Preregistration

Last updated on Monday, December 15, 2025 at 01:08 PM

Overview

The goal of this study is to investigate and compare sources of differential item functioning (DIF) in raw performance measures of executive functioning tasks and computational model parameters of those same tasks. Specifically, we will use the Drift Diffusion Model (DDM), which distinguishes between the efficiency of evidence accumulation, response caution, and non-decision time. The study will be based on data collected within the Longitudinal Internet Studies for the Social Sciences (LISS) panel (see Q12). For information on data availability (which requires a signed agreement with LISS), see Q7.

As this study uses data that have already been collected (although not yet accessed), this preregistration document is based on the template by Akker et al. (2021). There are six parts, which you can jump to following the links below:

Part 1 - Study Information

Q1: Preliminary title

"Does computational modeling yield invariant measures of executive functioning?"

Q2: Authors

- Stefan Vermeent[1]
- Meriah L. Dejoseph[2]
- Dana Miller-Cotto[3]

[1]Department of Evolutionary and Population Biology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands [2]Graduate School of Education, Stanford University, United States [3]Berkeley School of Education, University of California, Berkeley, United States

Q3: Research Questions

- **R1.** To what extent do demographic factors and childhood adversity predict DIF in raw performance measures of executive functioning?
- **R2.** To what extent do demographic factors and childhood adversity predict DIF in EF-specific cognitive processes (i.e., drift rate) versus ability-irrelevant processes (i.e., boundary separation or non-decision time), as derived from DDM?
- **R3.** How do mean-level group differences in EF scores change when comparing unadjusted to DIF-adjusted latent factors?

Q4: Hypotheses

- **H1.** DIF will be more prevalent and severe in measures of raw performance than in computational model parameters, reflecting greater susceptibility of traditional composites to construct-irrelevant influences tied to demographic background and adversity exposure.
- **H2.** When DIF emerges in computational model parameters, it will occur primarily in boundary separation and non-decision time, reflecting differences in response caution and non-decision processes (e.g., preparation and response execution speed), with relatively less DIF emerging in drift rates (reflecting evidence accumulation).
- **H3.** Adjusting for DIF will attenuate observed group differences in latent EF factor means, indicating that some disparities reflect measurement non-invariance rather than true ability differences.
- **H4.** We expect all demographic factors to be associated with at least some DIF in both factor loadings and item intercepts. Our investigation of childhood adversity is more exploratory due to a lack of studies investigating DIF related to adversity exposure

Go back to Overview.

Part 2 - Data Description

Q5: Dataset

For more information on the LISS panel, see https://lissdata.nl and https://website.lisspanel.nl (Dutch only).

We integrate data from the following LISS studies:

1. https://doi.org/10.57990/yrh7-j521 (fielded June 2024). Participants in this study completed all six cognitive tasks and the childhood adversity measures.

2. https://doi.org/10.57990/d9h4-jr36 (fielded July 2025). Participants in this study also completed the cognitive tasks and adversity measures. We will select participants from this study that *did not* participate in the study under (1). Thus, all participants will have data on one time point, either from the wave 1 or wave 2 dataset.
3. https://doi.org/10.57990/n52r-kq87 (fielded October 2023). Some participants in studies (1) and (2) already reported their level of childhood adversity in this study from 2023 (involving other cognitive tasks), and did not do so again in study (1) or (2). For these participants, we take their data on childhood adversity exposure from this study and merge it with their later cognitive data.

Demographic information (age, education, urbanicity) is collected in the monthly LISS background section https://doi.org/10.57990/qn3k-as78, which is a collection of basic information that panel members can update as necessary each time that they take part in a LISS study. The demographic variables used in this study are automatically appended to the data of studies (1-3) by LISS.

Final sample: Some people who participated in the cognitive studies (1) and (2) are from the same household, which violates the assumption of independence. We will randomly sample one participant per household and exclude the other members of the household. This should yield a sample size (prior to further exclusions) of N = 1000-1200.

Q6: Public Availability

All LISS data used in this project will be made openly available in the LISS data archive at the links listed under Q5 (note that at the time of submitting this preregistration, LISS was still processing the 2023 and 2024 data for publication in the archive). Researchers who want to access the data are required to sign a statement confirming that information about individual persons, households, etc., will not be released to others (go to https://statements.centerdata.nl for more information).

Q7: Data Access

Data will become available in the LISS data archive at https://www.dataarchive.lissdata.nl/study-units/view/1. At the time of submitting this preregistration, the wave 1 data are openly available (see DOI under Q5), but the other two studies are still being processed by LISS.

Q8: Date of Download

- Stefan Vermeent: Will access data after timestamping the preregistration.
- Meriah DeJoseph: Will not download/access data
- Dana Miller-Cotto: Will not download/access data

Q9: Data Collection

General information about recruitment of participants into the LISS study can be found at https://www.lissdata.nl/methodology. The inclusion criteria for both waves were that people are between 18 and 55 years old and have given permission for their data to be linked to data from the Dutch Central Bureau of Statistics (CBS; not relevant for this study). For this project, we select participants who participated in both waves.

Q10: Codebooks

Codebooks for the newly collected data will become available later at https://www.dataarchive.lissdata.nl/study-units/view and https://github.com/StefanVermeent/ddm_dif.

Go back to Overview.

Part 3 - Variables

Q11: Manipulated Variables

Not applicable

Q12: Measured Variables

MNLFA moderators

To measure potential DIF resulting from adversity exposure, we will include two self-reported measures:

Childhood deprivation. Perceived scarcity scale (Young et al. 2022), consisting of four items measuring perceived exposure to material deprivation prior to age 18 (e.g., "My family had enough money to afford the kind of food we all needed), rated on a scale of 1 ("completely disagree") to 7 ("completely agree"). We will compute an unweighted average of the seven items.

Childhood threat. Neighborhood Violence Scale (NVS; Frankenhuis and Bijlstra 2018; Frankenhuis, Young, and Ellis 2020), consisting of seven items measuring perceived exposure to neighborhood violence prior to age 18 (e.g., "Crime was common in the neighborhood where I grew up"), rated on a scale of 1 ("Completely disagree") to 7 ("Completely agree"). We will compute an unweighted average of the seven items.

To measure potential DIF resulting from demographic factors, we will include three measures:

Education. Measured as a LISS background variable, with six categories: (1) primary education, (2) Prevocational secondary education, (3) Senior years of senior general secondary education/pre-university secondary education, (4) secondary vocational education, (5) Higher vocational education, (6) University degree. We will treat education as a continuous variable (centered).

Urban character of place of residence. Measured as LISS background variable, based on the surrounding address density per km^2. One of five categories: (1) Extremely urban (2,500 or more); (2) Very urban (1,500 to 2,500); (3) Moderately urban (1,000 to 1,500); (4) Slightly urban (500 to 1,000); and (5) Not urban (less than 500). We will treat urbanicity as a continuous variable (centered).

Age. Measured as LISS background variable, in years (centered).

Cognitive measures

Raw performance. For the Flanker and Simon tasks, we will compute difference scores by subtracting the average response time on correct incongruent trials from the average response time on correct congruent trials. For the Color-shape, Global-local, and Animacy-size tasks, we will compute switch costs by subtracting the average response time on switch trials from the average response time on repeat trials (after removing incorrect trials and trials following an incorrect response).

DDM parameters. The DDM will be fit separately to each task in each wave. We will use individual-level hierarchical Bayesian estimation methods, with default broad and uninformative priors. For the inhibition and attention-shifting tasks, drift rate, boundary separation, and non-decision time will be allowed to vary across conditions. For the Posner task, we will estimate a single drift rate, boundary separation, and non-decision time across all trials. We will assess model convergence visually (through trace plots) and by computing the Gelman-Rubin statistic, which should be below 1.1. In addition, we will also use a simulation-based approach to compare model-implied response times and accuracy with empirical response times and accuracy.

Q14: Missing data

We will use Full Information Maximum Likelihood (FIML) to handle missing data in the SEM indicator variables. We will impute missing data in the MNLFA moderators using predictive mean matching.

Q15: Outliers

For each cognitive task, I will first remove responses longer than 10s. Next, we will remove:

- Trials with response times (RTs) < 250 ms
- Trials with RTs > 3 SD above the participant-level average log-transformed mean RT, separately for different task conditions (e.g., congruent and incongruent)
- Task data for which a participant performed at chance level (using the accuracy rate at the 97.5% tail of a binomial distribution if a participant would be purely guessing).

Q16: Sample Weights

Not applicable

Go back to Overview.

Part 4 - Knowledge of Data

Q17: Relevant Publications

1. Vermeent, S., Schubert, A.-L., DeJoseph, M.L., Denissen, J.J.A., van Gelder, J.-L., & Frankenhuis, W.E. (2025). Inconclusive evidence for associations between adverse experiences in adulthood and working memory performance. *Royal Society Open Science, 12*(1), 241837. https://doi.org/10.1098/rsos.241837

2. Vermeent, S., Schubert, A.-L., & Frankenhuis, W.E. (2025). Adversity is associated with lower general processing rather than executive functioning. *Journal of Experimental Psychology: General, 154*(11), 3010-3028. https://doi.org/10.1037/xge0001812

Q18: Prior Knowledge

In the first previous study (Vermeent, Schubert, Dejoseph, et al., 2025; see Q17) we used data from the study that was fielded in 2023 (wave 1 and wave 2—which are the focus of the current study—were not yet collected at the time the study was completed). The main independent variables (recent rather than childhood exposures to adversity) and dependent variables (performance on working memory tasks) were different from those included in the current study.

In the second previous study (Vermeent, Schubert, & Frankenhuis, 2025; see Q17), we used wave 1 included in the current study to investigate how recent and childhood adversity are associated with task-general and task-specific DDM parameters. From this study, we know that all three DDM parameters form their own coherent latent factor across all tasks and task conditions. We also know that childhood adversity is negatively associated with task-general drift rate as well as task-specific residual drift rates, and that childhood adversity was not associated with other DDM parameters. We did not investigate measurement invariance in this study, nor did we analyze raw performance measures. We completed the study before collecting the data of wave 2.

Go back to Overview.

Part 5 - Analyses

Q19: Hypotheses -> Statistical Tests

Establish factor structure

- **Raw performance:** We will compare two models: (1) a diversity EF model, in which we separately estimate an Inhibition factor and an Attention-shifting factor; (2) a unity EF model, in which all performance measures load on a single EF factor. We will compare these models in terms of the change in the Akaike Information Criterion (AIC). Based on previous work (Vermeent, Schubert, and Frankenhuis 2025), we expect the unity model to provide the best fit.
- *DDM parameters:* From previous work (Vermeent, Schubert, and Frankenhuis 2025), we know that all three DDM parameters (drift rate, boundary separation, and non-decision time) load on their own, single latent factor. Thus, we will estimate a single latent factor for each parameter loading on all manifest parameter estimates. To simplify the MNLFA procedure, we will start by constructing separate latent models for each parameter.

Selected models should have at least acceptable fit, assessed as a root-mean-square error of approximation (RMSEA) $< .08$, and a comparative fit index (CFI) $> .90$.

Establish configural invariance

Prior to the MNLFA analyses, we will test for configural invariance across different levels of each moderator. We will do so using the 'group' input in the *lavaan::cfa()* function. For education and urbanicity, we will assess configural invariance in each group. For the other variables, we will create categorical groups to allow for configural invariance testing. For childhood threat and deprivation, we will use a median split. For age, we will create a group of participants $<$ 45 years and a group $>$ 45 years, based on work showing that EF performance starts to decline around that age. We will apply the same fit indice cut-offs as listed above.

MNLFA analyses

We will fit all MNLFA models using the OpenMx R package (Boker et al. 2011), adopting code from (Kolbe et al. 2024).

Step 1. Assess full measurement invariance. To start, we will conduct an omnibus test for full measurement invariance, comparing an unconstrained configural model with a constrained scalar model.

- **Configural model:** freely estimate effects of all moderators on item intercepts and factor loadings; constrain mean impact to zero (for identifiability). Note that in all other models, the mean impact of moderators on latent factors will be freely estimated.
- **Scalar model:** Constrain effects of all moderators on item intercepts and factor loadings to zero; freely estimate mean impact. the effects of all moderators on item intercepts and factor loadings are constrained to zero, while freely estimating impact paths.
- **Test:** Model comparison through likelihood ratio test (`OpenMx::mxCompare(fitConfig, fitScalar)`. If the scalar model provides a significantly worse fit to the data than the configural model (at $\alpha = .05$), we will conclude that there is evidence for DIF and will proceed to the next steps.

Step 2. Select anchor indicators. Next, we will select anchor indicators, which will serve as (relatively) DIF-free reference points to ensure the stability of the measurement model.

- **All-but-one models:** We will fit a series of all-but-one models, in which only the DIF paths (intercept and loading) of a single indicator are freely estimated, while constraining all others to zero.
- **Test:** Compare each model with the scalar model through likelihood ratio tests.
- **Select anchors:** Per latent factor, we will select two indicators with the lowest DIF (in terms of X^2 test statistics) using a rank-based strategy (Woods 2009), following their recommendation to reserve around 20% of indicators as anchors.

Step 3. Assess partial invariance. We will assess partial invariance for each individual DIF path.

- **Anchor-only model:** Freely estimate all DIF paths, except for those involving the anchor indicators.
- **Anchors-plus-one models:** Fit a series of models that iteratively constrain the effect of a single moderator on a single model parameter (i.e., intercept or loading), in addition to the anchor variables).
- **Test:** Compare each of the anchors-plus-one models to the anchors-only model through likelihood ratio tests. We will retain all DIF paths with a significance level below $\alpha = .10$.

Step 4. Fit final models.

- **Final DIF selection raw performance model:** Fit a single model that simultaneously estimates the significant DIF paths from the previous step. We will apply a Benjamini Hochberg correction to the p-values of these moderation effects, and only retain the DIF paths with an adjusted p-value below .05.
- **Final DIF selection DDM model(s):** First, we will attempt to fit a model that combines the unidimensional factor models (for each DDM parameter) into a single model. Thus, while the respective factor structure for each DDM parameter will be identical to previous steps, they will be combined in a single model, with freely estimated covariances between the latent factors). This model would account for typical covariances between parameters, which could affect DIF. However, if such a combined model does not converge, we will fit final models for the separate DDM parameters (identical to previous steps).
- **Final models:** For both raw performance and DDM models, we will refit the final models including only the DIF paths that were found to be significant at an adjusted alpha-level of .05. These final models will be used for answering our hypotheses.

Q20: Predicted effect sizes

We do not have strong *a priori* predictions for effect sizes, nor do we have specified cut-offs for which effect sizes will be deemed meaningful. Because multiple factor loadings/intercepts can show DIF, the combined effect of small sources of DIF can add up in the latent factor estimation.

Q21: Statistical Power

Precise sample size requirements for MNLFA currently do not exist. However, some initial evidence shows that DIF estimation stabilizes around N = 200-500. Given that our factor models are relatively simple, we expect our sample size to be sufficiently large to detect DIF.

Q22: Inferential Criteria

Here, we specify how (non-)support for each hypothesis will be determined. A prerequisite of support for all hypotheses is that there is evidence of DIF in at least the raw performance model, which holds if the scalar model provides a significantly worse fit to the data than the configural model.

H1. DIF will be more prevalent and severe in measures of raw performance than in the drift rate parameter, reflecting greater susceptibility of traditional composites to EF-irrelevant influences tied to demographic background and adversity exposure.

- **Full support**: A higher quantity of significant DIF paths in the raw performance model compared to the drift rate model, AND standardized regression coefficients of moderators on item intercepts and factor loadings are larger in the raw performance model compared to those in the drift rate model.
- **(partial) Non-support**: A lower quantity of significant DIF paths in the raw performance model compared to the drift rate model, OR standardized regression coefficients of moderators on item intercepts and factor loadings are smaller in the raw performance model compared to those in the drift rate model.

H2. When DIF emerges in computational model parameters, it will more prevalent and severe in boundary separation and non-decision time, reflecting differences in response caution and non-decision processes (e.g., preparation and response execution speed), compared to drift rates (reflecting evidence accumulation).

- **Full support**: A higher quantity of significant DIF paths in the boundary separation and/or non-decision time model compared to the drift rate model, AND standardized regression coefficients of moderators on item intercepts and factor loadings are larger in the boundary separation and/or non-decision time model compared to those in the drift rate model.
- **(partial) Non-support**: A lower quantity of significant DIF paths in the boundary separation and/or non-decision time model compared to the drift rate model, OR standardized regression coefficients of moderators on item intercepts and factor loadings are smaller in the boundary separation and/or non-decision time model compared to those in the drift rate model.

H3. Adjusting for DIF will attenuate observed group differences in latent EF factor means, indicating that some disparities reflect measurement non-invariance rather than true ability differences.

- **Full support**: Estimates of the impact of demographic and adversity moderators on latent factor means are smaller in the final model (which accounts for significant DIF) compared to the scalar model (which constrains all DIF paths to zero), and their 95% confidence intervals do not overlap between models.
- **(partial) Non-support**: The impact estimates on latent means are not significantly smaller in the final model compared to the scalar models (i.e., their 95% confidence intervals overlap or the estimated impact is higher in the final model).

H4. We expect all demographic factors to be associated with at least some DIF in both factor loadings and item intercepts. Our investigation of childhood adversity is more exploratory due to a lack of studies investigating DIF related to adversity exposure

- **Full support**: At least one or more DIF paths involving each demographic factor is significant either within the raw performance model, and/or in one of the drift diffusion models.

- **(partial) Non-support**: One or more of the demographic factors do not show any DIF in the raw performance model nor the drift diffusion models.

Q23: Assumption Violations/Model Non-Convergence

There are three steps in which non-convergence could be a potential issue: (1) DDM estimation, (2) establishing the basic factor structure, (3) estimating MNLFA models.

DDM estimation. Non-convergence in the DDM models is possible but unlikely, given that we were previously successful at fitting these models to task data in wave 1. However, if we do run into non-convergence, we will first increase the number of (burn-in) samples. If that does not solve the issue, we will investigate which parameter(s) are causing non-convergence, and re-specify them to improve model convergence (e.g., estimating a single parameter across trials, instead of estimating different parameter values for each condition).

Basic factor structure. Non-convergence in the initial SEM models is also possible but unlikely, for the same reason. In the case of non-convergence, we will inspect the model to see which indicator(s) are causing non-convergence (e.g., by inspecting Heywood cases and modification indices), and will re-specify the model accordingly.

MNLFA models. Non-convergence poses a relatively bigger risk in the MNLFA analyses given the large number of parameters to be estimated. If MNLFA models including all preregistered moderators fail to converge, we will start with omnibus tests for DIF for each moderator separately. We will then drop the moderator that shows the smallest decrement in model fit, and reassess convergence when including all remaining moderators. If necessary, we will repeat these steps until the MNLFA models converge.

Q24: Reliability and Robustness Testing

See Q19: Hypotheses -> Statistical Tests for DDM model fit assessments.

Q25: Exploratory Analyses

The mean impact analyses combine demographic and adversity measures in the same model. We will interpret these effects as conditional effects rather than direct (causal) estimates. As a sensitivity analysis, we will refit the final models including (1) only the impact paths of demographic variables, and (2) only the impact paths of adversity variables, and compare the effect sizes with those in the full model.

Go back to Overview.

Part 6 - Statement of Integrity

We state that we filled out this preregistration to the best of our knowledge and that no other preregistration exists pertaining to the same hypotheses and dataset.

References

Akker, Olmo R. van den, Sara Weston, Lorne Campbell, Bill Chopik, Rodica Damian, Pamela Davis-Kean, Andrew Hall, et al. 2021. "Preregistration of Secondary Data Analysis: A Template and Tutorial." *Meta-Psychology* 5 (November). https://doi.org/10.15626/MP.2020.2625.

Boker, Steven, Michael Neale, Hermine Maes, Michael Wilde, Michael Spiegel, Timothy Brick, Jeffrey Spies, et al. 2011. "OpenMx: An Open Source Extended Structural Equation Modeling Framework." *Psychometrika* 76 (2): 306–17. https://doi.org/10.1007/s11336-010-9200-6.

Frankenhuis, Willem E., and Gijsbert Bijlstra. 2018. "Does Exposure to Hostile Environments Predict Enhanced Emotion Detection?" Edited by Rolf Zwaan and Christopher Madan. *Collabra: Psychology* 4 (1): 18. https://doi.org/10.1525/collabra.127.

Frankenhuis, Willem E., Ethan S. Young, and Bruce J. Ellis. 2020. "The Hidden Talents Approach: Theoretical and Methodological Challenges." *Trends in Cognitive Sciences* 24 (7): 569–81. https://doi.org/10.1016/j.tics.2020.03.007.

Kolbe, Laura, Dylan Molenaar, Suzanne Jak, and Terrence D. Jorgensen. 2024. "Assessing Measurement Invariance with Moderated Nonlinear Factor Analysis Using the R Package OpenMx." *Psychological Methods* 29 (2): 388–406. https://doi.org/10.1037/met0000501.

Vermeent, S., A.-L. Schubert, and W. E. Frankenhuis. 2025. "Adversity Is Associated with Lower General Processing Rather Than Executive Functioning." *Journal of Experimental Psychology: General* 154 (11): 3010–28. https://doi.org/https://doi.org/10.1037/xge0001812.

Woods, Carol M. 2009. "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement* 33 (1): 42–57. https://doi.org/10.1177/0146621607314044.

Young, Ethan S., Willem E. Frankenhuis, Danielle J. DelPriore, and Bruce J. Ellis. 2022. "Hidden Talents in Context: Cognitive Performance with Abstract Versus Ecological Stimuli Among Adversity-Exposed Youth." *Child Development*, 1493–1510. https://doi.org/10.1111/cdev.13766.