To the Editor of BMC-Psychiatry, or….

Dear Dr. Culverhouse and colleagues,

We are delighted to see pre-study publication in *BMC Psychiatry* (Culverhouse et al., 2013) of the research design and plans for meta-analysis of the literature on the interaction between serotonin transporter genotype and life stress, predicting depression (Caspi et al., 2003). We expressed concerns with the planned methodology from the beginning, and in mid-2012 we wrote to the group of researchers involved in the meta-analysis, to express two remaining concerns with the plan (see our letter at: http://www.moffittcaspi.com/sites/moffittcaspi.com/files/Letter_to_Culverhouse_JUNE2012.pdf).

Because these two concerns have not been adequately addressed in the final published protocol, we write again now.

**Issue 1: “Primary Analysis Plan 2” to study lifetime depression does not allow for establishing temporal order between stress and depression.**

The literature contains ample documentation that retrospective recall of lifetime depression is inadequate for research purposes. We reviewed this evidence in *Psychological Medicine* (Moffitt et al., 2010), and since then the inadequacy of retrospective recall of lifetime depression has been demonstrated again (Copeland et al., 2011; Takayanagi et al., 2014). Together these and prior papers show that lifetime reports overlook at least half of depression cases. Thus Plan 2 will wrongly assign individuals who have had depression to the non-depressed outcome group in the meta-analysis. Numerous publications have noted that retrospective recall of lifetime stressful life events is likewise unsuitable for research purposes (Monroe, 2008; Monroe & Reid, 2008). In GxE research, the inaccuracy of retrospective recall is particularly important. Simulation studies reveal that the difference between measurements that are unreliable (correlation with true score = 0.4) vs reliable (0.7) corresponds to a 20-fold difference in sample size. Thus, although measuring environmental exposure is costly, doing it well can pay for itself by reducing sample size (Moffitt et al., 2005).

Setting aside for the moment the inadequate quality of recalled lifetime depression and life events, the main reason that these measures should not be used in the forthcoming meta-analysis is that using lifetime measures precludes establishing temporal order between a hypothesized cause and a hypothesized effect. The hypothesis in question is that individuals with an at-risk serotonin transporter genotype are likely to develop depression after life stress and in response to it. The minimal criterion for a valid test of this hypothesis is a set of measures that can unambiguously establish that the stress came before the depression. When using lifetime measures one cannot simply make
the assumption that stress came before the depression, because there is a literature showing that individuals with depression tend to experience more stressful life events as a consequence of their mood disorder (Kendler et al. 1999). This well-known phenomenon is referred to in the literature as “stress generation” (Hammen, 2004). For example, depressed individuals have elevated rates of intimate partner violence and divorce. To use retrospective reports of lifetime depression is tantamount to using lifetime weight to test hypotheses about the cause of low birth-weight, or to use lifetime IQ to test hypotheses about causes of IQ decline in Alzheimer’s dementia; the measure sounds the same, but it is not. Timing is everything. The importance of temporal order in hypothesis testing in studies having observational designs is nicely explained in a powerpoint lecture "What Do Survey Data Really Mean? Considering Issues of Causality and Temporality in Survey Research," by Seth Noar (http://www.nidcr.nih.gov/Research/DER/BSSRB/PowepointPresentations/default.htm).

The *BMC Psychiatry* methods paper includes two plans: Primary Analysis 2, the original plan to study lifetime depression including all studies available, and Primary Analysis 1, a new separate plan to study only those data sets that can establish temporal order. We applaud the addition of Plan 1. However, we query why Plan 2 is still retained. This point about temporal order is not new; we addressed it empirically in our original *Science* paper reporting the GxE in question (Caspi et al., 2003). In that paper we estimated the GxE effect using a measure of life events that occurred prior to depression and we estimated it again using a measure of life events that occurred after depression. Results showed empirically that unless the stress came before the depression, the GxE finding was not observed. Culverhouse et al. carefully and rightly emphasize the importance of matching the design features of a replication analysis as closely as possible to design features of the original publication. However, Plan 2 not only fails to match the design of the original publication, it includes a design feature that the original publication tested and advised against.

We suspect that Plan 2 is retained only because it offers an attention-getting large sample size. To quote Culverhouse et al. (2013), “Our second set of primary analyses will involve larger sample sizes, including children and adults of all ages. The increase in sample size will result in increased power if there is a broad genetic association between 5-HTTLPR genotypes, stress, and depression. However, this comes at a cost; in these analyses, we give up the opportunity to investigate whether stress preceding depression was a potential cause of the depression, as relative timing of stress and depression may not be known, and thus will not be included in the models.” We anticipate that even if the more focused Analysis 1 (closer replication, smaller N) shows evidence of the interaction, the results of Plan 2 (lacking temporal order, larger N) will be those most likely to be highlighted by the authors, covered in the media, and remembered by readers, because Plan 2’s mega sample size exceeds 30,000 participants.
**Issue 2: The protocol excludes studies with N < 300.**

Discovery science in genetics requires large samples, but hypothesis-testing science does not necessarily. The Culverhouse et al. replication project is not discovery science, it is hypothesis-testing science. In hypothesis-testing science, the consideration of sample size is secondary to more primary considerations of quality of the measures and correctness of design. This order of priorities may be particularly true of hypothesis testing using a meta-analysis approach, as the approach itself provides more than ample sample size. Many of the best-designed studies for testing the GxE hypothesis in question have samples under 300; these smaller studies are significantly more likely to be prospective longitudinal and to utilize face-to-face interviews (Uher and McGuffin, 2010). In particular, studies of medical illness stressors overcome the problems of variable stressors between subjects and inaccurate retrospective assessment that compromise power in many other GxE studies. However these medical-stressor studies are typically small, and as a result the protocol plan has excluded them. Some studies the protocol includes, no matter how large, must be designated unsuitable for this project if their measures of stress and depression are weak, as is common when data must be collected through the post, telephone, or internet to contain costs of assessing a large sample. When it comes to measuring stress and depression, face-to-face clinical interviews have superior precision but are more expensive, usually necessitating smaller samples. Again, Culverhouse et al. have emphasized the importance of matching features of a replication analysis as closely as possible to features of the original published study. The original published study used face-to-face clinical interviews. Thus, the protocol plans to include studies that fail to match the design of the original publication in the key area of measurement, and most such studies have very large Ns. Moreover, as noted above (see issue 1), large-N studies are even more unsuitable if their designs do not allow establishing clear temporal order between hypothetical cause and outcome. The protocol’s over-emphasis on sample size of individual studies, coupled with exclusion of many well-designed studies for testing the hypothesis, is misguided.

The rationale given in the Culverhouse et al. protocol publication for exclusion of small studies is that more small studies have claimed positive findings. They note small-N studies run a risk of publication bias. Such bias emerges when a small-N study with a negative finding is more often “file-drawerered” because it is not deemed rigorous enough to constitute decisive rejection of the null, whereas a small-N study with a positive finding would be more often published because it was able to reject the null despite being under-powered). However, the simple fact that more small studies have obtained positive findings does not by itself constitute evidence of publication bias, particularly when there are systematic differences in quality between small studies and large studies. Moreover, it has been commented before that in relation to this particular GxE
finding, both researchers and editors have been quite keen to publish negative findings. Culverhouse et al. allow unpublished studies to submit data for the meta-analysis, and have trawled for these unpublished studies. As such, requiring N > 300 to prevent the file drawer problem does not seem necessary.

Our point about sample size is not new. We explained it in our American Journal of Psychiatry paper (Caspi et al., 2010), Uher et al. explained it in two publications (Uher & McGuffin, 2008; 2010), and Karg et al. also explained it in their meta-analysis (Karg et al., 2011). Yet, the meta-analysis protocol does not contain a justification of its choice of N=300 as a cut off for study inclusion. Why not 500, why not 200? According to PRISMA guidelines for reporting meta-analyses, those that aspire to be authoritative provide a rationale for their decision points, e.g., “Specify study characteristics used as criteria for eligibility, giving rationale” (http://www.prisma-statement.org/2.1.2%20%20PRISMA%202009%20Checklist.pdf).

Culverhouse et al. include an a priori plan to test for effects of study design features on heterogeneity in findings, and include a list of five design features to be tested. We applaud this approach. However, the list of design features to be tested omits sample size. We find this omission curious because sample size has been at the heart of debate in the literature about prior meta-analyses of this GxE. The heterogeneity analyses proposed by Culverhouse (cross-sectional vs. longitudinal, interview vs. questionnaire, specific stressor vs. undifferentiated stressor) are important analyses to guide the field going forward. Unfortunately, because so many high-quality longitudinal, interview-based, and specific-stressor studies have been excluded by the sample-size restriction, the results of the planned analyses will be difficult to interpret. Excluding small studies instead of testing for their putative bias on findings seems a missed opportunity for the Culverhouse team. In fact, our claim is not really that smaller studies are more desirable. Our claim is that the largest studies are least desirable because they have the worst measurement technology and in many cases have been unable to establish temporal order, which is rather different. Including a test of sample size as a heterogeneity factor could shed light on the veracity of our claim.

These two issues, temporal order and sample size, are not new to observational hypothesis-testing research. They apply to all observational studies, beyond the special case of GxE studies. Other meta-analyses of this GxE hypothesis have made these same methodological mistakes before, and these mistakes have been repeatedly pointed out in published articles in the past five years. As such, the protocol as published seems fundamentally an inexplicably flawed. We regret this missed opportunity to do something better.

Yours, Terrie Moffitt and Avshalom Caspi
References.


Weblinks:


http://www.nidcr.nih.gov/Research/DER/BSSRB/Pow erPointPresentations/default.htm

http://www.prisma-statement.org/2.1.2%20-%20PRISMA%202009%20Checklist.pdf