Reviewer 1 (anonymous)

This registered report will examine whether adversity is related to lower or intact working memory (WM) performance. The authors will isolate variance in performance related to WM capacity from variance in performance related to updating ability as the two measures covary. They will combine existing and new data, and estimate participants' exposure to neighborhood threat, material deprivation, and unpredictability as measures of adversity. Structural equation modeling will be used to analyze the relationship between adversity and WM measures.

The relationship between working memory and the experience of adversity is a topic of great practical and clinical importance, especially in psychology and related fields of research. However, the significance of this as basic research is not clear enough, and there appear to be some major problems, as discussed below.

Thank you.

1. First, to the best of my knowledge, most previous research has supported that adversity is negatively associated with performance on WM tasks and executive function. It is true that some other research also showed no relationship between adversity and WM performance, but only a few studies reported the positive relationship between adversity and WM/executive function. In fact, the authors relied on research by Young and colleagues for this aspect. If the authors wish to test two hypotheses (deficit-based and adaptation-based models), they should cite more research on the adaptation-based models and justify that the research question in this study is worth investigating.

As the reviewer notes, most previous research has supported that adversity is negatively associated with WM performance. The adaptation-based focus on enhancements in EF is relatively new. Yet, there are by now several studies documenting cognitive enhancements not just in WM updating, but also in attention shifting (Please see our response to the next comment below for more details).

As we outline in the introduction, deficit findings in WM are more prevalent in the literature for two possible reasons. First, studies have focused more on impaired than enhanced abilities. In the case of WM, most studies in the deficit literature have focused on WM capacity instead of WM updating (see P. 5 L. 11-18). Second, impairment and adaptation processes may simultaneously operate within the same task. In the case of WM, impairments in WM capacity could lower performance across all WM tasks, even if updating-specific processes are enhanced (see P. 8 L. 9-20).

Research on enhancements in WM updating is scarce, especially compared to the literature on WM impairments. We want to accurately reflect this difference in our theoretical background. Yet, existing enhancement findings are promising and deserve to be investigated. To develop a balanced view of how adversity is associated with performance, the field needs studies that can identify deficit and adaptation processes on a within-person level and compare

their relative strengths.

We now introduce adaptation-based theories more broadly, beyond the specific case of WM, on P. 6 L. 4-18 (new additions in bold):

"Adaptation-based theories assume that developmental processes tailor an individual's cognitive abilities to the unique challenges and opportunities posed by their environment. The link between adversity and cognitive abilities is further assumed to be specific; as different types of adversity (e.g., threat vs. deprivation) pose different challenges, they should (at least in part) shape cognitive abilities in different ways. ***For example, with regards to executive functioning, previous studies have found that people with more exposure to unpredictability (characterized by random variation in adversity exposure over space or time) and threat tend to be better at rapidly shifting their attention between tasks (Fields et al., 2018; Mittal et al., 2015; Steudte-Schmiedgen et al., 2014; Young et al., 2022; but see Nweze et al., 2021).*** *WM updating may be especially adaptive in unpredictable environments. WM updating allows people to maintain an up-to-date overview of the (changing) current state of the environment (Young et al., 2018). Additionally, improved WM updating performance has also been documented for threat exposure (Young et al., 2022). An enhanced WM updating ability could facilitate keeping track of and integrating signals that may potentially signal acutely threatening situations."*

2. In addition, whether their argument can apply to other aspects of executive function is unclear. Previous research consistently showed that adversity can negatively affect children' and adults' executive function performance. Is WM special?

Aside from WM updating, other EF abilities might be enhanced by adversity exposure as well. For example, there is increasing evidence that people with more exposure to adversity are better at shifting their attention between different tasks (a form of cognitive flexibility), compared to people with less exposure to adversity. The ability to rapidly shift attention allows people in adverse conditions to take advantage of fleeting opportunities and rapidly detect threats. Thus, WM updating is not "special", but might be part of a larger repertoire of developmentally-adapted abilities that enable people from adversity to deal with unpredictability and threat.

We now introduce adaptation-based theories more broadly, also focusing on findings relating to attention shifting, on P. 6 L. 9-14:

"For example, with regards to executive functioning, previous studies have found that people with more exposure to unpredictability (characterized by random variation in adversity exposure over space or time) and threat tend to be better at rapidly shifting their attention between tasks (Fields et al., 2018; Mittal et al., 2015; Steudte-Schmiedgen et al., 2014; Young et al., 2022; but see Nweze et al., 2021)."

3. Second, and relatedly, although the authors compare the two hypotheses from the perspective of work memory capacity and updating, there are several other aspects that should be considered to explain different results in previous studies. For example, the two models may differ in the age of the participants and the population. Research supporting the adaptation model (i.e., Young and colleagues' research) has generally included children, whereas research supporting the deficit model has included both adults and children. Furthermore, the adaptation model may only be supported in a specific population (e.g., US), whereas negative effects on WM and executive function are observed globally (Western, Asian, and African).

Differences between previous studies focusing on deficits and adaptations could certainly be the result of age differences. However, we would argue that the differences are smaller than the reviewer claims. While it is true that Young et al. (2022) focused on a sample of young adolescents, Young et al. (2018) observed WM updating enhancements in a sample of adults ranging from early to late adulthood (Mean age ~40). Likewise, a study by Mittal et al. (2015) found attention shifting enhancements in a sample of adults ranging from 18-64 years old. This is in addition to studies that focus on attention shifting in young children and adolescents. Thus, cognitive enhancements have been observed in samples with age groups that are comparable to the current study. We have added the information on sample ages in previous studies on P. 5 L. 21 and P. 6 L. 10

We agree that the current evidence is mostly based on people in Western countries. A notable exception is a study by Nweze et al. (2021), who found better WM performance in a group of institutionalized Nigerian children (9-18 years old), but no differences in attention shifting ability. A limitation of their study is that no explicit measures of adversity were included. Nevertheless, it is possible that adaptation-based theories may only be supported in Western populations, or that different cognitive abilities are adaptive in different cultural contexts (which we deem more likely). These are very important future extensions of this field of research but are outside the scope of the current study.

4. This point is important because the authors will be using data from participants between the ages of 18 and 55. If their results support one model, we cannot determine whether the measures of working memory or the age of the participants/population are critical factors.

As noted above, previous studies have found support for both deficits and enhancements in WM in comparable age groups. We agree, however, that much more work is needed comparing these theories using a wide variety of WM measures and across different age groups. We see our study as a step in this direction. Two strengths of our study are that (1) we are able to compare deficit and adaptation-based theories within the same participants, and that (2) we use multiple WM measures (whereas it is fairly common to use only one WM measure in this type of research). We plan to come back to this issue in the Stage 2 discussion section.

5. Third, the term "adversity" is ill-defined and somewhat vague. It encompasses a very broad range of indicators other than those used in this study. For example, some researchers use socioeconomic status to refer to it, while others refer to family criminal history or abuse, using the Adverse Childhood Experiences questionnaire. The authors must explain how the measures of adversity may affect the relationship between adversity and WM and justify why they chose their measure to address the research questions in this study.

We agree that there are different ways in which adversity has been defined in previous studies. In general, researchers measuring exposure to adversity often use one of two approaches: (1) a cumulative approach (e.g., the Adverse Childhood Experiences questionnaire), which counts the number of adversity exposures, with more distinct exposures meaning a higher level of adversity; (2) a dimensional approach, which focuses on core underlying dimensions of experiences that operate across many different types of exposures. Contemporary dimensional approaches often focus on threat vs. deprivation vs. unpredictability, and have shown that these dimensions have partly distinct effects on various outcomes, such as cognitive abilities.

In this study, we draw from the dimensional approach because it aligns well with one of the core assumptions of adaptation-based theories, namely, that specific types of adversity pose different challenges, and should therefore (at least in part) shape cognitive abilities in different ways (P. 6 L. 6-8). Previous studies have argued, for example, that the ability to rapidly update information in WM is particularly useful in environments that can change often and suddenly, as well as in environments where threats can arise at any moment. In contrast, WM updating does not offer a clear adaptive benefit in deprived, but otherwise predictable environments. Therefore, we expect WM updating to be positively associated with our measures of unpredictability and threat, but not with our measure of deprivation.

We now provide more justification for our operationalization of adversity on P. 4 L. 4-6.:

*"Living in adverse conditions, with prolonged exposure to intense stress, tends to have a profound and enduring impact on cognitive functioning (Farah et al., 2006; Sheridan et al., 2014; 2022)****. Although adversity can be defined in many ways, we follow contemporary models focusing on threat, deprivation, and unpredictability as key dimensions of adversity (Ellis et al., 2009, 2022; Mclaughlin et al., 2016; 2021)."***

We now also provide definitions of threat, deprivation and unpredictability on P. 9 L. 6-14:

"Threat refers to experiences involving the potential for harm imposed by others. We focus on perceived neighborhood violence, the extent to which an individual reports having been exposed to acts of violence in their neighborhood. Deprivation refers to having a low level of resources. We focus on perceived material deprivation, a (perceived) lack of access to material resources. Unpredictability refers to variation in material deprivation over time.

This definition is inspired by, but deviates from the harshness-unpredictability framework, in which unpredictability is defined as stochastic variation in harshness (age-specific rates in morbidity and mortality) over space and time (Ellis et al., 2009, 2022)."

We would like to point out that although it is true that studies sometimes use SES as an indicator of adversity, this is problematic in two ways. First, SES and adversity can operate through different mechanisms (e.g., enrichment vs. physiological stress) and have separable effects on cognition. Moreover, distinct types of adversity (e.g., threat vs. deprivation) may have different effects on cognition. Second, people with low SES have diverse experiences, both positive and negative, even if adversity is more common in this group.

6. Also, for the analyses, they will be calculating composite scores without analyzing the relationship between the measures (e.g., Neighborhood Threat Composite). They should analyze the relationship between the measures, such as confirmatory factor analyses.

In line with comments from both reviewers, we have decided to separate the perceived and objective measures of material deprivation, regardless of their correlation. This decision also reflects recent perspectives that suggest that perceived and objective indices of adversity likely have different effects on cognitive processes (Smith & Pollak, 2021). Hence, our model will include two measures of material deprivation and two measures of unpredictability, based on the measures of perceived scarcity and the income-to-needs ratio, respectively.

We have also updated our procedure for computing composite variables of the various adversity measures (i.e., the three threat measures and the perceived scarcity measure). For the mean level of perceived scarcity, we will first compute the average across time for each item separately. We will then examine correlations between the item averages to assess whether the items are measuring the same construct. If the correlation between all averaged items is equal to or larger than .60 (i.e., indicating a "strong" correlation), then we will compute a uniformly weighted average. If it is lower than .60, we will apply Principal Component Analysis (PCA) to the averaged items and extract only the first principal component score.

For unpredictability in perceived scarcity, we will follow an identical approach, but based on the coefficient of variation across time instead of the mean.

Similarly, for the threat measures (i.e., the two measures of perceived neighborhood crime and the measure of crime victimization), we will first compute the average across time for each measure separately. If the correlation between the averaged measures is equal to or larger than .60 (i.e., indicating a "strong" correlation), then we will compute a uniformly weighted average. If it is lower than .60, we will apply PCA and extract only the first principal component score.

For more detailed information on each construct, see P. 15-19.
Finally, we decided to adjust our measure of unpredictability upon further reflection by calculating it as the "coefficient of variation" (Key et al., 2017; Liu et al., 2021), which is the standard deviation across time divided by the mean (see P. 20 L. 2-8).

Reviewer 2 (Kathryn Bates)

I recommend this manuscript to be accepted at Stage 1 with just a few clarifications.

The rationale is mapped out clearly with a strong theoretical and methodological basis for the study. Research questions and hypotheses are outlined in detail.

Thank you.

1. What is the rationale for creating composite scores for adversity measures? Perceived and objective measures of crime, for example, might show different relationships to WM, and these nuances are lost in a composite score. Explanation of this choice would be useful.

In line with comments from both reviewers, we have decided to separate the perceived and objective measures of material deprivation, regardless of their correlation. This decision also reflects recent perspectives that suggest that perceived and objective indices of adversity likely have different effects on cognitive processes (Smith & Pollak, 2021). Hence, our model will include two measures of material deprivation and two measures of unpredictability, based on the measures of perceived scarcity and the income-to-needs ratio, respectively.

We have also updated our procedure for computing composite variables of the various adversity measures. See our response to comment 6 of Reviewer 1, as well as the corresponding changes on P. 15-19.

The reviewer uses the example of perceived and objective measures of crime, but we prefer to collapse across the measures of perceived neighborhood crime and crime victimization for three reasons. First, to keep the number of adversity measures going into the model manageable. Second, because both measures are self-reported, even though crime victimization asks about distinct events and not perceptions, per se. Regardless, we would argue that the measure of crime victimization is less 'objective' than the income-to-needs ratio. Third, because we do not have hypotheses about differences between perceived neighborhood crime and crime victimization, but see them both as different indicators of threat.

Finally, we decided to adjust our measure of unpredictability upon further reflection by calculating it as the "coefficient of variation" (Key et al., 2017; Liu et al., 2021), which is the standard deviation across time divided by the mean (see P. 19 L. 2-8).

2. In terms of the framework, how will you determine “lowered”, “intact”, and “enhanced”? Arguably you would need a repeated measures design with manipulations to determine whether WM was lowered or enhanced.

We agree that detecting within-person changes in performance over time would require a repeated measures design. However, what we are interested in here is detecting between-person differences as a function of adversity exposure (i.e., whether people with more exposure to adversity tend to show lowered/enhanced/intact performance relative to people with less exposure to adversity). We now explain this in more detail on P. 10 L. 1-10:

“We distinguish between three between-person data patterns: (1) lowered performance, (2) enhanced performance, and (3) intact performance. We define lowered performance as a statistically significant negative association between a type of adversity and WM capacity or updating (irrespective of effect size). We define enhanced performance as a statistically significant positive association between a type of adversity and WM capacity or updating (irrespective of effect size). We define intact performance as an association between a type of adversity and WM capacity or updating that has a standardized effect smaller than 0.1 and larger than -0.1—even if the effect is statistically different from zero—which we will test using Two One-Sided T-Tests (TOST) equivalence testing (see the ‘Primary analyses’ section; Lakens et al., 2018).

3. Moreover, how would you define an intact score?

Intact performance is defined as an association between any type of adversity and ability with a standardized regression coefficient that is statistically smaller than 0.1 *and* statistically larger than -0.1. We formally test this using Two One-Sided T-Tests (TOST) equivalence testing, using -0.1 and 0.1 as equivalence bounds. We consider any effect that reliably falls within this region to reflect practical equivalence, i.e., a difference that is practically equivalent to zero. Note that by using this statistical technique, we would not conclude intact performance by interpreting a non-significant effect as evidence for the absence of an effect. Instead, the TOST test allows us to conclude intact performance based on a significant effect.

We now mention equivalence testing on P. 10 L. 1-10, where we first talk about intact performance (see comment 2). In addition, we have provided more explanation of the TOST procedure on P. 26 L.21 – P. 27 L. 6:

“To statistically test whether small effects are practically equivalent to zero---suggesting intact performance---we will use Two One-Sided T-tests (TOST) equivalence testing (Lakens et al., 2018), using -0.1 and 0.1 as equivalence bounds. TOST equivalence testing allows us to conclude intact performance based on a significant effect, rather than erroneously interpreting a non-significant effect as evidence for the absence of an effect. We consider any effect that reliably falls within this region to reflect practical equivalence, i.e., a between-person difference in performance that is practically equivalent to

zero. TOST provides two p-values, one testing against the upper bound and one testing against the lower bound; we will report the largest of the two p-values."

4. Adjusting the language to reflect regression analyses would be more suitable, e.g., higher threat predicts lower WM capacity.

We agree that the language of *prediction* may be more suitable than the language of *association*. However, in the past, we have received reviewer feedback noting that readers might misinterpret the term *prediction* as denoting a causal relationship. Of course, we do not want to imply causation. For that reason, we prefer to use associative language. However, if the reviewer and recommender prefer predictive language, we are open to changing this throughout the manuscript.

5. The methods are mostly outlined in detail. The authors describe two datasets collected via the LISS with multiple timepoints. Will the 800 participants be randomly selected?

Most of the adversity measures come from the LISS data archive, collected across different studies and at different timepoints. In addition, we measure neighborhood crime and Working Memory in a new study which started in October 2023. At the start of the Method section, we distinguished between these two sources by referring to "the LISS data archive" and "newly collected data". However, we realize we were not using these labels consistently. We have now rewritten this section to distinguish between (1) the LISS data archive or (2) newly collected data. In addition, we have added the following Figure on P. 14 to make clearer which measures where collected at which timepoint:

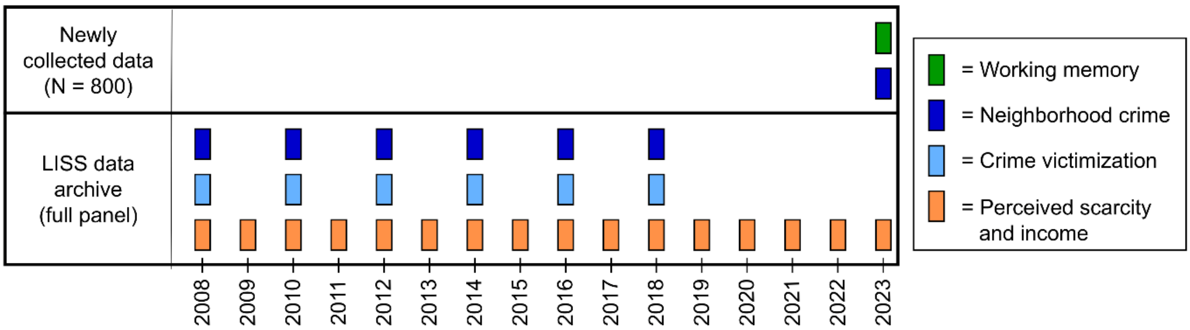

Figure 1. Overview of the different data sources used in this study. We distinguish between measures taken from the LISS data archive and measures that were newly collected in our own study starting in October 2023. Perceived scarcity and income were collected yearly in the full panel from 2008 – 2023. Neighborhood crime and crime victimization were collected across six waves between 2008 and 2018. In the newly collected data in October 2023, we collected data on neighborhood crime and working memory. Note that participants may not have data across all timepoints of the archived studies

because they joined the LISS panel more recently or because they did not participate in each wave.

Thus, the 800 participants were indeed a random selection of people who are currently active in the LISS panel and who had data on at least one wave of the Crime Victimization study in the LISS data archive (the light- and dark-blue bars in the Figure). We have rewritten the Participants section (P. 14) to better explain our sampling procedure.

6. And will they be selected from one or more timepoints? What if a participant has data from all time points and another only has data for one timepoint? This should be explained more clearly.

Participants will inevitably differ in the number of timepoints for which they have data on their level of crime exposure and resource scarcity. Each participant will have at least one wave of crime data, and, as the last wave of crime data was collected in 2018, five years of income data. We average across waves the same way for all participants regardless of the number of waves they participated. We will list this as one of the limitations in the Stage 2 discussion section.

7. It is also not clear what the final sentence (point 4) of the “Data access” section is referring to – what do you mean by later timepoint? It might be helpful to label the timepoints at the start of the method section and use those labels throughout.

We apologize for the confusion. As noted above, we realize that we were not referring to the different data sources using consistent labels, especially in the Data Access section. We have now rewritten this section to distinguish between (1) the LISS data archive or (2) newly collected data (P. 25). We also hope the new Figure makes the distinction clearer.

8. The number of trials is missing from the task descriptions; this is needed for future replication. For the updating task, 18 trials are referred to. Is this total number of trials or number of trials for the updating condition? Number of trials per condition and per task is needed. Can the authors explain why only 18 trials in this case? Is this enough trials to accurately capture performance?

We have now added information on the total number of memory items that participants have to recall on each task in addition to the number of trials (P.20 L. 20; P. 21 L. 13; P. 22 L. 7). Note that each trial consists of a sequence of memory items that participants should remember/update. The number of trials that we use are comparable to previous studies using these tasks (e.g., Young et al., 2018; Wilhelm et al., 2013; see Wilhelm et al. for reliability estimates). Of course, all else equal, having more trials is better in terms of reliability. However, our sample is more diverse in terms of age and SES background than many studies in this field. We were therefore careful to not make the tasks too long and exhausting, especially for people who are already experiencing more stress and might therefore be disadvantaged more by overly long and repetitive procedures.

9. The data analysis plan is well thought out with appropriate steps, control variables, and model fit checks.

Thank you.