Thoughts:

0. Try correlations with condition effects, not difference scores

0. Don’t respond to the reviewers in order, respond to IGD first since he asked for the most and the changes made in response to him are the most consequential.

1. We need to emphasize two effects: Cue x Task and Group x Cue x Task. Cue x Task because it is very strong, novel, and in line with NPS focus on basic work and advances in cog neuro.

2. A possibility for Fig 5: collapse across groups for the Q-S/animacy plot, since there is no Group x Cue interaction there. That will also bring out the nature of the Cue x Task effect. Probably worth computing Q-S/mobility across groups too (I bet we have both these graphs already and I’m just forgetting it).

3. For the question about distance between channels, here is the output of the spatial\_neighbors function with dist\_bw\_chans = 4:

max\_dist value of 4 corresponds to an approximate distance of 4.06 cm (assuming

a 56 cm great circle circumference head and that your electrode coordinates are based on an idealized spherical head with radius of 8.774327).

Min/Max distances between all pairs of channels (in chanlocs units): 1.460951/19.205354

Median (semi-IQR) distance between all pairs of channels (in chanlocs units): 12.082978 (3.162924)

Mean (SD) # of neighbors per channel: 6.2 (1.5)

Median (semi-IQR) # of neighbors per channel: 6.0 (1.0)

Min/max # of neighbors per channel: 1 to 9

Song et al. (2015) (Don Tucker’s group), in *J Neurosci Methods*, indicate that the mean distance b/w channels with the 128 channel EGI cap is ~2.7 cm. To test how sensitive our results are to this variable, I re-ran the MDD vs. HC Q/MI contrast with dbwc = 2 and got no significant results (sensible b/c that is lower than the mean distance and spatial\_neighbors indicates a min/max # of neighbors from 0-3), and with dbwc = 3, which gave me no significant results from 400-800 but a significant cluster from 800-1400 ms. Looking at the plots, the lack of significant effects from 400-800 ms is unfortunate because there is clearly a strong finding there, and follow-up tests on the waveforms in that time-window from electrodes identified by the dbwc = 4 analysis yield highly significant findings. Thus, it seems to me that dbwc = 4 is an appropriate setting for these data, keeping in mind that a mean interelectrode distance of 2.7 cm reflects data from areas where the spacing is much tighter than over the parietal scalp, where stretching is pronounced (e.g., over the occiput there is no stretching and the mean distance is notably smaller, which brings down the mean). (I’ve added a line to the text on this point.)

Visual inspection of Figure 6 should also limit concern about this, for two reasons. First, you can see that the clusters of significance closely follow the topographies where electrode-level significance is high—that is, you don’t see clusters that have a few electrodes in regions of high significance that spread to areas of low significance simply because the electrodes are proximal to each other. Second, you can see how strict the clustering assumption is by looking at the left parietal effect from 800-1400 ms and again from 1400-2000 ms; there is clearly strong activity there but it is very focal, and apparently not big enough to be considered robust. Thus, I think if one were to have a concern about this method it might be that it is overly conservative, not anti-conservative. But in this case we feel it’s worthwhile because the method provides the critical attribute—namely, a principled way to look across the whole head (important because we don’t know where the between-group effects may be).

4. Starns & Hicks (2005) include a “guess” response (so it’s not that unusual . . . I’m guessing I grabbed it from Simons and colleagues too so check there as well). “We also included a *guess* response because participants certainly should not have source information for items that they recognized only on the basis of guessing.”

5. Regarding a threshold for artifacts . . . in his 2014 book “An Introduction to the Event-Related Potential Technique”, Steve Luck reports using a threshold of 25% artifacts to reject datasets in his work with college students, but notes that “In our experiments on schizophrenia, we see a lot more artifacts (*in both the patients and the control subjects*), so we exclude subjects for whom more than 50% of trials were rejected [emphasis added]” (p. 210). It was on the basis of this statement that we adopted a criterion of 50% contaminate trials for rejection of datasets, and we have added a reference to Luck, 2014, to indicate this. Our use of 18 bad channels as an additional threshold is admittedly more arbitrary, but our work with pilot subjects indicated that for some subjects (e.g., those with very thick hair), signal quality at the base of the head was consistently poor as it was difficult to ensure good contact with the scalp. There is no limit to the number of poor quality electrodes that can be interpolated, but every interpolation entails the loss of data at one electrode. We felt that loss of up to ~15% of the data (i.e., 18 of 128 channels) was a reasonable threshold to use before declaring a dataset too noisy to retain.

6. I am looking for any other paper where a deep encoding condition is associated with worse performance in a conceptual vs. perceptual retrieval test, as is the case for us with respect to animacy words (worse for Question vs. Side). That’s an interesting an odd result, because you’d think that deep processing at encoding would support good conceptual retrieval—you had to think about the semantic properties of the object denoted by the word to answer the encoding question, but you didn’t have to think at all about Side and the Side placement is totally arbitrary, so you’d think you’d get Question > Side. Since you do not, I think this result must reflect interference at retrieval . . . you encoding the words okay in the animacy task, but when you respond to the Question condition you get mixed up vis-à-vis the mobility judgments and so that drives Question accuracy way down. Is there any precedent for that in the literature? Well, in Starns vs. Hicks (2005) Experiments 1A and 1B they did not control encoding strategy but they do show better source memory for font sizes than locations . . . but it’s not like font size is a semantic property of the words so I don’t think this really helps. This paper is a difficult read but its point is simple; in multi-source retrieval experiments, such as ours, people retrieve information about the two sources independently, and retrieving information about one source does not seem to cue memory for the other source. In their experiment 2, for instance, people learn two bits of source information for each word (font and location), and are either tested on both sources at one time, or at separate times (i.e., retrieve the font for every word, then go through again and retrieve the location for every word). You might have thought that you’d get some source-to-source cuing in the first condition, but you don’t—performance for either source is basically identical across these two testing formats. The bottom line is that they find a robust correlation b/w memory for the two sources—if you remember one accurately you’ll probably remember the other accurately—but they find no evidence that remembering one source helps you remember the other. It seems more likely that the correlation just reflects good encoding of the entire episode. **Actually, I think our data are consistent with this. Side information comes out just fine no matter what the encoding task, but Question information is affected . . . so they are separable bits of information about the episode.** “These results demonstrate independence in memory for different source dimensions in the sense that remembering one dimension does not alter the probability that other dimensions will also be retrieved.” “By demonstrating that an encoding variable increases memory performance on one source dimension without affecting another, the results of Light and Berger and of Marsh et al. suggest that contextual features are encoded and retrieved independently.”

Vogt and Broder (2007) provide a nice summary of Starns & Hicks (2005); when a source cue matches the encoding condition, source accuracy along that dimension improves—but critically, this benefit does not spread to the other source dimension. However, V&B argue that these results may be partially contaminated by the use of the average conditional source identification measure (ACSIM). Thus, they go on to use more complex procedures to address this issue, and the bottom line is they replicate S&H.

**Response to reviews of**

***The impact of depression on brain activity during source memory retrieval***

We were delighted to receive positive feedback from the reviewers, who wrote that “Barrick and Dillon present an excellent study of source memory retrieval in major depressive disorder” (Reviewer 3), that “The manuscript is very well-written, the aims of the study are clear, and the analyses are thorough” (Reviewer 2), and that “This is an interesting topic and a novel design with potentially informative outcomes” (Reviewer 5). We also appreciate the reviewers’ constructive criticisms. Below we provide point-by-point replies to each one; we have taken the critiques very seriously and done our best to address them in this substantially revised manuscript. We believe the paper is significantly improved, and we hope the reviewers will agree. To improve the flow of this letter, we respond to reviewers in an order that reflects the scope of their critiques, rather than the number they were assigned in the review process.

**Replies to Reviewer 5**

1. *“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.” This includes (a) the use of multilevel modeling (MLM) instead of conventional ANOVAs, which would be appropriate with the small number of subjects, nested conditions, and a balanced design, and (b) the decision to “combine the direction of the report with the confidence when analyzing source accuracy.” Thus, “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.*”

We thank the reviewer for this detailed feedback and apologize for the lack of clarity. We have become enamored with MLM models because of their sensitivity and because, by including *Item* as a random effect, it is possible to control for variability in the memorability of particular stimuli. However, we agree that MLMs are not needed for this relatively simple, balanced design. Moreover, we regret the decision to combine the accuracy and confidence variables into a single measure, which was done to facilitate the MLM analysis—as the reviewer notes, this made it difficult to understand the outcome of the experiment. Thus, we have taken the reviewer’s suggestion and now present separate results for accuracy and confidence, and we have removed the MLMs and replaced them with ANOVAs. We find it easier to parse the results with this data analytic strategy, and we hope the reviewer’s experience will be similar.

To point to the critical behavioral findings, Figures 4A and 4B depict a *Group* x *Cue* interaction for words from the mobility task (left panels), but a main effect of *Cue* for words from the animacy task (right panels). For words from the mobility task, depressed adults showed better accuracy under the Question versus the Side cue, but no such difference was seen in controls. By contrast, both groups showed poorer accuracy under the Question versus the Side cue for words from the animacy task. In our view this pattern clearly parallels what was seen in the ERP data, where Figure 9 shows a group difference in “Question minus Side” ERP amplitude over left parietal sites for words from the **mobility** task; no such effect was seen for the animacy task (see Figure 8, which shows that the “Question minus Side” ERP subtraction for words from the **animacy** task yield a broadly distributed negativity with a frontal focus). The confidence data showed a different pattern. As seen in Figure 5, the depressed group was significantly less confident than the control group under the Side cue for both encoding tasks—there was not the selectivity vis-à-vis encoding tasks that the accuracy data showed. Therefore, we interpret the ERP results shown in Figure 9 as more clearly tied to the accuracy effects shown in Figure 4 (left panels of A and B) than to the confidence data in Figure 5.

2. *“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 the each subject using signal detection or similar approaches . . . although it may be the case that isolating this* [bias] *from an accuracy effect may be impossible without the presence of new items in the design*.”

The review has put his or her finger on the major limitation of the design used in this experiment—namely, the lack of new items. The accuracy data shown in Figure 4A may reflect a bias-free accuracy difference engendered by the two encoding tasks, it may reflect bias to respond “mobility”, or it may reflect some combination of bias and accuracy-related effects. Given the lack of new items, we do not think we can cleanly tease apart these possibilities—and indeed, other experiments that have used similar methods simply report hit rates, as we now do (Bergström, Henson, Taylor, & Simons, 2013; Simons et al., 2008; Simons, Gilbert, et al., 2005; Simons, Owen, Fletcher, & Burgess, 2005). This limitation notwithstanding, we believe we are contributing to the literature by highlighting the influence of the encoding tasks on source accuracy. In the aforementioned studies, retrieval accuracy was analyzed as a function of cue (here, Question vs. Side), but the data were always collapsed across the different encoding tasks. Our results clearly show that the encoding task strongly affected participants’ ability to retrieve conceptual information. Specifically, source accuracy was characterized by a *Cue* x *Task* interaction, *F*(1,46) = 22.35, *p* < 0.001, because Question accuracy was much higher for words from the mobility task vs. words from the animacy task, *t*(47) = 5.65, *p* < 0.001, but Side accuracy was only marginally better for words from the mobility vs. animacy tasks, *t*(47) = 1.95, *p* = 0.057. We now discuss this point and the limitation associated with the lack of new items in the manuscript, on pages X and Y, respectively. Of course the detection of effects associated with unmedicated depression (*Group* x *Cue* interaction for accuracy in response to words from the mobility task; lower confidence on Side trials; modulation of left parietal “Question minus Side” ERPs to mobility words by group status) is another novel contribution of the study. Nonetheless, in a follow-up study that we are currently launching, we are including new items to directly address this limitation.

3. *“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 . . . 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 constitutes a memory or retrieval impairment. This seems inappropriate . . . lower confidence doesn’t mean that less evidence is recovered, nor does it mean that metacognitive resolution or calibration is impaired.”*

We thank the reviewer for this helpful critique, to which we offer a two-part reply. First, we have taken the reviewer’s suggestion and added descriptive and statistical data on confidence (as well as accuracy and RT) for the parity judgment on Odd/Even trials (see page X). Confidence on these trials was uniformly high and not remotely different between the groups, thus there is no evidence for tonically lowered confidence in our MDD sample. Second, by taking the reviewer’s advice and adopting a simpler ANOVA approach to the data, the picture with respect to confidence is hopefully more clear. As shown in Figure 5, the MDD group was less confident than the control group in response to the Side cue but not the Question cue, and this was not specific to either encoding task. Critically, this result does not mirror the accuracy data: the depressed adults were not significantly less accurate than the controls when responding to the Side cue (or the Question cue), and thus the reviewer’s point about lower confidence not tracking lower accuracy is well-taken. Consequently, on page X we now offer a more careful discussion of the confidence data.

4. *“In the discussion . . . the authors suggest that longer reaction times* [at encoding] *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.*”

The reviewer is entirely correct, and his or her critique makes it clear that our reasoning was not plainly stated. We did not intend to suggest a simple relationship between depth of processing and RT. However, given that both of our encoding tasks would typically be considered “deep” tasks, and given that both depend on consideration of the semantic properties of the items, we feel we are on reasonably safe ground concluding that the task that takes longer to complete entails deeper processing. Thus, while we do not wish to imply that RT can be used to sort deep from shallow tasks, as these may differ in many important respects, we do think RT is a fair proxy for depth of processing in tasks that make similar cognitive demands. As precedent for this argument, we note that when Dobbins & Wagner (2005) found slower encoding RTs for pleasant/unpleasant judgments than for animacy judgments, they suggested “that comparison of the former to the latter encoding trials would identify regions differentially engaged during sustained conceptual analysis” (p. 1773). Here, we are similarly arguing that the longer RTs observed for mobility versus animacy judgments in our data implies “sustained conceptual analysis” (i.e., deeper processing) for the former relative to the latter trials. We have attempted to clarify this point on page X.

5. *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 because: (1) only correct responses were analyzed; (2) “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”; and (3) there is no evidence that individual differences in ERP amplitude correspond to individual differences in source accuracy or confidence.*

We appreciate the reviewer’s comments and have re-phrased the problematic passages—please see the new text about X on page Y in the Discussion. Here we would like to draw attention to three additional points. First, please note that our discussion of parietal ERP effects has been simplified. In the original manuscript, we presented two analyses: a “classic” analysis in which we examined a small number of electrodes based on prior literature, and a mass univariate analysis that permits examination of every electrode simultaneously. We reported a group difference (MDD < controls) in parietal ERP amplitudes in the classic analysis, but—critically—the size of this difference did not vary by *Cue* and it was evident on Odd/Even trials. Because Odd/Even trials do not require retrieval from episodic memory, this group difference must not reflect a mechanism specific to episodic retrieval. Furthermore, because our *a priori* contrasts were “Question minus Odd/Even” and “Side minus Odd/Even”, this group difference was not detectable in the mass univariate analysis (i.e., the left parietal amplitude was shifted down in MDD vs. controls, but the differences between conditions were of similar magnitude in these contrasts). Because we are specifically interested on episodic retrieval, and because we argue that the mass univariate analysis is preferable for its sensitivity, we have removed the classic analysis and now only present the mass univariate approach. Consequently, we no longer discuss a depression-related deficit in recollection in as much detail.

Second, we take the reviewer’s point about the lack of strong negative effects of MDD, and thus we now put more emphasis on the link between the parietal ERP effects in Figure 9 and the fact that depressed adults performed surprisingly well in the condition that the ERPs capture (i.e., Question minus Side differences for words from the mobility task). Third, in the revised manuscript (page X) we note that there were positive relationships between left parietal Question minus Side ERP difference waves in the 400-800 ms and 800-1400 ms time windows, on the one hand, and Question minus Side source accuracy and confidence difference scores on the other. These relationships were modest—the strongest was between accuracy and ERP amplitude between 800-1400 ms, *r* = 0.28, *p* = 0.05—but they indicate that there was a relationship between left parietal ERPs and behavior.

6. *The authors acknowledge the lack of group differences when comparing either source task to the parity task, but they discuss the groups separately and this may give the false impression of meaningful group differences—it may not be necessary to plot these contrasts separately by group. Moreover, it may be useful to emphasize that the secondary ERP analysis—which was designed to parallel the behavioral results—was exploratory and thus may reflect Type I error.*

We have taken the reviewer’s advice on all these points and now present the Question minus Odd/Even (Figure 6) and Side minus Odd/Even (Figure 7) contrasts collapsed across the groups. We also added a note about the exploratory nature of the subsequent ERP analyses on page X of the Discussion.

7. *In the secondary ERP analysis, the authors compute Question minus Side difference scores separately for words from the mobility and animacy task. It is difficult to know what these contrasts isolate because, “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.”*

We hope that the basic goal of the secondary ERP analysis is clear in the revised manuscript: our intention is that the ERPs in Figures 8 and 9 should parallel the accuracy subtractions presented in Figure 4B. Most importantly, we are struck by the fact that depressed (but not healthy) adults showed a Question minus Side accuracy advantage for words from the mobility task, and this was paralleled by a *Group* x *Cue* interaction for left parietal ERPs elicited by words from the mobility task from 400-800 ms and 800-1400 ms, as shown in Figure 9. Similarly, both groups showed worse accuracy under the Question versus Side cue for words from the animacy task, and this was paralleled by broadly distributed and temporally sustained negative polarity potentials (Figure 8). We hope that the revised behavioral analysis and the consistent focus on the results from the multivariate analysis approach will make all these parallels clear.

We can only agree with the reviewer’s second point, which is that it is generally difficult (if not impossible) to isolate the contribution of bias to accuracy results in this type of source memory task. As noted in response to point 2 (above), the absence of new items is a clear limitation, and we now discuss that on page X. In fairness, this limitation is present in many prior studies of source memory, and we are extending those prior studies by highlighting the importance of the encoding tasks and by investigating the impact of unmedicated depression. As we noted earlier, we are currently following up on this effort with a new study that will include new words so that we can dissociate bias from accuracy, in line with the reviewer’s point.

8. *The response scale is a bit unusual in that participants are offered the opportunity to guess. “I would be particularly interested in seeing whether the two groups use this non-committal response option similarly across the three tasks” as depressed adults might be expected to guess more often than controls*.

In response to the reviewer’s question, we now present the guessing data in Figure 3. Analysis of these data did not reveal a main effect of *Group* or any interactions involving this factor. Instead, all participants guessed less in response to words from the mobility task (vs. the animacy task) and to the Question cue (vs. the Side cue). No participant guessed on a single Odd/Even trial. As shown in the figure, when we considered each cell of the design on its own, we only found a significant cue effect in one condition: the MDD group guessed significantly less often under the Question cue relative to the Side cue for words from the mobility task. This was not true for words from the animacy task, and it was not true for words from either task when the controls were considered alone. We do not wish to make too much of this result in light of the lack of significant interactions with *Group*. However, we describe it in the paper and highlight it with an asterisk in the figure because we believe it complements the accuracy data shown in Figure 4 and the ERP data shown in Figure 9 by providing further evidence that the combination of especially deep encoding (mobility task) and conceptual retrieval (Question cue) supported good performance in the MDD group. Finally, we note that prior studies of multidimensional source monitoring tasks also offer the “guess” response option (e.g., Starns & Hicks, 2005), and we have added references to those studies on page X.

9. *The design requires participants to switch, on a trial-by-trial basis, between parity judgments, conceptual retrieval, and perceptual retrieval. The authors may be able to analyze source trials as a function of the immediately preceding trials, to see whether there is any evidence indicating that the MDD group has more difficulty reconfiguring their cognitive set*.

We appreciate the reviewer’s suggestion—we had actually tried this kind of analysis prior to submitting the original manuscript. We coded trials as either “switch” or “no-switch” depending on whether the cues on the current and preceding trials matched, and then we ran *Group* x *Switch* ANOVAs on accuracy, confidence, RT, and propensity to guess. We found that confidence was lower (*p* = 0.005), RT was slower (*p* < 0.001), and accuracy was marginally lower (*p* = 0.06) on switch relative to no-switch trials; there was no effect of switching on the guess rate (*p* = 0.21). However, none of these ANOVAs yielded a main effect of *Group* (all *p*s > 0.08) or a *Group* x *Switch* interaction (all *p*s > 0.10). Because of the lack of group differences, and in light of the already large number of figures and tables included in the manuscript, we elected to omit this material. However, we will be happy to add it if the Editor feels it would be worthwhile to do so.

10. *“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.”*

We truly appreciate the reviewer’s thoughtful comment and the obvious effort that went into his or her review. Upon further reflection, we have come to agree with this final and critically important point—in the original manuscript, we overemphasized negative effects of depression on recollection despite rather limited evidence for such effects. We have tried to rectify this in the manuscript. In particular, although we note that that the MDD group was significantly less confident than controls in response to the Side cue, we place the greatest emphasis on the *Group* x *Cue* ANOVA seen for accuracy and left parietal ERPs in response to words from the mobility task. These are the most striking group differences in the study, and if anything they highlight better performance in the MDD group relative to the controls. Encouragingly, although this pattern is not what we expected, it is consistent with a substantial body of behavioral work produced by Paula Hertel and her colleagues. Thus, we interpret our results in light of those data, and suggest that our ERP results provide insight into the relevant neural mechanisms. Finally, given the mission of *Neuropsychologia*, we spend much more time discussing the striking differences in cue effects on accuracy across the two different encoding tasks, as the *Cue* x *Task* interaction was strong (*F*(1,46) = 22.35, *p* < 0.001) and has interesting implications for memory retrieval in healthy as well as depressed adults. We thank the reviewer again for his/her contribution, and hope that the substantially revised manuscript will be judged worthy of publication in *Neuropsychologia*.

**Replies to Reviewer 4**

1. *“The sentence ‘In short, memory retrieval is impaired in depression and enhancing it can bring lasting relief’ is unsubstantiated and needs clarification and references*.

We thank the reader for their careful reading of the manuscript. In response to this critique, we have edited the sentence (p. X) by adding references to work by Mark Williams, Tim Dalgleish, (and anyone else?), Filip Raes, and their colleagues. This work touches on the two topics mentioned in the sentence. First, several studies (i.e., X, Y, Z) provide evidence of retrieval deficits in depression, primarily in the context of autobiographical memory. Second, a more recent body of work has shown that training specifically designed to improve the precision of autobiographical memory can have very positive effects on depressive illness. Importantly, this is not simply due to increased retrieval of positive memories, or to the use of retrieved memories as a means for mood repair. Instead, it seems that enhancing the specificity of memory retrieval may involve enhancements in executive control that have broad benefits on mood and cognitive function in depression. We have attempted to clarify these points in the manuscript (see page X), and we thank the reviewer for highlighting this need.

2. *The authors used neutral stimuli; this may reduce confounds associated with mood congruent encoding, but it also reduces the relevance to depression and enhanced memory for negative material*.

The reviewer is completely correct: by using neutral stimuli, we cannot speak to the emotional memory bias in depression (i.e., enhanced memory for negative stimuli, impaired memory for positive stimuli). This was a conscious trade-off on our part. We are very aware of the emotional memory bias in depression and have conducted empirical work on it (e.g., Dillon, Dobbins, & Pizzagalli 2014; Dillon & Pizzagalli, 2013) and advanced an argument (Dillon, 2015) regarding its neural basis. However, we also know of data linking depression to hypofrontality (e.g., Mayberg, Lewis, Regenold, & Wagner, 1994), and thus we reasoned that we might be able to detect a depression-related memory deficit for neutral material provided the retrieval test used was sufficiently difficult (as successful performance would depend on frontal circuits that might not be adequately active in depression). As the reviewer knows, we were only partially successful: we found the expected group difference (controls > MDD) in confidence for responses to the Side cue, but the robust group difference in accuracy that we expected did not materialize. In the revised manuscript we devote more space to speculating about the lack of stronger group differences in memory accuracy (see page X). Importantly, and speaking directly to the reviewer’s point, in that place we specifically suggest that future functional imaging studies of retrieval use emotional stimuli in order to see the expected memory deficit in MDD. In fact, such studies are already underway in our laboratory.

3. *“It is unclear whether the depressed participants were actually clinically depressed at the time of testing . . . 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 clinical psychiatrist or clinical psychologist?”*

The patients were clinically depressed when they were tested. As described in the manuscript (page X), we used a screening procedure to identify individuals who were in the midst of a Major Depressive Episode and who had a BDI-II score of at least 14 (the published cut-off for minimal depression). On the day of the experiment, every participant was evaluated using the MINI International Neuropsychiatric Interview 6.0, and participants in the MDD group had to meet criteria for a current depressive episode (comorbidity with generalized anxiety, social phobia, and specific phobia was allowed). We also re-administered the BDI-II on the day of the study; our participants were, on average, moderately depressed (BDI-II, mean±S.D. = 25.38±8.69). Importantly, participants in the MDD group were unmedicated. Although studies of unmedicated samples are somewhat rare, our group at the McLean Center for Depression, Anxiety and Stress Research (Director: Diego Pizzagalli) has been recruiting unmedicated MDD samples for many years (D.G. Dillon & Pizzagalli, 2013; Daniel G Dillon, Dobbins, & Pizzagalli, 2014; Pizzagalli et al., 2009). Importantly, we do not recruit medicated adults and then ask them to stop taking their medications. Rather, we find adults who have chosen not to use medication for one reason or another (e.g., because they prefer to rely on psychotherapy). Although recruiting a sizable unmedicated MDD sample can take a significant amount of time, we believe it is worthwhile because otherwise medication use would confound the investigation of group differences.

4. *There was a large number of exclusions* (*n* = 10 for controls, *n* = 2 for depressed), *and the authors have not commented on (a) the fact that more controls than depressed participants were excluded or (b) the rationale for the criteria they used to reject datasets* (18 or more bad channels, artifacts on 50% [or more] of trials).

We thank the reviewer for paying close attention to these points about EEG data quality. As the reviewer noted, there was a notable group difference in the number of datasets excluded; we were surprised that we had to exclude more data from controls than from patients. Exclusions were based entirely on EEG data quality, and we have no ready explanation for the group difference—but it seemed to us that the patients were generally more compliant with instructions and more engaged in the experiment than the controls were, perhaps because the study was described as investigating the neural basis of memory deficits in depression.

Importantly, we used the same criteria (> 18 bad channels, more than 50% of trials contaminated with artifact) to evaluate EEG data quality in both groups. In his 2014 book “An Introduction to the Event-Related Potential Technique”, Steve Luck reports using a threshold of 25% artifacts to reject datasets in his work with college students, but notes that “In our experiments on schizophrenia, we see a lot more artifacts (*in both the patients and the control subjects*), so we exclude subjects for whom more than 50% of trials were rejected [emphasis added]” (p. 210). It was on the basis of this statement that we adopted a criterion of 50% contaminated trials for rejection of datasets, and we have added a reference to Luck’s book on page X to indicate this. Our decision to use 18 bad channels as an additional threshold is not based on a prior precedent, but our work with pilot subjects indicated that for some subjects (e.g., those with very thick hair), signal quality at the base of the head was consistently poor as it was difficult to ensure good contact with the scalp. Interpolation is the common method of dealing with bad channels, and there is no limit to the number of poor quality electrodes that can be interpolated, but every interpolation entails the loss of data at one electrode. We felt that loss of up to ~15% of the data (i.e., 18 of 128 channels) was a reasonable threshold to use before declaring a dataset too noisy to retain. We have added text explaining this rationale on page X of the revised manuscript.

5. “*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 . . . might result in quite large clusters. How does cluster-based permutation control for such assumptions?”*

We thank the reviewer for this excellent set of questions. The mean distance between electrodes in the 128 channel EGI nets that we used is ~3 cm (Song et al., 2015). Because the net tends to stretch out over parietal scalp regions that are of great interest given their known role in recollection, we used 4 cm as our threshold. We also used the mass univariate toolbox program *spatial\_neighbors* to check the consequences of our decision, and it indicated that the mean number of neighboring electrodes with this setting ranged from 1 to 9, depending on the electrode’s location, with mean and median = 6 electrodes. If the reviewer examines the EGI electrode map in the supplement and considers the left parietal electrodes, he or she will see that those are typically surrounded by 6 other electrodes—this means that the 4 cm setting captures the layout of the net without being too generous (e.g., left parietal electrode 42 is surrounded by electrodes 36, 41, 47, 52, 54, and 37, which should all be detectable with our setting). Nonetheless, it is true that we could have chosen a multiple comparisons correction procedure with greater localizing power—for instance, we could have applied a Bonferroni correction, setting a critical alpha value of 0.05/128 = 0.0004 and evaluating every electrode in isolation. Alternatively, Dr. David Groppe (developer of the mass univariate software) has provided a permutation-based procedure that is more powerful than Bonferroni correction but still allows detection of effects at single electrodes (Groppe, Urbach, & Kutas, 2011a, 2011b). Either approach would have improved localization power relative to the cluster-based permutation approach (also developed by Dr. Groppe) that we used.

However, we believe the cluster-based approach is better suited to our data and to ERP studies of episodic retrieval in general. As Groppe and colleagues noted, the cluster-based permutation is “possibly the most powerful mass-univariate procedure for detecting the presence of [broadly distributed] effects” (p. 9), which are the kinds of effects typically seen in studies of episodic retrieval. Indeed, we respectfully suggest that the main issue with mass univariate analysis is not overly large clusters, but rather lower power due to correction for multiple comparisons. For instance, if the reviewer looks at the middle panel of Figure 6, which shows the “Question minus Odd/Even” contrast from 800-1400 ms, he or she will note a swath of positive amplitude activity over left parietal electrodes that is not significant despite *t*-values greater than 2. For instance, electrode 59, which is just posterior to 10-20 position P3, has *t* = 2.75. Considered alone, this electrode would reveal a significant difference, as would several of its neighbors, but because of the correction for testing at 128 electrodes, this electrode is not significant in our cluster-based analysis. Because tests with greater localizing power demand an even higher level of significance for single electrodes, this kind of problem would be worse with any other multiple comparisons correction we could choose. In other words, the cluster-based permutation correction we used maximizes our ability to detect true effects without being overly liberal.

6. ‘*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.”*

We thank the reviewer for this comment. As the reviewer notes, significant activity was seen over left frontal cortex in the “Question minus Side” contrast when participants were confronted with words from the animacy task, but not when words from the mobility task were presented—in the latter case, strong activity over left parietal cortex was observed (left frontal activity was also seen in the “Question minus Odd/Even” contrast). Because the “Question minus Side” accuracy subtractions revealed significantly better performance for words from the mobility versus animacy task in both groups, *t*s > 2.7, *p*s < 0.02, we speculated that the left frontal activity might reflect additional cue elaboration needed to generate candidate memories following poorer encoding, or possibility post-retrieval monitoring or selection. In the revision, we now emphasize that the first mechanism—additional cue elaboration—is more likely, based on a recent functional magnetic resonance imaging (fMRI) paper by Han and colleagues (n.d.). In this paper, Han and colleagues used orthogonal contrasts to dissociate cue elaboration from the need to monitor and select from several candidate memories, and they showed increased left ventrolateral PFC activation for elaboration but not monitoring/selection. Although of course there are pronounced differences between fMRI and ERP research, to our knowledge this work offers the strongest empirical basis for the left frontal activity we observed.

7. *“The main focus of the discussion should be to discuss the value and feasibility of reduced parietal amplitude in the presence of relatively intact behavioural performance. Are the authors speculating that this finding is a vulnerability marker for achieving remission?”*

We appreciate the reviewer’s points. As noted in our response to Reviewer 5’s 5th critique (see above), we now put greater emphasis on the link between relatively intact source memory and increased left parietal ERPs in the Question/mobility condition for the MDD group—this seems to be consistent with what the reviewer is driving at here. Regarding the question about whether these ERP effects can be considered a marker for achieving remission, the answer is “no”—this is a very interesting idea, but because our MDD group was currently depressed we do not think our results speak to the possibility of remission. Given the complexity of the research problem, we feel it is best to be circumspect and focus the discussion on the relationship between the ERPs and the behavior that was collected during the experiment, setting aside the important issue of vulnerability markers (or biomarkers) for studies specifically designed to address those topics.

**Replies to Reviewer 3**

1. *The last two paragraphs of the Introduction are difficult to follow, and it may be better to focus the discussion on psychological constructs (e.g., depth of encoding) first and bring in the neural circuitry second*.

We appreciate this comment and can see how early introduction of neural systems could add complexity to an already complicated discussion of the psychological constructs that mediate encoding and retrieval. In the revision, we have tried to take a more sequential approach of the kind the reviewer suggests—please see page X. The additional space provided by *Neuropsychologia* makes this easier to do, and hopefully the result is a more readable paper.

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.”*

We appreciate this question, which gets to the core of the manuscript. The strong *Cue* x *Task* interaction that we observed (which was further modulated by the presence of a *Group* x *Cue* interaction for words from the mobility task) was unanticipated, as the reviewer notes—in the prior imaging studies of multidimensional source retrieval of which we are aware, there is little to no discussion of dependency on encoding task. Therefore, finding that accuracy under the Question cue varied so strongly by encoding task was striking. We regret any superficial treatment of this result, and in the revised manuscript it is now very clearly emphasized (see pages X and Y). Also, we now spend more time describing the fact that we saw similar patterns not only in the accuracy data and the left parietal ERPs, but also in the guessing data, where the MDD group guessed significantly less in response to words from the mobility task presented under the Question cue (the RT and confidence data do not show the same *Cue* x *Task* effects). We hope that this treatment makes it plain that these task-dependent effects constitute a major contribution of the current study relative to what has already been accomplished in this area.

3. *The authors argue that depressed adults have difficulty with the task, but if this is so it’s not very strong as performance is similar across the groups. There appears to be a stronger link with confidence, which suggests that neural signatures of accuracy and confidence (or “perceived error likelihood”) might be dissociated. “To test these accounts, ERPs could be related to individual differences in behavior.”*

The reviewer’s point is well-taken; upon re-reading the manuscript alongside these reviews, we realize that we overstated the negative impact of depression on performance. Consequently, we have toned down such claims and have tried to offer a more accurate account of the results in the revised manuscript. Regarding the reviewer’s point about confidence, it is true that the depressed group was significantly less confident than the controls in response to the Side cue, and this raises the possibility that the significant group difference in left parietal ERP amplitudes in the Question/mobility condition (Figure 9) could reflect confidence as much as accuracy. We see two ways to try to resolve this issue. The first is simply to ask whether the pattern of ERP results more faithfully reflects the results for accuracy or confidence? Here we think the answer is clear: the ERP results more closely track the accuracy data than the confidence data. Specifically, if one compares the bar graphs at the bottom of Figure 9 with the left panel of Figure 4A, one can see that in both cases the MDD group shows a relative advantage (higher accuracy, more positive left parietal ERPs) for Question versus Side responses, while controls show the opposite: lower accuracy and lower ERP amplitudes for Question versus Side. By contrast, the left panel of Figure 5A shows greater confidence for Question versus Side responses in both groups. Thus, the ERPs appear to track accuracy a bit better than confidence.

The second method is the one the reviewer recommends—one can look for correlations between ERPs and either accuracy or confidence. Here the answer is a bit less clear. We found that “Question minus Side” difference scores for accuracy and confidence in response to words from the mobility task were significantly correlated across the groups, *r*(46) = 0.42, *p* = 0.002. This relationship suggests that it may be difficult to conclusively tease apart these two factors. And indeed, in both time windows where there were significant group effects for left parietal “Question minus Side” ERPs (Figure 9), we found weak correlations with both “Question minus Side” accuracy and confidence difference scores (**400-800 ms**:accuracy, *r* = 0.18, *p* = 0.21; confidence, *r* = 0.27, *p* = 0.06; **800-1400 ms**: accuracy, *r* = 0.28, *p* = 0.05; confidence, *r* = 0.28, *p* = 0.055). Not surprisingly, direct contrasts of the strength of these correlations did not reveal a significant difference in either time window. In summary, our attempt to link the left parietal ERPs to accuracy or confidence based on correlations was unsuccessful: the two behavioral measures are interrelated, and the ERPs appear sensitive to some mixture of both.

Putting these two analyses together, we have tried to advance a cautious line of argument in the revision (see page X). Specifically, we indicate that the ERPs appear to track the accuracy pattern a bit better than the confidence data, but note that the correlational analysis does not suggest a very strong separation. If a more definitive separation of accuracy vs. confidence in MDD can be made, it will have to accomplished in a follow-up study that is better designed to tease apart these two factors.

4. “*Do MDD show less different confidence ratings for correct vs. incorrect responses, compared to controls?”*

This is a very interesting question that we had not thought to ask! In response, we computed a *Group* x *Accuracy* (hit, miss) ANOVA on the percentage of high confidence responses. Unsurprisingly, we found a massive effect of *Accuracy*, *F*(1, 46) = 152.83, *p* < 0.001, as participants were highly confident on a higher percentage of hit (63.92±18.38%) vs. miss (35.60±24.33%) trials. However, neither the main effect of *Group*, *F*(1, 46) = 1.35, *p* = 0.25, nor the *Group* x *Accuracy* interaction, *F* < 1, was significant. Thus, it appears that the separation in confidence levels for hits versus misses was similar in depressed and healthy adults. Thank you for suggesting this interesting analysis.

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

This is another very interesting question by the reviewer, to which we are inclined to answer “Yes”. **STOPPED**

Replies to Reviewer 1

1. *The selection and matching of the HC and MDD groups is a strength. Do the effects still hold when covarying education or IQ*?

Thank you for this question. As you noted, the groups are closely matched and do not differ on years of education or IQ, as estimated by WTAR scores.

**References**