariates in step 1."

4) I did not think the references to ego depletion were needed given the uncertain empirical status of that approach and the uncertain status of how that theory related to the specific "trait" measure of self-control used in this paper.

We have revised this section (p. 8) and have excluded the references regarding ego depletion.

5) What was the correlation across Wave 1 and Wave 2 and how many people had scores at both waves?

We now include this information (p. 13):

"For 307 participants, self-control values were available at both times (telephone wave 1 and face to face wave 2). Since the two measurements showed a substantial correlation ($r(307) = .51$, $p < .001$), a mean was calculated and used in the analyses. Given the poor reliability, this correlation can be considered as good."

6) The idea of including a large correlation table suggested by Reviewer A strongly resonated with me.

We appreciate this suggestion and have included a correlation table according to Reviewer A's recommendation (see Table S1, p. 49-50).

7) Reviewer B had concerns about considering SES as an intelligence variable (see e.g., p. 26 in the paper). I strongly agree, and I think that is probably an unwise framing. I think it is fine to frame this as testing the moderating effects of SES, fluid intelligence, and self-control. Indeed, I would be quite specific about what moderators were tested (e.g., there was only one personality characteristic considered so it would be clearest to just talk about self-control rather than personality per se).

We agree that it is more appropriate to talk about the specific constructs as assessed instead of the broader realms they are part of or associated with. Thus, we rephrased this throughout the manuscript (namely, fluid intelligence, socio-economic status, and self-control). Moreover, we revised the title to mirror this decision (new title: Reexamining the relationship between shift work and health behavior: Do fluid intelligence, socio-economic status, and self-control moderate the relation?). We further

subdivided the previous paragraph "Education and cognitive ability as potential moderators" into two separate ones addressing (1) fluid intelligence as moderator and (2) SES as moderator within the new paragraph "Socio-economic status as a potential moderator" (p. 7):

"Socio-economic status as a potential moderator

Past research also suggests a link between shift work and socio-economic status. Wang and colleagues (2012), for instance, reported a higher risk of being in the lowest third of the socio-economic status distribution for women who ever worked night shifts compared to women who never worked night shifts. A low(er) socio-economic status might further be associated with unhealthy behavior patterns such as higher rates of smoking (e.g., Pomerleau et al., 1997). This might be due to a reduced consciousness regarding the effects of one's actions and, hence, a reduced probability to choose healthier behavior patterns (Neumark et al., 2003). Thus, health behavior patterns of (shift working) employees might differ depending on a person's socio-economic status."

8) I like the constraints on generality statement paper from Simons, Shoda, and Lindsay (2017) and encourage you to include a dedicated section about such constraints in the Discussion. For example, I think non-German readers might benefit from an even broader perspective on German public policies about work and health care. In other words, the null results here might not apply to countries with different government policies and health care systems and I think additional discussion about boundary conditions is worthwhile. (Threads of such constraints are in the Discussion, so this would involve bringing them together and adding even more).

We would like to thank you for this helpful reference and included a COG statement in the discussion section where we added some further discussion about the boundary conditions (p. 37):

"Constraints on Generality (COG)

It must be noted that the generalizability of the present findings might be restricted due to international differences regarding labor or health policies and health care systems. As mentioned earlier, Germany is a country with broad regulations both regarding the labor law to protect the health and well-being of employees as well as health policies to protect health and well-being of every individual. For instance, labor law regulations in Germany are rooted at the country level and, thus, apply equally to all federal states – whereas in the USA, for example, such regulations can vary from state to state. Moreover, in Germany, there is a highly developed health care system that is – in its basic and many additional services – mostly independent of one's financial or social standing and provides a low-threshold service for every citizen. The statutory health insurance represents an essential part of this health care system. In addition, since 2009, health assurance has been mandatory for all citizens residing in Germany (for more detailed information on the German health care system, see Busse & Blümel, 2017). This means that a solid health care system is provided equally for everyone and only a number of additional services are not covered by

the insurer and must be paid by the insured person itself. Thus, the results found in this study may not be applicable to other countries with less strict labor protection laws and health policies or with lower availability of healthcare services and more difficult accessibility to the healthcare system for every individual."

Due to the recommendation of reviewer B to shorten the manuscript, we kept this section rather short and provide reference for further reading.

UNIVERSITÄT DES SAARLANDES

REVIEWER A COMMENTS

General comments and summary of recommendation

This paper presents an analysis of the relationship of shift work, cognitive ability, and measured self-control with self-reported health behaviors across various demographic groups. The paper uses a large sample. There are many strengths to the study, but as presented, there are several important issues with the data analysis and interpretation that should be addressed before publication.

We thank reviewer A for this evaluation and will respond to his/her comments below.

Control variables and data handling

1) (p. 14) In your section reporting results of the MCAR test, please refer to the results as suggesting that male missing data "occurred _completely_ at random" and that female results "did not occur _completely_ at random". Specifying „completely" is important, as „Missing at Random" has a different meaning (i.e., that some of the other predictors in the data can account for the missingness; this is the assumption you make with the female sample). However, Little's MCAR test should generally not be used to determine a missing data handling strategy. This approach violates assumptions about the sampling distributions of the imputed data. If predictor variable are available that can potentially account for missingness, it is better to use these variables in multiple imputation, rather than to perform this MCAR test screening procedure.

To address this comment and to avoid confusion, we added that missing data did not occur completely at random (p. 15):

"Little's MCAR test (Little & Rubin, 2002) indicated that missing data did not occur completely at random ($\chi^2 = 189.51$, $p < .001$)."

As we no longer analyze the data separately for males and females, MCAR-results are reported for the whole sample now.

2) You state that "Data were winsorized (Sheskin, 2003) – except for likert-scaled data (namely, health behavior, subjective health, and self-control) – and corrected for linear effects of age.", but you do not provide any justification for these procedures. I see no reason to winsorize your data. Please report results using the observed data. Please provide a clear justification for controlling for age (cf. Spector & Brannick, 2011, https://doi.org/10.1177/1094428110369842). Please also report results not controlling for age.

We re-ran the analyses with the observed data and excluded the analyses as well as the corresponding results obtained with winsorized data. To address another comment above, we also re-ran analyses with age (and gender) as a covariate. We now report the results of these analyses. Compared to the previous

analyses, the main results have not changed – we did not find any meaningful main or interaction effects.

3) Your regression tables state "For the female subsample, all variables were z-standardized before analyses." Why only for the female sample? Please be consistent and report both unstandardized and standardized results for both samples. See below for comments on reporting and interpreting unstandardized results in this study because of the inherent meaningfulness of the measured outcome variables.

Z-standardization in the female sample was used due to multiple imputation. We acknowledge that this was inconsistent and have now z-standardized all variables prior to our analyses for the whole sample. We added a comment for the readers on p. 15:

"Prior to the analyses, all variables were z-standardized."

Moreover, we now provide unstandardized results of the analyses in the Appendix (see Table S2 - Table S4). We highly appreciated this comment, as it provides a more complete picture. We included comments for the readers on p. 17, p. 21, and p. 25):

"Results for the unstandardized variables can be found in Appendix (S2)."
"Results for the unstandardized variables can be found in Appendix (S3)."
"Results for the unstandardized variables can be found in Appendix (S4)."

Interpretation of results

1) The paper focuses almost entirely on statistical significance in interpreting the results. This is a problem because most if not all of the reported effects appear to be practically negligible, even if they are „statistically significant". Beyond whether the effects are statistically significant, please focus your discussion on the size of the effects. For example, I would look at all of zero-order shift work–outcome relationships reported on page 14 and conclude that shift work has negligible to small relationships with all health behvaiors, even the one that is nominally significant. The work of (Funder & Ozer, 2019, https://doi.org/10.1177/2515245919847202) can be helpful in interpreting effects sizes. They reviewed distributions of effect sizes observed in psychological research and found that the 25th, 50th, and 75th percentiles for effect sizes correspond to correlation values of r = .10, .20, and .30, respectively. These might be interpreted as rough generic benchmarks for „small", „medium", and „large" effects.

As mentioned above with respect to the editor's comments, we found the work by Funder and Ozer (2019) to be a very helpful resource for interpreting effect sizes. We added this information in the introduction and have focused more on effect sizes in the revised discussion section.

Additionally, all three of your criteria are meaningful in themselves. I recommend you focus on the _unstandardized_ regression coefficients as indices of the effects of shift work in terms of number of cigarettes smoked, amount of alcohol consumed, or estimated number of preventative heath behaviors. This „real-world“ anchoring of your results might better illustrate the meaningfulness (or not) of the impact of your predictors on these outcomes. For the coefficients with fluid intelligence and self-control, you could partially standardize the coefficients only for the intelligence/self-control variable to estimate the effect of a 1 SD/15 IQ point difference.

In the revised version, we included unstandardized regression coefficients and provided several tables of the unstandardized results as supplement material (see above). However, as we did not assess the number of cigarettes smoked or the amount of alcohol consumed, we were not able to address this comment even though we agree with the reviewer that such information would be helpful.

2) Throughout the paper, you report statistical significance tests separately by group (e.g., separately for men and women for all analyses; differences in correlations between subjective health and behaviors across gender and shift work samples). As written, based on these analyses, you appear to conclude in various places that variables are related in one group, but not the others (e.g., you conclude that shift work is related to alcohol use in men, but not women; you conclude that subjective health is related to smoking in male day workers, but not other groups). These analyses do not permit this inference. A difference in statistical significance p value across groups _does not imply_ a statistically or practically significant difference in the relationship across groups. To make an inference that variable relationships differ across groups, you must (1) focus on the difference in estimated effect sizes across group and (2) formally test the differences in relationships. For example, in the zero-order shift work–outcome relationships reported on p. 14–15, gender does not appear to be a substantial moderator of any result. However, as written, the discussion appears to suggest that results are different for men/women for the shift work–alochol relationship, which isn't really the case. The effect appears to be fairly negligible in both cases. A rough calculation based on your reported β values and the group sample sizes shows me that the $\Delta\beta = .09$ [95% CI .03, .15]. While this is technically „statistically significant“, the difference is practically very small, and both effects remain fairly negligible in size (just on opposite sides of zero). Similarly, looking at Table 4, the correlations among variables are nearly identical across shift and day workers, so the conclusion that there is a relationship between subjective health and behaviors only for male day workers is not valid. (Also, these p values appear to be incorrect? A correlation of -.21 with a sample size of 348 (female shift works) should have a confidence interval of [-.31, -.11] and a p value < .0001, not p = .101 as reported. Throughout the paper, please carefully evaluate whether differences across genders or other groups are statistically and practically significant to avoid erroneous suggestions that a relationship is present in one gender but not in the other. A brief examination of results in Table 2 suggests few to no major differences in effect sizes across genders. The authors provide no justification for gender-

stratified analyses, and I see no reason to have expected relationships to differ across groups, so I recommend reporting results only for the full gender-pooled sample.

We appreciate these helpful explanations very much. Since we followed the recommendation to re-run our analyses (as described above), we did not revise the earlier interpretation but rather interpreted the new results found for the full sample.

3) For the group comparisons of mean differences (e.g., across shift work groups), please report and interpret effect sizes (mean differences and confidence intervals, Cohen's d values and confidence intervals) in addition or in lieu of the t test results.

We now include means and standard deviations for the compared groups as well as Cohen's d and the confidence interval in addition to the results of the t-test (p. 29):

"The compared groups did not differ significantly regarding the perception of their subjective health (*t*(2,588) = 0.84, 95%-CI[-4.35, 16.35], *p* = .408, *d* = -0.06). Mean evaluation of subjective health was -0.05 for the shift working group (*SD* = 1.01) and 0.01 for the day working participants (*SD* = 1.00)."

Reporting/presentation of results

1) For transparency and reproducibility of your analyses by readers, please provide a table of means and standard deviations for each variable and a full correlation matrix among all variables. Include the multiplicative composites for all of your interaction variables (e.g., shift work x fluid intelligence). These multiplicative variables are needed for readers to be able to reproduce your results from the correlation matrix.

To address this comment, we included means and standard deviations for each (interaction) variable in Table S1 (p. 49-50). As reported above, Table S1 also provides a correlation matrix.

2) In your regression tables, for clarity, please include step 1 results in the table, as well as in the text. At first, I had thought that Step 1 was a control variables step.

To enhance clarity and readability, we included all steps in each table. As the first step is a control variables step now, we did not include the results in the text.

3) To aid readability, in your rgression tables, place the R^2 and ΔR^2 results directly as rows in the table, rather than in the table note. Additionally, I recommend reporting R and √(ΔR^2) instead of their squared values to aid interpretability (Funder & Ozer, 2019, https://doi.org/10.1177/2515245919847202). Please also provide confidence intervals for R and √(ΔR^2) (or R^2 and ΔR^2), e.g., using the methods described by (Alf

& Graf, 1999, https://doi.org/10.1037/1082-989x.4.1.70) or (Shieh & Kung, 2007, https://doi.org/10.3758/bf03192963).

We appreciate this comment! Hence, we decided to report R and √(ΔR^2) rather than the squared values throughout the whole manuscript. We further included them as well as the corresponding confidence intervals as rows in all regression tables to aid interpretability. Please see p. 15-29 as well as p. 51-59 in the manuscript for all changes based on this comment.

4) In your regression tables, the confidence interval appears to be for b, not Beta. Please place it after SE b for clarity. Please also provide a confidence interval for Beta, using the method described by (Jones & Waller, 2013, https://doi.org/10/gckfx4)

As Beta and b are the same for the standardized results, we included one confidence interval for both in the Tables reporting standardized results. We also included a clarifying note (see p. 15-16, p. 18-20, p. 22-24, and p. 26-87). For the unstandardized results in the supplement material section, we now provide confidence intervals for both, b and Beta (see p. 51-59).

Minor comments

1) (p. 12) A sentence or footnote briefly describing the SOEP study as a nationally representative survey of socioeconomic, health, and psychological variables in Germany would be useful for readers unfamiliar with these data.

To address this comment, we added a footnote on p. 11:

“The SOEP study is an ongoing nationally representative longitudinal survey of socioeconomic, health, and psychological variables in Germany. For more information, see Goebel et al. (2019).”

2) To help ensure that Zotero is able to keep your citations update to date as you write and potentially change citation styles, I suggest entering prefixes like „for a detailed overview of this project, see” here into the “Prefix” field in the Zotero citation window, rather than directly typing into your document.
See https://www.zotero.org/support/word_processor_plugin_usage#customizing_cites for details.

Thank you for this helpful suggestion. We adapted each citation that was affected (e.g., on p. 4, p. 5, etc.) without highlighting changes in grey.

Figures/tables/data availability

The tables are in many places unclearly laid out. No correlation matrix is provided, so the analyses cannot be reproduced.

We acknowledge that the tables were unclearly laid out and addressed this point by revising and extending them (e.g., on p. 15-16 or p. 18-20). As described above, we added a correlation matrix to enhance reproducibility (see Table S1, p. 49-50).

Ethical approval

The study uses a large publicly available dataset (https://www.twin-life.de/publikationen/). No information about ethical approval is provided in this paper.

We added information about the ethical approval of the TwinLife study in the section "funding information" (p. 40):

"The TwinLife study received ethical approval from the German Psychological Association (protocol numbers: RR 11.2009 and RR 09.2013)."

Language

The writing is clear.

UNIVERSITÄT DES SAARLANDES

REVIEWER B COMMENTS

General comments and summary of recommendation

1) I would like to commend the authors' thorough review of shift work and health behaviors. I learned a lot about this topic from this reading.

We would like to thank reviewer B for the positive evaluation. We will respond to further comments below.

2) My general suggestion is for the authors to take a research synthesis approach to summarizing past research instead of individually listing the study results. It might be a personal stylistic preference, but unless there is a strong rationale, I am not sure if statistical results of individual studies such as sample size and odds ratios are necessary to be included. Overall, I think the authors made a good case that the existing evidence is mixed. But I think their presentation can be improved to make the paper a more enjoyable read.

We agree with the reviewer that a research synthesis approach to summarize past research is absolutely reasonable. Hence, we decided to adapt our introductory section accordingly and removed the majority of individual studies. To aid readability, we further excluded information such as sample size or Odds Ratios of the remaining individual studies.

3) I think individual differences cognitive ability as a moderator is an interesting one. However, I am not completely sold by the hypothesizing. Intuitively, it seems to make sense that maybe people with higher cognitive ability use healthier coping strategies. But is this based on theory? Is there past research that support this prediction in other contexts? I would like to see a stronger argument made for this hypothesis.

In the introduction, we edited the section about fluid intelligence as a moderator and added further evidence (i.a., from contemporary psychological literature on cognitive epidemiology). However, since we also agree with Reviewer B regarding the note on the length of the manuscript, we have kept the addition short and, thus, restrict our addition to two major research findings from the field of cognitive epidemiology as well as one reference regarding coping strategies (p. 6):

"Fluid intelligence as a potential moderator

The wide-spread view seems to be one of "variable shiftwork as a blue collar phenomenon" (Gordon et al., 1986, p.1,226). However, Gordon et al. (1986) reported that "education was not clearly related to the probability of being a shift worker" (p.1,226). Bearing in mind the heterogeneity of the occupational groups of shift workers, it makes sense that the IQ distribution is not as restricted as was originally assumed. However, literature in this area is limited to a relatively small number of studies and should therefore be interpreted with caution.

Regarding measured intelligence and health-related behavior patterns, such as alcohol consumption, evidence is mixed with a large proportion of studies revealing a negative relationship (e.g., Sjölund et al., 2015). Batty and colleagues (2007), for example, reported a negative association between child IQ scores and the prevalence of ever having smoked as well as heavy alcohol consumption in adulthood. Moreover, Wraw and colleagues (e.g., 2018) showed that a higher IQ in youth was associated with a lower likelihood of – amongst other things – heavily consuming alcohol or smoking in middle age. In the context of shift work, this link appears to be especially important. Even if shift workers relied on less healthy behavior patterns such as consuming alcohol as, for example, a coping strategy, those with higher cognitive ability may show better health behaviors or might use healthier strategies to cope with stressful working conditions, weakening the link between shift work and negative health-related behavior patterns. This assumption is further supported by the work of Minehan et al. (2008), who showed an association between cognitive ability and coping strategies in the prediction of drug use and by certain empirical models emphasizing an association between substance use and avoidance coping strategies (e.g., Ebata & Moos, 1991)."

4) I appreciate the power analysis. Could the authors also report the power of the moderation hypotheses?

To address this comment, we conducted an additional power analysis for the full moderation model and included the results in the manuscript (p. 10):

"For the full model (two control and six predictor variables; see step 5 of the hierarchical regression analyses below) of the moderation analyses, a total sample size of at least N = 759 was required to detect a small effect size (f^2 = 0.02) with an alpha of .05 and a power of .80."

5) I realize the authors are interested in gender differences. From a presentation perspective, gender is another variable in the study. Treating male and female respondents as different 'subsamples' seems unnecessary.

As described above, we addressed this comment by re-running the analyses without splitting the sample by gender.

6) The internal consistencies of the moderator measures are concerning. Are the Cronbach alphas listed on page 13 for the subtests of fluid intelligence? The authors summed the subtests scores – what was the overall alpha for the summed scale? Internal consistencies for the self-control measure is also a bit troubling. How many items were there?

We apologize for the confusion. Yes, the Cronbach's alphas listed are for the subtests of fluid intelligence. We now added the overall alpha for the summed scale (p. 12):

“Cronbach’s alpha across all scales was .75.”

Self-control was assessed using three items of the SCS-K-D (Bertrams & Dickhäuser, 2009; p. 12).

We agree that the internal consistency of .58 for the short scale is on the low end of acceptable reliabilities and address this now in the limitations section (p. 37):

“Moreover, although not untypical for very short scales, the internal consistency for our three-item self-control measure is rather low. Unfortunately, time constraints prevented the use of the full 13-item inventory (SCS-K-D; Bertrams & Dickhäuser, 2009).”

7) The authors used SES as a ‘indicator of cognitive ability’ – is this theoretically appropriate? What is the convergence between SES and actual cognitive ability in the sample? Given the authors have an actual measure of cognitive ability, I am not sure why the authors need to use an indicator like SES. Relatedly, the authors are using SES and ‘education’ interchangeably. I am not sure of SES is the same as education. Instead of equating SES with intelligence, why not discuss how SES might be uniquely related to shift work and health behaviors? I think that is an interesting question on its own merit.

This point was also raised by the editor and we agree that it is more appropriate to treat SES not as a proxy for education or intelligence, but as a unique variable of interest. As can be seen above, we rephrased this and added a new paragraph introducing SES as a potential moderator.

8) The authors analyze the data separately for male vs. female. Why not use a combined sample and use gender as a predictor in the regression equation and then check for moderation or simple slopes separated by gender? The authors’ sample is adequately powered for this. The authors moderation hypotheses do not seem to be contingent on gender, so separating the data into male vs. female seem to unnecessary.

We re-ran analyses including gender, shift work x gender, gender x each moderator, and shift work x gender x each moderator in each hierarchical regression analyses (see above).

9) In addition to significance tests, the authors should present the difference in shift vs. non-shift work in more meaningful metrics. Given that the variable is dichotomized, I suggest the authors present raw and standardized differences in health behaviors (e.g., Cohen’s d) between shift vs. non-shift work groups.

Based on this recommendation, we conducted mean-difference analyses to examine differences between shift and day workers regarding their smoking behavior, alcohol consumption, and preventive health care behavior. Results can be found in the exploratory analyses-section (p. 33):

"Differences between shift and day workers regarding their health behavior

To further exploratory investigate whether shift workers and day workers differed regarding their health behavior, we conducted mean-difference analyses.

The compared groups significantly differed regarding their smoking behavior (t(570.547) = -4.49, 95%-CI[0.16, 0.36], p_{corr} = <.001, d = 0.25) as well as their preventive health care behavior (t(2,588) = 3.41, 95%-CI[0.08, 0.28], p_{corr} = <.001, d = 0.18). Mean of the shift working group was 0.22, respectively -0.15 (SD = 1.12, respectively SD = 0.99) and -0.04, respectively 0.03 for the day working participants (SD = 0.97, respectively SD = 1.00). The groups did not differ significantly regarding their alcohol consumption (t(646.221) = 1.37, 95%-CI[-0.03, 0.17], p_{corr} = .258, d = 0.07). Mean of the shift working group was -0.06, (SD = 0.94) and 0.01 for the day working participants (SD = 1.01)."

10) Results on shift work and subj health: please report effect size information (cohen's d). And again, I am not sure if splitting the samples on gender make sense here. I recommend analyzing the whole sample, and maybe look at gender as a moderator.

We now report effect sizes (see above). We also re-ran the analyses, including gender (as well as interactions with gender) in these analyses.

11) Occupational status was first mentioned on page 24. How is this variable measured? Table 5 did not give much clarity. The authors noted that they conducted "mean difference analysis" for table 5. I am not sure what that means. The authors report a f statistic – is this in the context of regression?

We now provide more information on the assessment of occupational status on p. 13:

"Occupational status

Occupational status was measured with the following question: "What is your current occupational status?" (TNS Infratest Sozialforschung, 2014). Participants were advised to indicate this in relation to their main activity with one of the following response options "blue-collar worker", "white-collar worker", "civil servant", "self-employed", and "apprentices / trainees / interns"."

Moreover, for more clarity we now state that we conducted ANOVAs to examine differences in health behavior related to shift work for different occupational statuses (p. 31):

“Table 7 shows the results from the ANOVAs conducted.”

12) In sum – the authors are examining an interesting topic in a large dataset. I think some of the writing and the analyses could be made much more efficient. I feel the paper could be shortened by 25% without losing too much content and still maintain a coherent presentation of findings.

We are grateful for this positive summary. We agree with him/her about the length of our manuscript and have shortened several sections where possible. However, as we also added some shorter sections (e.g., on effect sizes or COG) and included additional and newly edited tables to enhance readability, clarity, and reproducibility, the total number of pages of the manuscript has not decreased.

Figures/tables/data availability

they are adequate

Ethical approval

Na

As mentioned above, we added an information about the ethical approval of the TwinLife study in the section “funding information” to address this point.

Language

Is the text well written and jargon free? yes