Biological Psychiatry
MS Number: BPS–D–16–01637
Title: The Impact of Depression on Brain Activity
During Source Memory Retrieval

Dear Dr. Dan Dillon,

Thank you for submitting your work to Biological
Psychiatry. It has been reviewed by experts in the
field, and their comments are enclosed. Based on the
comments of the reviewers, I regret that we will not
be able to accept your paper for publication. I realize
that this is disappointing news. In many cases such
as this, we understand that you would likely be able
to address many of the reviewers' concerns. However,
limited journal space and a large number of
submissions require us to reject over 90% of all
submitted papers. As a result, we often are forced to
make difficult decisions and reject manuscripts with
considerable merit.

I hope the reviewers' comments are helpful to you as
you prepare the paper for publication elsewhere. I'm
sorry the decision was not more positive, and hope it
does not discourage you from submitting new work
to Biological Psychiatry in the future.

Biological Psychiatry is a member of the Neuroscience

Peer Review Consortium. The Consortium is an alliance of neuroscience journals that have agreed to accept manuscript reviews from each other. If you submit a revision of your manuscript to another Consortium journal, we can forward the reviews of your manuscript to that journal, should you decide this might be helpful. You can find a list of Consortium journals and details about forwarding reviews at http://nprc.incf.org.

For your guidance, the reviewers' comments are included below.

Thank you for giving us the opportunity to consider your work.

Yours sincerely,

Dr. John H. Krystal
Editor
Biological Psychiatry

Comments:
Deputy Editor's Note: I have now received comments from five individuals with expertise relevant to this manuscript. They agree that the topic areas are important and the use of ERPs to examine these memory processes is relatively novel in relationship to depression. However, at the same time, they have a number of very significant concerns, with many of these concerns shared across reviewers. These

these concerns shared across reviewers. These concerns include the need to have an introduction that much more clearly outlines the literature, making important distinctions between rote memory and autobiographical memory, which have some dissociable psychological and neural mechanisms. A second major concern is that the paradigm used may not have been particularly sensitive to memory deficits in general, and not in MDD, which creates concerns about how to interpret the ERP data when there are no behavioral effects. As noted by one of the reviewers, the fact that confidence was more associated with MDD may not reflect memory per se, but rather the types of cognitive biases often found in depression. Further, these ERP results do not seem to relate to important clinical facets of MDD in your data. There were also a number of suggestions about the statistics, including concerns about power with the MLM approach and the need to present clearly parallel analyses of the ERP and the behavioral data. Further, several reviewers asked for a direct examination of the association between the ERP data and behavior. Several reviewers also raised the concern that sleep quality was introduced, but not sufficiently justified or explained and it was not clear how multiple comparisons were handled in that analysis. Further, as clearly noted by the fifth reviewer, it is not at all clear that the data actually support the primary conclusions of a memory deficit explanation of the differences between the MDD and control

participants.

Reviewer #1: This is a potentially interesting study of the role of encoding instructions on recognition, and whether depression interacts with encoding strategy to better illustrate key features of major depressive disorder. IT also highlights key ERP findings in relation to memory processes.

There are a number of strengths.
1. The design of the task to look at deep vs shallow encoding
2. the selection and matching of MDD and HC groups – do the effects still hold when covarying IQ or education?
3. careful analysis structure for the performance and ERP markers.
4. It seems that the last paragraph of the behavioral results is most interesting. Deep encoding improves memory for MDD and light encoding does not. With a few more subjects, this effect could be more strongly interpreted.
5. It would be worthwhile to see how well recollection and confidence values are correlated, by group (and interaction). Does the ERP effect relate more to one than to the other?
6. The slower RTs as justification for "deeper" processing is interesting, but could use some more justification.
7. Inclusion of the hit rate analysis showing similar

7. Inclusion of the hit rate analysis showing similar effects. But why does it seem like mobility is not really working for HCs?

I have some concerns
1. how the literature in memory is presented (e.g.,
a. difference between autobiographical and rote memory in depression. There is a great deal of literature on rote memory recall difficulties in depression. There is less on autobiographical. But the introduction mixes and merges these two ideas in ways that are difficult to follow and does not cover the larger rote memory difficulties in MDD.
b. lack of clarity in distinguishing recognition and recall with a broader concept of retrieval (cued vs uncued)
2. this is a cued recognition paradigm, and as evidenced by the results, elicits a near ceiling effect in performance. That is very different than the free recall autobiographical and rote memory concerns reported by patients and replete in the literature. The same literature illustrates how recognition memory is frequently not impaired in MDD. This leads to a broader concern about how the ERP data is interpreted. For successful recognition, there are some parietal differences between MDD and HC for Mobility. And mobility seems to enhance relative memory for question, if potentially diminished for animacy.
3. the interpretation of the ERP findings in light of low power – is something important being missed

here?
4. lack of a linkages between the ERP and the key behavioral outcome in MDD. If there were more errors, we'd have a better grasp of whether this group difference is a reflection of enhancement or not.
5. More broadly, there is a great deal assumed that the reader should kow about this line of research and prior work by this group (and be careful if you use the word replicate prior findings by our group that these were based upon entirely distinct samples – e.g., not a carryover of the same HC group).

Minor details

Figure 5 lacks some clarity and details. Where is demarcation of group in the figures? In general, the figures and captions do not integrate as well with the text. Just needs to be tightened up in the figures, e.g., labeling columns for time windows in both HC and depressed. Figure 4 organization does not allow for easy comparison of groups.

Reviewer #2: The authors present behavioral and ERP data recorded during a memory task among MDD patients and healthy controls. While depression is strongly associated with memory impairment, this

strongly associated with memory impairment, this phenomenon has not been well-studies using neuroimaging techniques. As expected, the MDD patients were less accurate and less confident in their responses overall, with one notable exception: MDD patients were highly accurate for the most difficult condition, conceptual source memory following judgments of mobility during encoding. This pattern was also mirrored by the ERP activity: MDD patients exhibited blunted left parietal activity overall during recollection, but relatively intact activity specifically for the conceptual/mobility condition.

The manuscript is very well-written, the aims of the study are clear, and the analyses are thorough. The findings of relatively normal performance and neural activation in MDD specifically in the context of sustained attention are striking, and they raise important new questions about the nature of memory impairment in MDD and how it may be targeted in treatment. I have a few suggestions to further improve the quality of the manuscript:

1.   The authors very carefully analyze the Group x Cue x Encoding Task interaction in the accuracy data, and then go on to show a very similar pattern of effects in the ERP data. It is not clear to me, however, that the authors formally tested for the three-way interaction in the ERP data. If I understand Figure 5 correctly, they present Cue/Encoding contrasts separately for the two groups, as well as the group

contrast for the Mobility condition. These all seem to represent follow-up tests that would decompose the three-way interaction. I recommend also reporting the three-way interaction itself.

2.   Were the ERP and accuracy data correlated within the MDD group? Given that the question/mobility condition elicited relatively normal performance and left parietal activation, it would be important to know if these two effects are, in fact, associated with each other.

3.   I'd like to see the authors grapple with the role of comorbid anxiety a bit more. Secondary anxiety disorders were allowed within the MDD group, but the prevalence rates are not reported. Anxiety symptom severity was assessed with the MASQ, but relationships with task performance and ERPs are not considered. For example, does the effect of sleep quality hold when taking anxiety into account? Does anxiety also predict ERPs, over and above sleep quality/depression?

4.   Related to the point above, the findings with sleep quality are potentially interesting, but I do not think they receive adequate attention in the text. Why was sleep quality of interest for this particular study? I think it is a sensible choice given that both mood and anxiety disorders are characterized by sleep disturbance, but the authors' logic should be stated explicitly within the Introduction.

explicitly within the introduction.

Other, minor points:

–Introduction, p. 4: In the last paragraph, the authors hint at a Group x Cue x Task interaction ("one of our tasks promoted deeper encoding than the other...and the MDD group was quite accurate"), but do not formally state it until later on in the manuscript. I think this is worth stating more explicitly here, too.

–Results, p. 12: "These data suggest that recollection was strongest under the Question cue and reduced in MDD." This statement combines two different main effects (i.e., increased overall for Question, blunted overall for MDD), and reads as if it was an interaction (i.e., blunted in MDD specifically for the Question condition).

–Table 1: I recommend adding effect sizes to this table.

Reviewer #3: Barrick and Dillon present an excellent study of source memory retrieval in major depressive disorder. The most interesting aspect of the study is that the design of the paradigm provides a very rich data set, allowing a number of different analytic possibilities. I have a number of comments on the manuscript.

1.   The last two paragraphs of the introduction are tricky to follow, as they seem to assume that the neural systems (e.g. left PFC) are necessarily involved in certain kinds of encoding. It might be easier to stick to hypothesis based on psychological constructs in the first instance (e.g. depth of encoding), and then explore the mapping to neural systems subsequently (e.g. parietal, PFC systems).

2.   In addition, the last paragraph implies that the mobility task operates as a kind of confound, or at least, un-anticipated encoding manipulation. This leads to a slightly more superficial treatment of the task-dependent finding than it might deserve, and also seems to be overly dependent on the relatively spared behavioral finding – when other factors (a priori, or RT/confidence) might also be influential in coming to such a conclusion.

3.   The authors argue that the depressed participants generally have difficulties with the task, but this does not seem to be particularly strongly represented in overall performance (error rate). The fact that confidence is better associated may therefore reflect cognitive biases (meta-cognitive?) rather than memory deficits. Indeed, the finding implies that neural signatures of confidence (or perceived error likelihood) might be distinguished from responses which reflect the capacity to retrieve accurately. To test these accounts, ERPs could be

accurately. To test these accounts, ERPs could be related to individual differences in behavior.

4.   Do MDD show less difference confidence ratings for correct vs incorrect responses, compared to controls?

5.   Might the parietal ERPs reflect a capacity or effort to visualize the source material during retrieval?

6.   The PSQI correlations reminded me of the work of Jensen/Mazaheri and others showing that slow ERPs are related to asymmetric alpha oscillations. If drowsiness is related to altered alpha oscillations, one might expect PSQI to be associated with slow ERPs as is observed. It may also relate to the attention proposal mentioned in the introduction. Either way, this finding is not greatly discussed.

Minor points:
1.   If mobility vs animacy has such a big effect in controls, shouldn't ERPs be presented broken down by this factor too (rather than just fig 5)?
2.   Were reaction times and error rates auto-correlated across time? Did the statistical models correct for this?
3.   A voltage change on an ERP does not really reflect 'activation' see page 13, or at least this would be controversial.

4.   Page 13: 'in controls, no significant effects were seen in any time window': this is ambiguous.
5.   Abstract: 'slasting'
6.   Page 4: 'loses'

Reviewer #4: Page 3: The sentence "In short, memory retrieval is impaired in depression and enhancing it can bring lasting relief" is unsubstantiated and needs clarification and references.

Page 3: The authors used neutral stimuli. While they argue that this avoids confounds associated with mood congruent encoding, it also makes the study less relevant to depression and to memory for negative material.

Page 5: From the participants and self-report it is unclear whether the depressed participants were actually clinically depressed at the time of testing. For example, the participants had no medication use in the past two weeks (6 weeks fluoxetine, six months for neuroleptics). It is unusual for the patients in a study of clinical depression to be off medication unless they are remitted. Were these patients clinically evaluated at the time of the study by a clinical psychiatrist or clinical psychologist?

Page 5: There was a large number of exclusions

Page 5. There was a large number of exclusions (n=10 for controls, n=2 for depressed).

Other minor points are:
The authors had an a priori rejection criterion for bad channels. However, they have not provided rationale for the rejection process. They have not commented on the mismatch between the groups with regards to the number of participants excluded (10 versus 2).

All electrodes within 4 cm of each other were regarded in the same cluster. The authors should specify the rationale for the clustering more in detail. For 128 channel EEG, 4 cm within each other might result in quite large clusters leading to generous localisation assumptions.
How does cluster-based permutation control for such assumptions?

The authors speculated on the left prefrontal activation during the retrieval but the data was not relevant. A more comprehensive explanation is needed to explain how they related the reduced parietal ERP amplitudes and the left prefrontal activity.

The main focus of the discussion should be to discuss the value and feasibility of reduced parietal amplitudes in the presence of relatively intact behavioural performance. Are the authors speculating that this finding is a vulnerability marker for

achieving remission?

Reviewer #5: This is the first review of 'The impact of depression on brain activity during source memory retrieval' by Barrick and Dillon. The authors compare controls and individuals with major depressive disorder (MDD) during the retrieval of source memories targeting either prior semantic (mobility or animacy ratings of study materials) or perceptual experiences (right or left side of screen during study). The behavioral data suggested differences between the groups during semantic source attribution and the analysis of concurrent event related potentials suggested differences between the groups for only one class of prior study materials when comparing semantic to perceptual source retrieval tasks. This is an interesting topic and a novel design with potentially informative outcomes. I'm neither an expert in MDD nor ERPs and so will keep my questions or concerns largely focused on design and source memory issues.

I must admit I found the report quite hard to follow for several reasons, most of which have to do with the unusual analysis approach to the behavioral data. For instance, the authors appear to combine the direction of the report with the confidence when

direction of the report with the confidence when analyzing source accuracy. They then also consider confidence in isolation. Presumably these two analyses are somehow dependent. Regardless, the source accuracy analysis is fairly non-standard and it makes it difficult to isolate response bias from accuracy differences across the groups. It would be easier to understand the accuracy differences if for each of the source tasks (location or semantic) the authors calculated a discrimination and bias score for each subject using signal detection or similar approaches. For example, for semantic judgments they could present the percent correct (or d') and the percent mobility (or C or Beta) for each subject. This would allow the reader to parse whether the two groups were
similarly accurate and similarly biased for the questions. Based on Table S1 it seems it may be the case that both groups demonstrate a bias towards 'mobility' conclusions although it may be the case that isolating this from an accuracy effect may be impossible without the presence of new items in the design.

The MDD subjects appeared to be generally less confident than the controls during the source tasks but the authors didn't appear to consider confidence during the parity judgments and the design suggests the subjects rated confidence during this task as well. This would be useful to know since it would either suggest a tonically low confidence across cognitive

domains or a lowered confidence specifically during memory judgment. Regardless, I believe the authors describe or imply that this confidence difference constitutes a memory or retrieval impairment. This seems inappropriate since the data do not demonstrate that confidence is differentially predictive of success at the level of trials for the two groups. In other words, a generally lower confidence doesn't mean that less evidence is recovered, nor does it mean that metacognitive resolution or calibration is impaired.

The authors use multilevel modeling (MLM) for the behavioral analysis. However, with this small number of subjects (in which conditions are nested) and a balanced design this seems unnecessary. Although they indicate that it makes the consideration of covariates easier, I don't quite follow the reasoning because covariates can be considered using standard multiple regression applied to the accuracy or bias score of each subject. Moreover, I found it extremely difficult to follow the MLM because the exact equations/models being considered are never literally presented in the text and the nested model comparisons are not presented in tables. Perhaps this was done to save space, but understanding MLMs really is aided by specifying the model equations illustrating the fixed and random components. I also think the MLMs (but am not certain) are different from how judgments are often modelled in memory (e.g., Wright, Horry & Skagerbeg 2009; Wright &

(e.g., Wright, Horry & Skagerbey 2005; Wright &
Carlucci 2011). In these
approaches the DV is usually the response on each
trial and the key IV is the class of the item which is
modelled as a random effect. So for example, in the
semantic source, responding animacy might be coded
as 1 and mobility as 0. The IV, would be the class of
the item on each semantic source trial (animacy or
mobility) and thus the coefficient of this IV is the
degree to which the item origin governs the
responding of the subject (viz. accuracy). It is
modelled as random since subjects will obviously
randomly vary in their individual accuracy. The
intercept in such a model, which is also modelled as
random, reflects the bias of the observer towards one
or another classification (either with or without other
covariates accounted for). Presumably it would be
easiest to apply such an approach separately to the
semantic source, location source, and parity
judgment tasks…testing for a group interaction
within each. If the authors insist on using MLM it
would seem that
this more standard approach would make it easier to
follow the results.

In the discussion of their findings the authors
suggest encoding differences across the materials
linked to reaction times. That is, they suggest that
longer reaction times equate to deeper processing of
materials. This is problematic as it is fairly easy to
demonstrate that a shallow task can elicit a longer

reaction time than a deep task. For example, alphabetic sorting of the first and last letter of words is both slower and shallower than pleasantness ratings of the same materials.

In the consideration of ERPs the authors indicate that observed ERP differences over parietal areas meant that 'recollection was strongest under the Question cue and reduced in MDD.' This is problematic for several reasons. First, the ERPs are only conducted on correct reports. Second, I thought the behavioral data did not show an accuracy deficit for the MDD participants and thus claims that the ERPs correspond to group differences in the quality of memory evidence seem strained. Finally, this sort of reverse inference would be more compelling if the component in question was shown to correlate with performance at the level of individuals. Thus a larger response would need to be shown to predict increased accuracy (or certainty) for individuals within both groups, but be generally lower for one versus the other group. Without this it is just as likely the effect reflects global differences in the certainty of responding (even in the parity task) or some other third
variable explanation unrelated to memory evidence per se.

In the ERP analysis discussing Figure 4 the authors acknowledge that there were no differences across the groups for either source task when the parity task

the groups for either source task when the parity task was used as a baseline for subtraction. Yet the authors discuss the two groups separately and where the activations were spatially different across the groups. This could give the impression that these differences are meaningful, which the direct comparison demonstrates they are not. Thus it is not clear that the separate plotting of the groups is useful. After this they acknowledge that a secondary analysis was conducted. It might be useful to emphasize that this analysis was exploratory and hence increases the type I error possibility. During the secondary analysis in Figure 5 they compare Semantic minus location queries at test for each class of semantic task items from study (collapsed across previous side), I think. I had a bit of a hard time figuring out exactly what this analysis isolates, however, the fact
that it is conducted on a single source of items may mean that it is impossible to isolate bias and retrieval phenomena. Again, without decomposing behavioral performance into separate accuracy and bias scores it is hard to tell. Regardless, in any source memory design, one cannot identify definitively the relative degree to which a successful judgment to a specific class of study item is because of retrieval or a tendency to favor that classification during the task.

If I understand the design, the subjects are switching randomly between parity, semantic and location source judgments at test. If so, this may afford some

interesting exploratory analyses based on the possibility that it may be more difficult for the MDD subjects to dynamically reconfigure source monitoring sets (or differences in proactive versus reactive cognitive control). Here a useful analysis (behavioral and ERP) may be to consider source trials conditioned by the immediately preceding trials. Is there any evidence for example suggesting greater difficulty going from location to animacy (or visa versa) in one versus the other group?

The design illustration in Figure 1 and the methods suggest that the response scale is a little unusual in that participants are not required to commit to a judgment on the trials. That is, it spans from sure of one classification to sure of the other using 5 values. This means that '3', the center value, corresponds to no actual decision with respect to the question posed to the participant. That is, it is perfectly neutral with respect to the query. I would be particularly interested in seeing whether the two groups use this non-committal response option similarly across the three tasks as one might assume that MDD subjects might be more prone to not rendering conclusions when given the opportunity to avoid doing so.

Overall this is an interesting topic considered with useful methods. However, the authors seem to lean towards a memory deficit explanation of MDD/control differences despite the fact the data don't appear to support such a conclusion unless the

behavioral findings are somehow disconnected from the interpretation of the ERPs. So while I think they have uncovered some interesting differences I'm having a hard time grasping their functional significance. Perhaps a more standard approach to the behavioral analysis might help clarify these interesting ERP effects.