Thoughts:

1. 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 from 5th to 1st because our responses to Reviewer 5 had important consequences for the analysis and so are best laid out first.

**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 simple, balanced design. Moreover, we regret the decision to combine accuracy and confidence into a single measure—as the reviewer notes, this made it difficult to understand the results. Thus, we have taken the reviewer’s suggestion and now present accuracy and confidence separately, 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 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). In particular, if the reviewer compares the left panel of Figure 4A with the bar graphs at the bottom of Figure 9, he or she will notice the correspondence: for both accuracy and ERPs, the effect is Question > Side for the depressed group but Question < Side for controls. 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 seen for accuracy—and the pattern is Question > Side for all participants; the reversed pattern (Question < Side) is not evident in either group. 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 each subject using signal detection or similar approaches . . . although it may be the case that isolating* [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. As shown in Figure 4A, accuracy under the Side cue did not differ strongly between the two encoding tasks, *t*(47) = -1.95, *p* = 0.057, but accuracy under the Question cue was markedly worse for words from the animacy vs. mobility task, *t*(47) = 5.65, *p* < 0.001.This 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 studies that used similar methods simply report hit rates, as we 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 conceptual—but not perceptual—source accuracy. In the aforementioned studies, retrieval accuracy was analyzed as a function of cue (here, Question vs. Side), but the data were collapsed across different encoding tasks. Our results clearly show that the encoding task strongly affected participants’ ability to retrieve conceptual information. 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 this work.

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 controls in response to the Side but not the Question cue, but this was not specific to either encoding task. This does not mirror the accuracy data, which showed a *Group* x *Cue* interaction for words from the mobility task, and which did not show a significant controls > MDD effect under the Side cue (or the Question cue). 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 entails deeper processing. Thus, while we do not wish to imply that RT can be used to sort deep from shallow tasks, which 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, 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 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., left parietal ERP amplitude was reduced 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. 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 studies by highlighting the importance of the encoding tasks and by investigating the impact of unmedicated depression. We are currently following up on this effort with a 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 found a significant cue effect in just 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 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 revision. 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 we saw, and if anything they highlight better performance in the MDD group. 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 more time discussing the striking differences in cue effects on accuracy across the two different encoding tasks, which has implications for memory retrieval in healthy as well as depressed adults. We thank the reviewer again for his/her contribution, and hope the 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 (Dalgleish & Werner-Seidler, 2014). 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 (Dalgleish et al., 2007). 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 (enhanced memory for negative stimuli, impaired memory for positive stimuli). This was a conscious trade-off on our part. We are very aware of biased emotional memory in depression and have conducted relevant empirical work (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 detect a depression-related memory deficit for neutral material provided the retrieval test was sufficiently difficult (as successful performance would depend on frontal circuits that might not inadequate to the task in MDD). 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 expected group difference in accuracy did not materialize. In the revision we devote more space to discussing about the lack of stronger negative effects of depression on memory accuracy (see page X). Speaking directly to the reviewer’s point, in the same place we specifically suggest that future neuroscientific studies of retrieval use emotional stimuli to see the expected memory deficit in MDD. In fact, such studies are 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 in the midst of a Major Depressive Episode with a BDI-II score of at least 14 (the published cut-off for minimal depression). On the day of the EEG session, everyone was evaluated with 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 moderately depressed (BDI-II, mean±S.D. = 25.38±8.69). All participants in the MDD group were unmedicated. Although studies of unmedicated samples are relatively 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, but instead find adults who have chosen not to use medication at all or for a substantial period of time. Although recruiting an unmedicated MDD sample is difficult, it is worthwhile because otherwise medication use confounds 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 attending to the EEG data quality. We were surprised that we had to exclude more data from controls than patients. We have no ready explanation for the group difference, but it seemed to us that 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, > 50% of trials contaminated by artifact) to judge EEG quality in both groups. In his 2014 book “An Introduction to the Event-Related Potential Technique”, Steve Luck reports a threshold of 25% artifacts for rejecting 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” (p. 210). It was on the basis of this statement that we adopted a criterion of 50% contaminated trials for rejection, 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 reflects the fact that for some subjects (e.g., those with very thick hair), signal quality was consistently poor as it was difficult to ensure good contact with the scalp. Interpolation is the method for dealing with bad channels, but every interpolation entails data loss at one electrode. We felt that loss of up to ~15% of the data (i.e., 18 of 128 channels) was a reasonable threshold for rejection. We now explain this 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 these excellent questions. The mean distance between electrodes in the 128 channel EGI nets is ~3 cm (Song et al., 2015). Because the net tends to stretch out over parietal scalp regions that were of 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 overly generous. For example, left parietal electrode 42 is surrounded by 6 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 improved localization relative to the cluster-based permutation approach (also developed by Dr. Groppe) that we used.

However, we believe the cluster-based permutation approach is better suited to our data and to ERP studies of episodic retrieval in general. As Groppe and colleagues (2011a) noted, this approach 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 concern with mass univariate analysis is not large clusters, but rather low power due to strict 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 patch of positive 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 analysis. Because tests with greater localizing power demand an even higher level of significance for single electrodes (analogous to needing a smaller *p*-value for voxelwise vs. cluster-based FWE correction in fMRI), this problem would be worse with any other multiple comparisons correction we could use. In other words, the cluster-based permutation method we used provides appropriate FWE-control while retaining as much power as possible.

Finally, given the concern just raised the reviewer may wonder why we used the mass univariate approach at all—why not simply test the effects at a handful of electrodes as in many prior ERP studies, thus obviating the need for multiple comparisons correction? The reason is that with 128 electrodes and a novel research question (how does depression influence source retrieval?), we did not have strong enough *a priori* hypotheses to guide a comprehensive analysis of the data. We were especially interested in parietal electrodes that have been consistently implicated in recollection, but we were also interested in the late posterior negativity that typically has a posterior focus but also extends over left frontal cortex during conceptual retrieval. The mass univariate approach allowed us to examine all these effect simultaneously, in the context of a principled approach to multiple comparisons correction.

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. 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 words from the mobility task—in the latter case, strong activity over left parietal cortex was observed (left frontal activity was also evident 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. On page X of the revision, we now indicate 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 post-retrieval monitoring and selection, 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, this work offers an empirical basis for linking the left frontal activity we observed to cue elaboration.

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 fifth 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 (see page X)—this seems to be consistent with what the reviewer is driving. Regarding the question about whether these ERP effects can be considered a marker for achieving remission, the answer is “no”—this is an interesting idea, but because our MDD group was currently depressed we do not think our results speak to the remission. Given the complexity of the research problem, we feel it best to be circumspect and focus the discussion on the relationship between the ERPs and the behavior collected in 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 adds complexity to an already involved discussion of the psychological constructs that mediate encoding and retrieval. In the revision, we take a more sequential approach of the kind the reviewer suggests—please see page X.

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 *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 indeed unanticipated—in the prior imaging studies of multidimensional source retrieval of which we are aware, there is little discussion of dependency on encoding tasks. Therefore, seeing accuracy under the Question cue vary so strongly by encoding task was striking. We regret superficial treatment of this result, and in the revised manuscript it is now 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 this clarifies that these task-dependent effects constitute a major contribution of the current study relative to what has already been done.

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 agree that we overstated the negative impact of depression on performance. Consequently, we have toned down such claims and offer a more balanced account of the results in the revision. Regarding the 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 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. 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, not confidence.

The second method is the one the reviewer recommends—one can look for correlations between ERPs and accuracy or confidence. Here the answer is a bit less clear. As described on page X, we found that “Question minus Side”/*mobility* difference scores for accuracy and confidence were significantly correlated across the groups, *r*(46) = 0.42, *p* = 0.002, indicating that it may be difficult to tease apart these two factors. Indeed, in the time windows with significant group effects for left parietal “Question minus Side”/*mobility* ERPs (Figure 9), we found weak correlations with both 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 these correlations did not reveal a significant difference in either time window. In summary, our attempt to link the left parietal ERPs to either accuracy or confidence via 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 advanced a cautious argument in the revision (see page X). Specifically, we indicate that the ERPs track accuracy more closely than confidence, but we also note that the correlations link the ERPs to both accuracy and confidence. A more definitive separation of accuracy and confidence in MDD will have to accomplished in a follow-up study 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 an interesting question 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, the separation in confidence levels for hits versus misses was similar in depressed and healthy adults.

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

This is another interesting question. The specific psychological processes that mediate parietal activity seen during episodic retrieval have been the focus of much discussion, with arguments made for evidence accumulation in the service of decision-making, top-down attention to the products of retrieval searches, and the online representation of retrieved material in an episodic buffer, among many other candidates (Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; Gonzalez et al., 2015; Rugg & Curran, 2007; Rugg & Vilberg, 2013). The key role of the parietal lobes in various forms of mental imagery is also well-known, and the generation of mental images (as opposed to their manipulation) draws heavily on circuits in the left hemisphere (Farah, 1984, 1989), which is where strong retrieval success effects are typically seen (Rugg & Curran, 2007). Consistent with this, in a recent fMRI study in which we instructed healthy and depressed participants to generate mental images in the service of reappraisal, we saw strong activation of left posterior parietal cortex (Daniel Gerard Dillon & Pizzagalli, 2013). Therefore, it seems possible to us that mental imagery might be involved in one or more of the processes proposed to underlie parietal recollection effects (e.g., representation of retrieved content might involve vivid imagery, and directing attention to retrieved content might also involve imagery). Unfortunately, we cannot tease apart these processes, especially as it is now clear that left parietal regions in close proximity—namely, the intraparietal sulcus and the superior parietal lobule—make separable contributions to retrieval (Gonzalez et al., 2015); the contributions of these spatially neighboring regions are not dissociable in scalp-recorded ERPs. Thus, we cannot advance a strong argument but we have added a reference to mental imagery on page X to reflect the reviewer’s interest, which we share.

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 related to the attention proposal mentioned in the Introduction. Either way, this finding is not greatly discussed.”*

A major shortcoming of the original manuscript was that our discussion of the PSQI data was very truncated due to space limitations. In the revision we devote more attention to this result (see page X). We appreciate the reviewer’s suggestion about a link with alpha oscillations, as well as the pointer to Jensen and Mazaheri’s work (Mazaheri & Jensen, 2010; van Dijk, van der Werf, Mazaheri, Medendorp, & Jensen, 2010), which we were not aware of previously. If we understand correctly, the central idea is that stimulus-evoked changes in the amplitude of the peaks or troughs of oscillatory activity can give rise to slow ERPs, which contrasts with both (a) the argument that ongoing oscillatory activity simply cancels out when one forms ERPs and (b) the argument that ERPs are generated via a phase-reset of the ongoing activity. Moreover, Jensen and Mazaheri argue that the alpha rhythm is particularly important for inhibiting task-irrelevant cortical regions in order to enhance performance. Thus, we interpret the reviewer as suggesting that more drowsy depressed participants unhelpfully generate high amplitude alpha rhythms over the left parietal cortex, which might explain the negative correlation between chronic sleep disruption (as measured by PSQI) and the amplitude of the Question minus Side ERP effects (for words from the mobility task) seen over left parietal scalp. This is a fascinating suggestion, and we briefly acknowledge it on page X in the revision. We feel that a brief mention is appropriate, because the idea is complex and properly testing it would depend on conducting a time-frequency analysis, which we have not yet done (and which is not the focus here). Consequently, in our discussion of the PSQI data we embrace the reviewer’s first comment and focus more heavily on the psychological constructs (i.e., drowsiness) than on the underlying neural mechanism. However, we are currently engaging in time frequency analyses and will seek to incorporate this excellent suggestion into that work.

7. “*If mobility vs. animacy has such a big effect in controls, shouldn’t ERPs be presented broken down by this factor too?”*

We provide the information the reviewer is seeking in Figures 8 and 9 of the revised manuscript, which focus on Question versus Side comparisons for words from the animacy and mobility tasks, respectively. For both tasks, we computed “Question minus Side” difference scores and submitted the whole-scalp data to mass univariate analysis. Because we saw no group differences for words from the animacy task, we computed another mass univariate analysis across the groups—this is presented in Figure 8. Directly addressing the reviewer’s question, we plot the waveforms for Question and Side trials from the most significant electrode(s) in each cluster in each time widow. Figure 8 shows that for words from the animacy task, the Question trials were associated with a pronounced and long-lasting negativity relative to the Side trials. By contrast, we found a group difference in the “Question minus Side” subtraction for words from the mobility task in the 400-800 ms and 800-1400 ms intervals, and these are plotted in Figure 9. The key result here is that on Question trials, the MDD group generated a left parietal ERP that is very similar to what was observed in controls, but on Side trials the response in the depressed adults was much weaker. We think that the waveforms provide an important complement to the scalp maps, and we hope the reviewer will agree. Again, we note that the pattern of results across Figures 8 and 9 is very similar to what was observed for source accuracy (a strong *Cue* effect but no group difference for animacy words versus a *Group* x *Cue* interaction for mobility words).

8. “*Were reaction times and error rates auto-correlated across time? Did the statistical models correct for this?”*

Because some reviewers expressed confusion over the nature of the behavioral findings, we replaced the linear mixed models that were used in the original manuscript with separate ANOVAs focused on guessing (Figure 3), accuracy (percent correct; Figure 4), confidence (Figure 5A), and RT (Figure 5B). We did not enter RT as a covariate in any of the other ANOVAs. However, in response to the reviewer’s question, we ran a *Group* x *Cue* (Question, Side) x *Encoding Task* x *Accuracy* (hit, miss) ANOVA on RTs. We found strong effects of *Cue*, *F*(1, 43) = 144.86, and *Accuracy*, *F*(1, 44) = 50.83, reflecting slower RTs on Question vs. Side trials and for misses relative to hits. Critically, no effect involving *Group* approached significance, *F*s < 1.5, *p*s > 0.23. This suggests that the *Group* x *Cue* interaction seen for accuracy in response to words from the mobility task, as well as the group difference in confidence under the Side cue, should not be confounded with group differences in RT. However, we acknowledge that the ANOVA approach does not permit the sensitive analysis of trial-by-trial dynamics afforded by the linear mixed models.

9. “*A voltage change on an ERP does not really reflective ‘activation’ (page 13), or at least this would be controversial.”*

Thank you for this careful reading—we avoid referring to ERPs as indexing “activation” in the revised manuscript.

10. *On page 13, ‘no significant effects were seen in any time window’ is ambiguous, in the Abstract, ‘slasting’ is a typo, and on page 4, ‘loses’ is a typo*.

We appreciate the careful reading and have revised (or cut) the passages in question.

**Replies to Reviewer 2**

1. *The authors showed a Group* x *Cue* x *Encoding Task interaction in the accuracy data and a very similar pattern of effects in the ERP data, but it is not clear that they formally tested for the triple interaction in the ERPs; I would recommend doing this*.

We appreciate the reviewer’s excellent suggestion. Please note that the critical ERPs shown in Figure 5 of the original manuscript, and now presented in Figure 9, actually reflect a two-way *Group* x *Cue* interaction rather than a three-way interaction. As the reviewer noted, we computed Question minus Side difference waves separately for words from the animacy and mobility tasks in each group, thus isolating the *Cue* effect at each level of task and group, and then we computed between-group contrasts at each level of task, observing group differences only for words from the mobility task. Note that we did not compare the *Cue* effects across the two tasks before looking for group differences, which would have probed the three-way interaction. However, we can address this issue in another way—having identified a cluster of left parietal electrodes that show a *Group* x *Cue* effect for words from the mobility task (Figure 9), we can simply extract data for the animacy task from those clusters and then run *Group* x *Cue* x *Task* (mobility, animacy) ANOVAs in each time interval. We did this and found significant *Group* x *Cue* x *Task* interactions for 400-800 ms, *F*(1, 46) = 8.23, *p* = 0.006, and for 800-1400 ms, *F*(1, 46) = 5.09, *p* = 0.03. As described in the manuscript on page X, in both cases the *Group* x *Cue* interaction was significant for words from the mobility task, *Fs* > 14.3, *p*s < 0.0005. By contrast, analysis of responses to words from the animacy task revealed only a main effect of *Group*, *F*(1, 46) = 3.89, *p* = 0.05 from 800-1400 ms due to reduced ERP amplitude in depressed adults. In summary, there was a triple interaction in both intervals, and the *Group* x *Cue* interaction was restricted to words from the mobility task. Given the complexity of the manuscript and the fact that this analysis does not change the conclusions presented, we have opted to omit it. However, if the reviewer and editor feel it should be added to the manuscript, we will be happy to do so.

2. “*Were the ERP and accuracy data correlated within the MDD group?*”

Thank you for this excellent question. When we considered the data across both groups, we found modest correlations between source accuracy, confidence, and left parietal ERP amplitudes for the “Question minus Side” contrast for words from the mobility task (**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* = 0.05), as described on page X. However, when we restricted these correlations to the MDD group, nothing approached significance, all *ps* > 0.12 (see page X). Given the apparently modest relationship among these variables, analysis in the MDD group considered alone is apparently not sufficiently powered to detect it.

3. “*I’d like to see the authors grapple with the role of comorbid anxiety a bit more.” What were the prevalence rates of anxiety disorders in the MDD group? Also, did anxiety—as measured by the MASQ—influence behavior or ERPs? “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?”*

We thank the reviewer for these questions, which have resulted in some new findings. First, assessment with the MINI revealed that, of our 24 depressed adults: two met criteria for GAD in the past 6 months; 2 reported agoraphobia in the past month; 2 reported social anxiety in the past month; 2 reported panic attacks in the last month; and 7 reported having panic attacks at least once in their lifetime. We have added this information in the note to Table 1.

Second, we took the reviewer’s suggestion and investigated individual differences in anxiety as measured by the MASQ. We conceptualized the MASQ-GDA (general distress due to anxiety) and MASQ-AA (anxious arousal) scales as control measures because—as the reviewer likely knows—the classic account is that anxiety may impact priming but it typically leaves episodic memory intact, while the reverse tends to be true in depression (for no effect of anxiety on episodic memory, see also the meta-analysis by Burt, Zembar, & Niederehe, 1995; Williams, Watts, MacLeod, & Mathews, 1988). Thus, we did not expect anxiety to influence source accuracy. However, we discovered that MASQ-GDA (*r*  = -0.42, *p* = 0.04)and MASQ-AA (*r* = -0.47, *p* = 0.02) were negatively related to “Question minus Side”/*mobility* accuracy in the MDD group. These were the only self-report measures that showed a significant relationship with accuracy, but a negative correlation with the BDI-II approached significance, *r* = -0.35, *p* = 0.09. Indeed, when we ran stepwise regressions predicting “Question minus Side”/*mobility* accuracy with BDI-II entered in Step 1 and either MASQ-GDA or MASQ-AA entered in Step 2, neither regression yielded a significant effect for the anxiety measures (χ2 < 2.7, *p*s > 0.11). This is not terribly surprising, as these three self-report measures were highly correlated (*rs* > 0.68, *p*s < 0.0002). We describe these new analyses in the manuscript on page X; although they do not support an argument for a selective effect of anxiety on memory, they certainly indicate that anxiety should be investigated in future studies of memory in psychopathology. Finally, we took the reviewer’s suggestion and looked for a relationship between MASQ-GDA and MASQ-AA and the ERPs plotted in Figure 9, but we did not find any such relationship (|*r*|s < 0.17, *p*s > 0.44).

4. *Why did the authors include a measure of sleep quality in the study? It is a natural fit given the role of disrupted sleep in mood and anxiety disorders, but the specific rationale for measuring sleep in this experiment was not given in enough detail*.

Unfortunately, space limitations made it difficult for us to describe our rationale for including the PSQI in the original manuscript. We do so in the revision on page X. Briefly, because sleep plays an essential role in concentration and episodic memory, we reasoned that fatigue might disrupt performance in our source memory task—and because sleep problems are common in depression (and other psychiatric disorders), we were concerned that group differences in sleep quality might confound any results associated with depression.

5. *In the Introduction, the authors hint at the Group* x *Cue* x *Task interaction for accuracy but do not explicitly describe it until later—it is worth stating it clearly here as well*.

Thank you for this suggestion; we now clarify the nature of the key interactions in the Introduction, on page X.

6. *On page 12 of the Results, the statement that “These data suggest that recollection was strongest under the Question cue and reduced in MDD*” *combines two main effects and reads as though there was an interaction, which there is not*.

The reviewer is correct and we apologize for the lack of clarity. This passage has been revised, please see page X.

7. *I recommend adding effect sizes to Table 1*.

We have taken this suggestion—thank you!

**Replies to Reviewer 1**

1. “*There are a number of strengths: (a) the design of the task to look at deep vs. shallow encoding; (b) the selection and matching of MDD and HC groups (do the effects still hold when covarying IQ or education?); (c) careful analysis structure for the performance and ERP markers; (d) . . . deep encoding improves memory for MDD and light encoding does not. With a few more subjects, this effect could be more strongly interpreted; (e) 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? (f) the slower RTs as justification for ‘deeper’ processing is interesting, but could use some more justification; (g) inclusion of the hit rate analysis showing similar effects.  But why does it seem like mobility is not really working for HCs?”*

We thank the reviewer for these kind words and appreciate the care he or she took with the review. Before moving on to his or her concerns (listed below), we will respond to the questions included among this statement of strengths:

* Regarding **b**, we recomputed the key behavioral analysis (i.e., the *Group* x *Cue* ANOVA on accuracy in response to mobility words) with years of education as a covariate (education is preferable to the WTAR scores because—as noted in Table 1—WTAR scores cannot be computed for non-native English speakers, whereas we have education data for all subjects). Including education as a covariate actually strengthened the key *Group* x *Cue* interaction, which went from *F* = 6.45, *p* = 0.02, to *F* = 8.39, *p* = 0.006.
* Regarding **e**, we now report on correlations between source accuracy and confidence—please see page X. We focus on “Question minus Side”/*mobility* difference scores as this is where the most interesting effects arose. Across the groups, we find a positive correlation between accuracy and confidence, *r* = 0.42, *p* = 0.003. We also found positive correlations between the accuracy and confidence difference scores and the amplitude of left parietal ERP difference scores at the electrodes shown in Figure 9 (**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* = 0.05). As the reviewer can see, these correlations are not negligible but they are not very strong (and one is quite weak). On the basis of these results it is difficult to say whether the ERPs track accuracy or confidence, but we lean towards accuracy based on another piece of evidence: the pattern of ERP results more closely matches the pattern seen with accuracy. For instance, if the reviewer compares the bar graph at the bottom of Figure 9 with the accuracy data in Figure 4A, he or she will see the resemblance: in the MDD group there is a Question > Side effect for accuracy and ERPs, but in controls the pattern is reversed (Question < Side). By contrast, the confidence data show a Question > Side effect for both groups. Clearly this is not conclusive evidence and we do not wish to make too much of it, but on page X we articulate this reason for favoring an accuracy account of the ERP effects.
* Regarding point **f**, the reviewer’s comment (as well as a similar comment from Reviewer 5) made it clear that we did not plainly articulate our idea about RT and depth of encoding. We describe our thinking in more detail on page X of the revision. Briefly, both encoding tasks used in our study are canonical “deep” tasks and presumably engage similar cognitive processes related to semantic analysis of the words. Given the similarity of the tasks, we feel it is reasonable to assume that the task associated with longer RTs probably requires more extensive—i.e., “deeper”—processing. This is similar to an argument advanced by Dobbins & Wagner (2005), who also noted fast RTs for animacy judgments (see our response to Reviewer 5’s question 4). Importantly, we did not mean to suggest that RT could generally be used to separate deep from shallow encoding, as shallow and deep encoding tasks often differ greatly on several dimensions (e.g., emphasizing phonological or perceptual rather than semantic properties of words). But since we have two deep tasks that require similar processes, we think longer RT provides reasonable evidence of deeper processing for mobility vs. animacy judgments.
* Regarding point **g**, the mobility task supported good performance in both groups. For instance, there was a strong effect of *Encoding Task* on guessing, *F*(1, 26) = 12.34, *p* = 0.002,as everyone guessed less frequently to words from the mobility task (6.02±3.89) versus the animacy task (7.46±3.56) (see Figure 3). Similarly, there was an effect of *Encoding Task* on confidence, *F*(1, 46) = 6.72, *p* = 0.01, with a greater percentage of high confidence responses for words from the mobility task (59.68±23.21%) versus the animacy task (56.16±23.68%). The reviewer may be referring to the *Group* x *Cue* interaction for accuracy (Figure 4A) and the left parietal ERPs (Figure 9), but here it is important to note that the key is the group difference in response to the cues: the MDDs benefit from the Question cue vs. the Side cue, whereas the controls show slightly (non-significantly) worse performance under Question vs. Side. In other words, its not that the mobility task did not work for the controls, it is simply that the MDD group did especially well for words from the mobility task shown under the Question cue.

2. *I am concerned with how the memory literature is presented . . . there is a great deal of literature on rote memory recall difficulties in depression and less on autobiographical memory, but these two literatures are mixed in the Introduction in a way that (a) is difficult to follow and (b) does not do justice to the larger topic of rote memory difficulties in MDD*.

We appreciate this careful reading of the manuscript and apologize for the lack of clarity in the Introduction. In the revision, we devote significantly more space to the topic of rote memory difficulty in depression (see page X), and strive to be clear about the precise aspect of memory we are discussing (e.g., see line X on page Y). We agree that a thorough and accurate literature review is key, and we hope the reviewer will agree that this improved review highlights the gap that our study was designed to fill—namely, the lack of quality neuroscientific studies of disrupted episodic retrieval in depression.

4. *The manuscript does not carefully distinguish between recall and recognition within the broader concept of cued/uncued retrieval*.

We apologize for this and now provide important background information on page X.

5. *“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.”*

The reviewer is touching on several important points here. Beginning with the last point first, the reviewer has picked up on a novel finding that is emphasized more heavily in the revision than in the original: the *Group* x *Cue* interaction for accuracy in response to words from the mobility task contrasts with a very strong *Cue* effect but for words from the animacy task. For the mobility task, the cue effect was relatively modest (and in different directions for the MDD and control groups), but for animacy words accuracy was substantially and significantly worse under the Question versus Side cue in both groups. Another way to say this is that retrieval under the Side cue was similar following both encoding tasks, while accuracy under the Question cue was much worse following the animacy task. We believe this is an interesting, novel result because it shows that retrieval of conceptual information (Question cue) was strongly shaped by the encoding task, whereas retrieval of perceptual information (Side cue) was not. This dependency on encoding has not been described in previous imaging studies of multidimensional source memory. We are not yet sure why the result emerges, but we speculate that it reflects weaker encoding in the animacy task together with greater interference on Question versus Side trials: both encoding tasks require semantic processing but the mobility task requires a bit more (judging by the longer encoding RTs), and we suspect that the participants have difficulty sorting out the source of retrieved content under the Question cue and tend to mis-attribute it to the mobility task. The differences in neurocognitive processing at retrieval for the animacy versus mobility tasks are quite evident when one compares Figures 8 and 9, which show the Question minus Side difference waves for the animacy and mobility tasks, respectively. For the animacy task, this subtraction elicits a very strong relative negativity with a fronto-central focus, whereas for the mobility task the ERPs are centered around left parietal cortical regions consistently linked to recollection. Again, without additional data (e.g., fMRI data) we cannot be sure of the neural circuits that give rise to these results, but they are very obviously different and we suspect that the fronto-central negativity observed for the animacy task reflects increased activity in the prefrontal and anterior cingulate cortex that is necessary for dealing with the interference and responding correctly to the Question cue. Again, this is simply speculation and must await a follow-up study using fMRI for confirmation or rejection.

Regarding the first two points raised by the reviewer, it seems that the linear mixed models we presented in the original manuscript caused some confusion. If the reviewer looks at Figure 4A, he or she will see that source accuracy (percent correct) is not at ceiling; it ranges from 65.18±14.74% (Question/*animacy* accuracy in the MDD group) to 82.11±9.13% (Question/*mobility* accuracy in the MDD group). Also, this is not a cued recognition paradigm as there are no new words and no “Old/New” judgments were made. Rather, all the words are old and the task was to retrieve the conceptual and perceptual sources associated with each one—this is essentially cued recall. These two important clarifications notwithstanding, we take the reviewer’s basic point that performance in the MDD group is not as bad as one might have expected—indeed, as we expected. On page X of the revision, we devote space to thinking about why the MDD group performed as well as it did.

6. *“The interpretation of the ERP findings in light of low power—is something important being missed here?”*

We agree with the thrust of the reviewer’s question and there is no doubt this work would be improved with more participants. However, recruiting unmedicated depressed participants is not trivial and our samples (*n* = 24 per group with clean data) are reasonably large as judged by current standards. Also, while we concede that we do not have the power to detect small effects, with these samples we have successfully: (1) identified a strong relationship between conceptual (but not perceptual) retrieval and encoding tasks; (2) ruled out the possibility of large negative effects of MDD on source memory for neutral material, at least in this design; but (3) shown that MDD can nevertheless modulate the relationship between encoding, retrieval, and left parietal activity as measured with ERPs. We believe this provides a useful starting point for future studies in larger samples.

7. “*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.*”

We agree with the reviewer that it would be ideal to find correlations between behavior and ERPs that discriminated between the two groups, but we note that there are correlations between accuracy, confidence, and left parietal ERPs when the groups are considered together (see our response to comment 1e, above). Furthermore, the ERP data shown in Figure 9 (top and middle panels) make it clear that the MDD group generates a similar neural response to controls on Question/*mobility* but not Side/*mobility* trials. Unfortunately, there were insufficient errors to form clean ERPs, but we note that many imaging studies of source memory focus exclusively on correct responses (e.g., Bergström et al., 2013; Dobbins & Wagner, 2005).

8. *“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 were based upon entirely distinct samples – e.g., not a carryover of the same HC group).”*

In the revision we strive to provide a clear account of the background literature, but the reviewer need not be concerned about prior work from our group as this is our first manuscript focused on source memory retrieval in healthy or depressed adults. We also appreciate the need to be careful about replication claims, but note that we claimed to have replicated prior work by Bergström, Henson, Taylor, and Simons (i.e., not our group), so carryover of participants is not a concern. To state it directly, our healthy and depressed participants were specifically recruited for this study and no data from this manuscript have appeared in any other publication.

9. “*Figure 5 lacks some clarity and details . . . 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.*”

We apologize for any lack of clarity. There are several new figures in the revision, and we have done our utmost to clearly indicate the source of the data (e.g., with respect to the groups and time windows for the ERPs).

**References**