# 4 Data collection

To address our research question, 5 independent experiments were conducted on each participant. The first of which - the *Wisconsin Card Sorting Task (WCST)* - is the primary experiment of interest: assess the subjects ability to learn the reward associated with unique stimuli.

The experiments, described above, were implemented sequentially as follows:

1. Wisconsin card sorting task (WCST)
2. N-back task
3. Corsi block span task
4. Backward Corsi block span task
5. Navon task

The experiments were conducted utilising Psytoolkit - a research first state-of-the-art free Linux distribution for conducting Psychological research [21]. Participants were chosen to represent a diverse population of individuals and demographics. The experiments were conducted online.

# 5 Data encoding

The high level objective of this analysis is to use both an individual's demographics data & performance on other executive function tasks to model their performance in WCST. Natural hierarchies exist, thus Bayesian methods are used to capture the variation over (nested) groups in explanatory variables. The primary consideration is whether or not to represent the response data as:

**A.** Polynomial categorical choice data.
**B.** Batch trails to measure the percentage of correct sco

The former represents the data in it's true form, and will be modeled by estimating the state-value pairs (in an RL fashion) as done in the related literature. The latter - inspired by the model-free analysis - captures each group of trials as a accuracy percentage, continuous values between 0 and 1, thus allowing us to model the data as spatial-temporal profiles (most suitable for Bayesian Regression or Gaussian Processes).

The former is chosen as it more closely relates to existing literature, as well as offering the aforementioned neuroscience theoretical interpretations.

## 5.1 Response Variable

The response data is captured as:

$$Y_p^t = \text{WCST performance for participant } p \text{ at time } t \tag{1}$$

If we exclude approach B. from above (batching trails average performance bins), there are two plausible representations for this response variable:

$$Y_p^t = \begin{cases} \text{Two polynomial dummy variables capturing the matching rule employed.} \\ \text{A binary indicator corresponding to the feedback received.} \end{cases} \tag{2}$$

Mathematically:

$$Y_p^t = \begin{cases} \{color, shape, number\} & for\ p \in [1, P],\ t \in [0, 100] \\ \in [0, 1] & for\ p \in [1, P],\ t \in [0, 100] \end{cases} \tag{3}$$

The subject does not have access to the correct choice, but only receives binary feedback. It is thus more natural to encode the response as a binary indicator variable. This also conveniently directly relates to the RL theory discussed in the previous chapter.

## 5.2 Independent variables

Our study consists of multiple distinct experiments & as a consequence we have a plethora of rich, detailed, data sources that require pruning in order to fit a parsimonious model.

The following subscript and superscript are employed throughout:

$$\begin{aligned} p &: participant \\ t &: time \end{aligned} \tag{4}$$

### 5.2.1 Demographics data

Participant demographic information is available:

$$\begin{array}{rll} X^a_{p,t} : & age & numeric \\ X^g_{p,t} : & gender & nominal \\ X^h_{p,t} : & handedness & nominal \\ X^i_{p,t} : & income & ordinal \\ X^e_{p,t} : & education\ level & ordinal \\ X^c_{p,t} : & computer\ hours & ordinal \\ X^{d-rt}_{p,t} : & response\ time\ to\ complete\ the\ demographics\ questions & ordinal \end{array}$$

Whilst they may be in the analysis - either as explanatory variables in the observation model or to pool variance across sub-populations - they are auxiliary and considered secondary to the neurological and psychological information.

### 5.2.2 Neuropsychological data

More saliently, a number of auxiliary neuropsychology experiments were conducted to gauge different executive functions for each participant. Each experiment offers verbose metadata, & needs to be summarised to represent the performance of each experiment.

The follow summarized experimental data was extracted:

$$\begin{array}{rll}
X_{p,t}^{nback} : & N - Back\ task\ accuracy & numeric \\
X_{p,t}^{nback-rt} : & N - Back\ reaction\ time & numeric \\
X_{p,t}^{fitts} : & Fitts\ task\ accuracy & numeric \\
X_{p,t}^{corsi} : & Corsi\ span\ achieved & nominal \\
X_{p,t}^{navon} : & Navon\ task\ accuracy & numeric \\
X_{p,t}^{navon-rt} : & Navon\ reaction\ time & numeric \\
X_{p,t}^{wcst-rt} : & WCST\ reaction\ time & numeric
\end{array}$$

### 5.2.3 Theoretical biological parameters

Additional to the raw data, there are some theoretical (biological) parameters - informed by the literature: capturing the learning rate $\alpha$ and exploratory temperature parameter $\beta$ - that require unique identification:

$$\begin{aligned}
\alpha &: learning\ rate \\
\beta &: exploration\ parameter
\end{aligned}$$

These are the most salient parameters in the analysis: constituting the learning model and offering direct psychological interpretations. These parameters are discussed at length in the methodology chapter.

## 5.3 General Abstract Model

The covariates can be combined in some design matrix. Similarly, the response can vectorized.

$$\begin{aligned}
\mathbf{X} &: \{X_{p,t}^{a}, X_{p,t}^{g}, ..., X_{p,t}^{wcst-rt}\} \\
\mathbf{Y} &: \{y_1^1, y_1^2, ..., y_p^t\}
\end{aligned} \tag{5}$$

Thus any model is some special case attempting to approximate the true unknown function $f$:

$$\mathbf{Y} \sim f(\mathbf{X})$$

# 6 Outlier removal

## 6.1 WCST

In preprocessing the data, outliers ought to be removed from our primary experiment: WCST. In addition to incomplete data being removed, subjects with significantly low scores ought to be excluded. Our study does not consider individuals with severe mental deficits or impairments, and as such it is exceedingly unlikely that healthy individuals achieve very low score in the WCST - given it's simplicity - and thus low scores are more likely a consequence of negligence or falsely completing the task as quickly as possible to receive payment.

We thus ought to set a threshold, marking subjects whom score under which as null and void, disregarding their data from the analysis as it does not capture a try reflection of the individuals cognitive ability.

The same logic may be applied to many other tasks, though is most important for the WCST as this forms the basis under which many other dimensionality reduction and variable selection techniques are utilized.

**Decision threshold**: The decision was taken to remove all participants whom scored under 0.4 for the WCST task. This is a plausible threshold given the complexity of the task, and removes about 10% of the participants.

$$\lambda_{wcst} = 0.4$$

Figure 6 and 7 provide a visual illustration of the threshold.

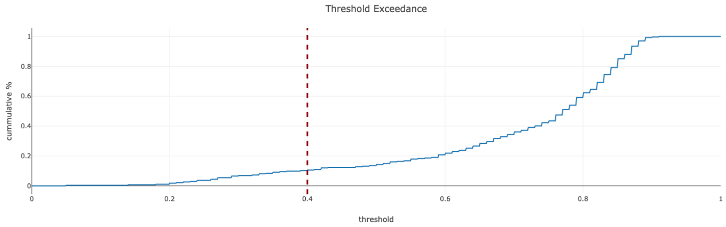


Figure 6: Cumulative probability density of WCST aggregate performance scores.

.

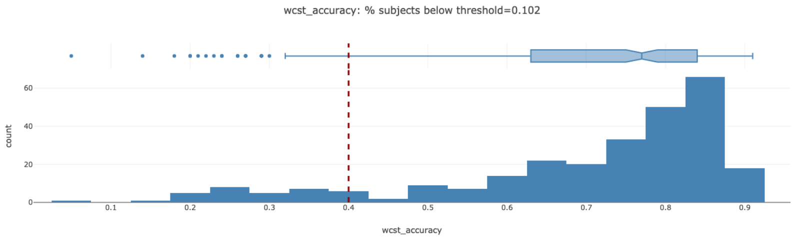


Figure 7: Distribution of WCST aggregate performance scores.

.

## 6.2 Navon task

We wish to examine the data generated by the Navon task in order to further simply the covariates summary statistics. The Navon task requires participants to identify patterns on either a *global* or *local* scale - assessing one's attention to detail at various levels. The questions arises: are there significant differences in global and local performance? If not, values may be aggregated.

As seen in figures 8 & 9 visually the distributions of Navon performance scores are very similar. Note that we are less concerned with the data labelled "None" as this is the absence of either a global or local matching rule. Nonetheless it is included for completeness.

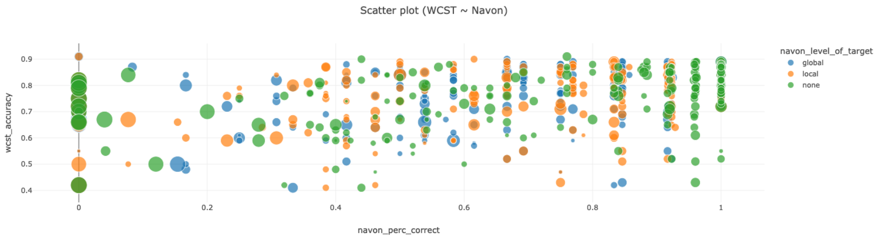


Figure 8: WCST performance as a function of different Navon scores.

.

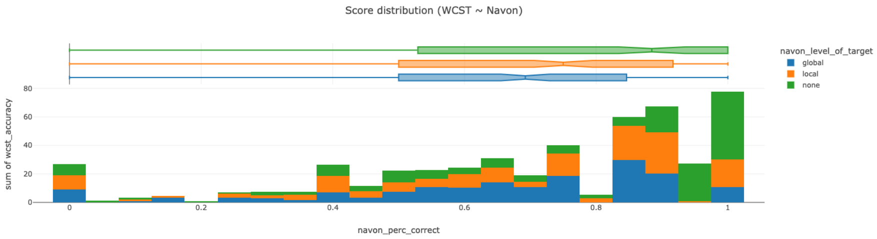


Figure 9: Distribution of Navon scores.

.

We are concerned with two distinct hypothesis regarding the relationship between Navon task performance and WCST performance:

1. Does the Navon task offer explanatory power when modeling the WCST?
2. Can any insight be drawn between the relative performance of global or local Navon performance?

To assess these hypothesis statistically, we summarize the Navon task data by the following three covariates:

$$\begin{aligned} X_{p,t}^{navon-GL} &: Navon\ task\ global - local\ performance && numeric \\ X_{p,t}^{navon-agg} &: Navon\ task\ aggregate\ score && numeric \\ X_{p,t}^{navon-rt} &: Navon\ reaction\ time && numeric \end{aligned}$$

This encoding allows us to capture potential variation between localized attention mechanisms across participants.

## 7 Dimensionality reduction

In order to converge to a parsimonious, plausible, model it is of utmost relevance that we reduce the dimensionality of the covariate space. Model selection techniques - as discussed in the methodology - are utilized as a form of variable selection however we can first reduce the number of covariates by performing some statistical analysis to assess the the empirical relevance of individual covariates.

It is salient to separate *psychological* and *demographic* variables as our research is primarily interested in psychological relationships in the data, whilst demographic information is purely auxiliary: and thus psychological variables require greater statistical rigor in their analysis.

**Model free analysis of non-linear mechanics**: the nature of the model design is inherently nonlinear. Proving additional complexity performing these pre-model (model free) analysis. In order to employ statistically robust techniques, variables are examined with respect to participants aggregate WCST score: *that is, the percentage of correct choices over the entire task*. This offers a simpler summer statistic to scrutinize the data. Naturally, this may negate the nonlinearities in the data. To compensate for any shortcomings, both linear and non-linear techniques are employed in covariate assessment.

**Covariate prioritization**: Given the sparse dimensionality of the dataset, it likely that the optimal model configuration utilizes a small subset of the covariates. Further, the core variables selection techniques discusses in methodology (likelihood ratios) requires one to order the variables beforehand. We thus need to establish an order of variable importance before modeling the data. 4 schools of techniques are used to rank the covariates:

1. Theoretical relationships
2. Linear relationships
3. Non-linear relationships
4. Ensemble relationships

Again, for our purposes it is salient to separate demographic and psychological covariates, with great emphasis on the latter.

**Demographic data**: Whilst only secondary information in our study, we do have access to participant demographic information. This information is additional and we are not concerned with deeply examining every possible interaction effect, however certain demographics (such as wealth or age) may correlate with certain cognitive performance metrics. As such one ought to determine whether or not a demographic variable may offer any explanatory power in the model.

**Psychological data**: Each experiment is ~~tasked with approximating some cerebral function:~~ conducted with the aim of assessing some executive function. As the space of possible model is exponential in the number of covariates, ~~it our~~ it is pragmatic to first establish the order of (both theoretical and empirical) relevance of each psychological covariate. That is, the relevance of each (approximated) executive function to our response variable: WCST performance.

## 7.1 Theoretical variable ranking

Our core objective is to contrast the empirical evidence with well understood neuropsychological literature, & thus we naturally begin with a theoretically informed basis.

### 7.1.1 Demographic data

There is no compelling argument to be made of the relationship between the social-economic information and WCST. A purely empirical approach is taken for this subset of covariates.

### 7.1.2 Neuropsychological data

Working Memory (WM) is the most relevant construct to our analysis. Whilst Wisconsin Card Sorting Task is not validated as an official working memory (WM) psychological assessment, two WM phases are observabled:

- Rule search.
- Rule application.

Which roughly correspond to:

- *Semantic WM*: one must maintain the rule search order in WM.
- *Visual WM*: one must maintain the card features in WM and find the matching card.

Notably, in practice the dichotomy is not as well defined as illustrated here: because the search phase requires visual WM and the application phases research at least remembering what the correct rule is. Nonetheless the distinction is useful in associating the WCST with the additional experiments.

**Task prioritization**: Both the *Nback* and *Corsi block span* tasks tap WM, they take precedence over the *Navon* and *Fitts* tasks. *Nback* maps to semantic WM, and *Corsi block span* maps to visual WM.

We are fundamentally interested in how subjects maintain and apply rule order in their WM. It naturally follows that semantic WM takes precedence over visual WM. The psychological tasks can be ranked in theoretical relevance to WCST as follows:

1. Nback
2. Corsi block span
3. Navon
4. Fitts

## 7.2 F-test: linear variable ranking

Each covariate is examined with respect to aggregate WCST scores individuals. The first, and simplest, comparison is to conduct an simple linear regression F-test to capture linear dependence.

**F-test**: A simple linear regression F-test fits a simple linear model to the data (in our case, the WCST aggregate performance as a linear function of the selected covariate) and then measures the relative variation explained producing an F-statistic:

$$F = \frac{explained\ variance}{unexplained\ variance}$$

This statistic is assumed to follow an F-distribution and thus can be tested for statistical significance.

In the discrete case, the calculation is performed:

$$F = \frac{\sum_{i=1}^{K} n_i \left(\bar{Y}_i - \bar{Y}\right)^2 / \left(K-1\right)}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2 / \left(N-K\right)}$$

Where $K$ is the number of groups in the discrete class; $N$ is the number of data points; $\bar{Y}$ is mean response in the dataset & $\bar{Y}_i$ is a mean response of group $i$. In the continuous case cross correlation is utilized.

Whilst theoretically robust (offering statistical founded interpretations) the F-test is severely limited only capturing independent linear relationships.

## 7.3 Mutual Information: nonlinear variable ranking

Originating in information theory, mutual information (MI) captures the mutual dependence between two random variables [4]. More technically, it quantifies the amount of information (in shannons bits, nats or hartleys) obtained about one random variable by observing the other [4]. MI is intrinsically linked to entropy of a random variable: which itself quantifies the expected ”amount of informaton” held in a random variable (analogous to variance of the random variable in the expectation) [4].

**Statistical intuition**: In statistics and mathematical data analysis, mutual information can quantify non-linear relationships in random variables [2]. More specifically, MI determines how different the joint distribution of a pair of random variables $p(X, Y)$ is from the product of the marginal distributions $p(X) \otimes p(Y)$; that is, the expected value of the pointwise mutual information (PMI) [2].

**variable ranking**: As illustrated in figure 10, MI is capable of capturing nonlinearities in the data. In our case, the relative scores of MI are of interest, provide a method for ranking covariates by the amount of mutual information with the response.

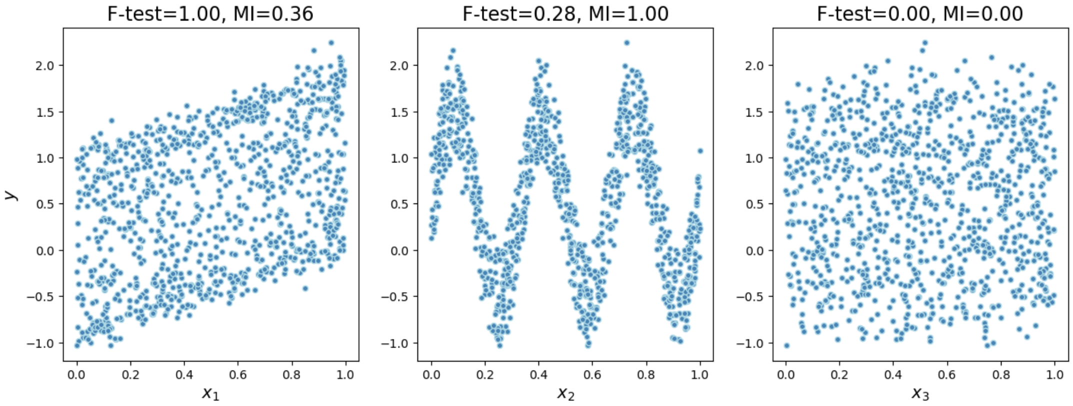


Figure 10: An illustration of non-linear dependence. The figures plot the relationship between 3 sets of random variables: the first with a high-variance linear relationship; the second with a low-variance non-linear relationship; and the third without any variable dependence. Both the F-test p-values and (normalize) mutual information are reported. It is clearly illustrated that MI is capable of capturing non-linear dependencies in the data by quantifying their joint probability density function relative to the tensor product of their individual density function.

.

**Computation**: Formally, the MI of two jointly continuous random variables is defined:

$$I(X,Y) = \int_{\mathrm{x}} \int_{y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

Where the intrals are replaced with summations for discrete variables.

**Kullback–Leibler divergence**: Intuitively is it straightforward to show the relationship between MI and another salient information-theory derived quantity the Kullback–Leibler (KL) divergence. The KL divergence $D_{KL}(P||Q)$ - or relative entropy - is an asymmetric statistical difference that measures how one probability distribution $Q$ differs from another $P$ [2].

The divergence of $P$ from $Q$ can be thought of as the expected excess "surprise" from using $Q$ as a model with the actual distribution over the space is $P$ [4].

For distributions $P$ and $Q$ of a continuous random variable, the relative entropy is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{p(q)} dx$$

**MI and $D_{KL}$**: MI can then be shown to be the $D_{KL}$ from the product of the marginal distributions $p(x) \otimes p(y)$ of the joint $p(x,y)$:

$$I(X,Y) = \mathbb{E}_Y \left[ D_{KL} \left( p_{X|Y} \; || \; p_X \right) \right]$$

Quantifying the conditional dependency (after accounting for variation in the reference covarate) [4].

## 7.4 mRMR: Maximum Relevance Minimum Redundancy: ensemble methods

**Intuition**: Maximum Relevance Minimum Redundancy (mRMR) is a prominent and highly successful feature selection algorithm that offers a succinct and intuitive framework to curate a subset of covariate that are not only most predictive, but also the most uncorrelated and distinct [24]. Many modern machine learning applications mine a vast plethora of sample covariates that require scientifically robust scrutiny and pruning in order to build generalizable and scalable models [24]. In order to arrive at interpretable/actionable solutions feature redundancy ought to be carefully considered in addition to purely (linear or nonlinear) correlation with the response variable [9]. In essence, mRMR balances feature importance and mitigates feature redundancy by ranking covariates according to some score that captures both metrics:

$$f(X_i) = \frac{\phi(Y, X_i)}{\delta(X_i, X^{'}_{i})}$$

Where $\phi(.)$ is some function that captures the relevance of feature $X_i$ when predicting response $Y$ and $\delta(.)$ is some function that quantifies the redundancy between feature $X_i$ and the set of compliment features $X^{'}_{i}$. Any plausible function set may be used to quantify these metrics, in the way the original algorithm can readily be extended to nonlinear and non-parametric functions [24].

**Formalization**:Whilst many variance have been thoroughly tested on both real-world and simulated datasets - including nonlinear kernel transformations and Shannon negative entropy - two variants prove consistently superior and thus are used in our analysis [24].

**1. FCQ (F-test correlation quotient)**:

$$f^{FCQ}(X_i) = F(Y, X_i) / \frac{1}{|S|} \sum_{X_s \in S} \rho(X_s, X_i)$$

Where $F(.)$ is the F-statistic score for relevance; $S$ is the compliment set of covariates (excluding feature $i$) and $\rho(.)$ is the Pearson correlation. Note that the $Q$ is used to denote the use of a quotient as apposed to a simple subtraction between (normalized) relevance/redundancy scores.

**2. RFCQ (F-test correlation quotient)**:

$$f^{RFCQ}(X_i) = I_{RF}(Y, X_i) / \sum_{X_s \in S} \rho(X_s, X_i)$$

Where $I_{RF}(Y, X_i)$ is an importance score computed by random forest feature selection [24].

**Random forest**: Decision trees are a simple discriminative algorithm that attempts to model dichotomous response variable by sequentially separating input data by their covariate values [10]. In the case of discrete covariates, the tree separates the data at each possible permutation, and selects the optimal configuring by computing the accuracy of the prediction and utilising the split that best fits the data [10]. In the case of continuous predictor variables, each possible splitting value is computing and thereafter the same model score (denoted the Gini importance or mean decrease impurity) is used to local the optimal split [10]. The algorithm is flexible and naturally extends to both (probabilistic) multinomial data and regression settings [10].

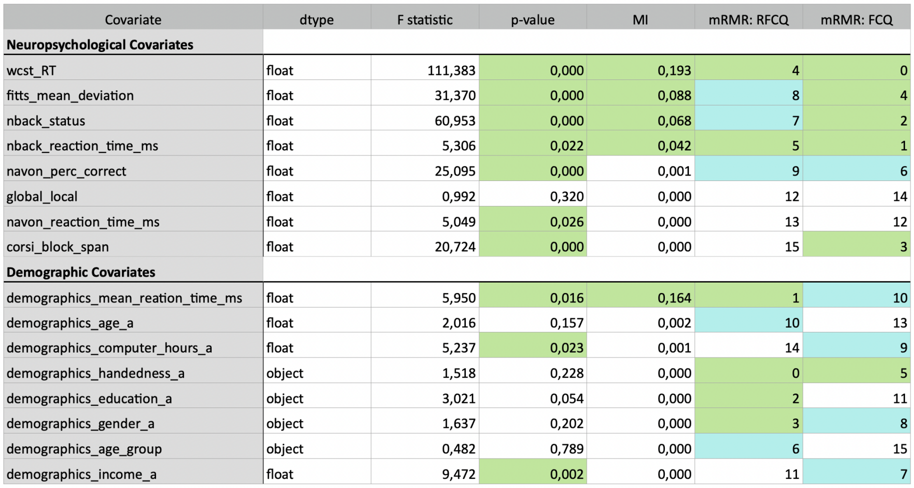
Whilst simple and easy to implement, classical decision trees are greatly flexible and thus regularly overfits any particular data instance. The widespread success of the algorithm is a consequence of leveraging statistically robust techniques to bootstrap aggregate performance [10]. By randomly sampling bootstrap-samples of the dataset and aggregating the predictions of many independently fit decision trees, statistically robust (in the limit) results can be achieved [10]. This algorithm is known as a Random Forest.

**Random forest feature selection**: Each node of each decision tree represents a dichotomous split in the data over a single covariate. As such, a natural extension is to use decision trees - and thus random forests - for feature selection. The (nonlinear) predictive power of any particular variable can be assessed by quantifying the deterioration of the algorithms performance in it's absence [10]. By this approach random forest offers a robust nonlinear random sampling approach to feature importance ranking and feature selection [10].

**Utility**: the two selected mRMR instances (FCQ and RFCQ) are utilized to aid our analysis by offering another variation of variable ranking.

## 8 Implementation of covariate selection

Each of the above methods are fit on our summary datasets. The response vector is the average *wcst* score earned by a participant over the entire trail. The results, available in table 1 are used to select and rank features. The **F-test** & **MI** the primary variables in our consideration. mRMR provide supportive information, capturing redundancy, however because we will test nested models it is instead favourable to be fairly liberal in variable selection and instead focus on variable ranking.



| Covariate | dtype | F statistic | p-value | MI | mRMR: RFCQ | mRMR: FCQ |
|---|---|---|---|---|---|---|
| **Neuropsychological Covariates** | | | | | | |
| wcst_RT | float | 111,383 | 0,000 | 0,193 | 4 | 0 |
| fitts_mean_deviation | float | 31,370 | 0,000 | 0,088 | 8 | 4 |
| nback_status | float | 60,953 | 0,000 | 0,068 | 7 | 2 |
| nback_reaction_time_ms | float | 5,306 | 0,022 | 0,042 | 5 | 1 |
| navon_perc_correct | float | 25,095 | 0,000 | 0,001 | 9 | 6 |
| global_local | float | 0,992 | 0,320 | 0,000 | 12 | 14 |
| navon_reaction_time_ms | float | 5,049 | 0,026 | 0,000 | 13 | 12 |
| corsi_block_span | float | 20,724 | 0,000 | 0,000 | 15 | 3 |
| **Demographic Covariates** | | | | | | |
| demographics_mean_reation_time_ms | float | 5,950 | 0,016 | 0,164 | 1 | 10 |
| demographics_age_a | float | 2,016 | 0,157 | 0,002 | 10 | 13 |
| demographics_computer_hours_a | float | 5,237 | 0,023 | 0,001 | 14 | 9 |
| demographics_handedness_a | object | 1,518 | 0,228 | 0,000 | 0 | 5 |
| demographics_education_a | object | 3,021 | 0,054 | 0,000 | 2 | 11 |
| demographics_gender_a | object | 1,637 | 0,202 | 0,000 | 3 | 8 |
| demographics_age_group | object | 0,482 | 0,789 | 0,000 | 6 | 15 |
| demographics_income_a | float | 9,472 | 0,002 | 0,000 | 11 | 7 |

Table 1: Covariate feature selection metrics. Covariates are separated into *neuropsychological* and *demographic* data. The p-value column (p-value associated with F-test) are marked green if significant at a 5% level. Mutual Information (MI, a purely relative figure) is marked green if in exceeding of 0.01. The two columns providing the results of the *mRMR* feature ranking mark the features $0-to-5$ as green and $6-to-10$ as blue.

**Selected demographic variables**: It is clear from this meta-analysis that demographic information offer little to no meaningful cognitive differences in the population. *Handedness*, *education* *level*, *gender* and *age* are not significant in the F-test nor produce meaningful MI results and as such can be dropped from the analysis moving forward. *Income* appears to have a significant linear relationship with the WCST, and may be included as a low priority nesting in the model. *Computer hours* and *Demographic reaction time* also show significant linear relationships. These variables, however, likely capture the same underlying generative process: the former being a self-described account of how frequently one uses computers, but the latter capturing the time taken to complete the demographic information which is likely to constitute attentiveness, alertness and how comfortable one is with the machine. Demographic reaction times boast a lower p-value, offer a substantially larger MI, and rank highly on the RFCQ variant of mRMR feature ranking. For this reason reaction times are included (assumed to incorporate computer hours) and computer hours are dropped from the analysis.

3 of the demographic variables that show no significant relationship with WCST ranking highly on the RFCQ variant of mRMR. One may falsely assume spurious correlation in the manor under which random forest decision trees are fit, however it is highly likely that, instead, the discrete nature of these covariates result in inflated importance in the compilation of decision trees as a function of a much smaller permutation space when compared with the continuous variables. We do not feel that these finding circumvents the negligible F-test and Mutual Information scores and thus these variables are dropped from the analysis.

**Selected neuropsychological variables**: *Navon global vs local* variable yields no relationship with the WCST response (supporting our prior analysis) and can be dropped from the analysis moving forward. Whist producing significant F-statistic p-values, both *Nback* and *Navon* RT (reaction time) score very low mutual information. *Navon RT* ranks low on both mRMR variants and can be excluded. *Nback RT* can be contrasted with *WCST RT*, as it scores lower on all metrics it may be excluded as multiple *RT* variables are superfluous. All remaining variables should be included in the analysis and ranked according to their theoretical relationship with the WCST (in the absence of strong statistical evidence to the contrary).

# 9 Model architecture