Recent Evidence for Reinforcement Learning in the Multi-Armed Bandit Task

Methods

These data were analysed in both wide and long formats using RStudio (Version: 2024.12.0 Build 467) running R version 4.4.2. `magrittr` (Bache & Wickham, 2022) and `tidyverse` (Wickham, 2023) were used to clean the data prior to analysis. Most models make use of the coding matrices from `codingMatrices` (Venables, 2023). Mixed effect models from the `afex` package used the Kenward-Roger approximation method unless otherwise specified (Singmann et al., 2024). Effect sizes were calculated using the `effectsize` package(Ben-Shachar et al., 2020). Analysis of the results was supported by `emmeans` (Lenth, 2024). Plots were graphed using the `ggplot2` package, including raincloud plots from `sdamr` (Speekenbrink, 2022). Exploratory factor analysis and parallel analysis was completed using the `psych` package (Revelle, 2024).

Reinforcement Learning

As participants continue to play the game, they are more likely to select the optimal option. Since the data was separated into blocks of sixty trials, the first hypothesis was that players performed better in later blocks of trials compared to earlier blocks.

To explore the data, a raincloud plot of the results grouped by trial block was created (see Fig. 1). Visual inspection suggested a general positive trend supporting the hypothesis. The data in each group was approximately normally distributed, implying

that linear models based on the normal distribution should be suitable. The *a priori* Levene's test indicated that there were unequal variances among the conditions,

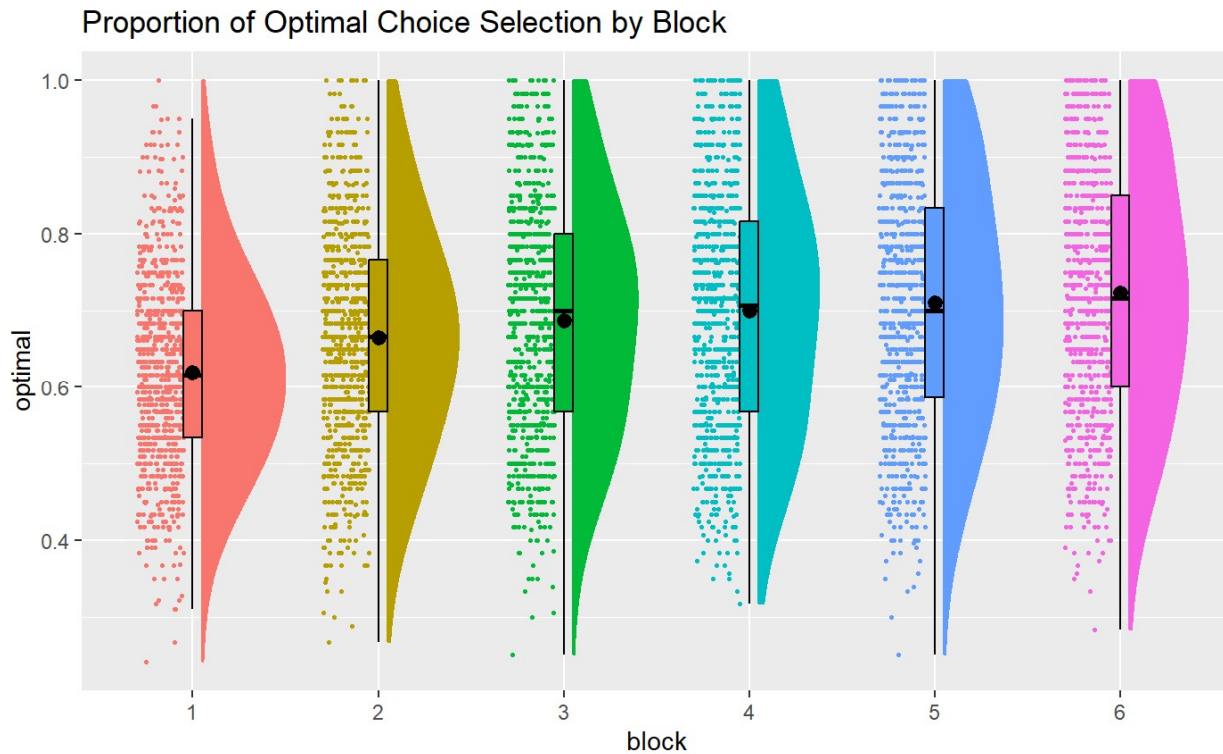

Figure 1. The proportion of optimal choice is separated by block. The distributions are arranged in raincloud plots. There is a general increasing trend.

$F(5,5274) = 22.441,\ p < 0.001$, the ratio between the largest and smallest variances was less than 4; therefore, the results of the analysis are likely robust despite the deviation (Maxwell et al., 2018).

The trial blocks were converted into factors which then had a difference coding matrix applied to them for an analysis of variance. The initial model investigated how the proportion of optimal choices made varied with the dependent variable, block number (see Table 1). The ANOVA results indicated a significant main effect of block, $F(5, 5274) = 54.71, p < 0.001,\ \eta^2 = 0.05$. The difference between blocks 1 and 2 was significantly different, $b = 0.045,\ t(5274) = 6.256,\ p < 0.001$. The difference between

Table 1: Summary of ANOVA Results

	Estimate	SE	t	p(>|t|)
Intercept	0.619	0.005100	121.455	<0.001
Block 2 - 1	0.045	0.007212	6.256	<0.001
Block 3 - 2	0.023	0.007212	3.158	0.0016
Block 4 - 3	0.013	0.007212	1.797	*0.0724*
Block 5 - 4	0.011	0.007212	1.493	*0.0724*
Block 6 -5	0.012	0.007212	1.725	*0.0846*

blocks 2 and 3 was also significantly different, $b = 0.023, t(5274) = 3.158, p = 0.0016$. This suggests that participants increase the proportion of optimal choices throughout the first 180 trials before stabilizing for the last three trials.

However, this model doesn't account for individual differences in the participants. Participant ID numbers were converted into a factor using an effect coding matrix. A second mixed-effects linear model computed using the `afex` package included a patient-wise random intercept to ameliorate this issue (Singmann et al., 2024). This model also found a significant effect of block on proportion of optimal choices, $F(5, 4395) = 148.26, p < 0.001, \eta_p^2 = 0.14$. Inspecting the fixed effects reveals that the difference between each block and the previous is significant and positive, which supports the hypothesis (see Table 2). The increases between blocks fall between 0.01 and 0.05, with the largest increase being from the first block to the second block, $M = 0.045, 95\%\ CI\ [0.037, 0.054]$. The increases after the third block are smaller, suggesting that learning slows down at this point.

Table 2: Summary of Mixed Effects ANOVA Results					
Random Effects	**Variance**			**Standard Deviation**	
ID	0.014441			0.1202	
Fixed Effects	**Estimate**	**SE**	**df**	**t**	**p(>|t|)**
Intercept	0.619	0.005100	176.3	121.455	<0.001
Block 2 - 1	0.045	0.004381	439.5	6.256	<0.001
Block 3 - 2	0.023	0.004381	439.5	3.158	0.0016
Block 4 - 3	0.013	0.004381	439.5	1.797	*0.0724*
Block 5 - 4	0.011	0.004381	439.5	1.493	*0.0724*
Block 6 -5	0.012	0.004381	439.5	1.725	*0.0846*

From these models, there is evidence to suggest that by completing more trials, participants are learning to select the optimal choice using a reinforcement paradigm. Participants will plateau in their learning about halfway through the experiment.

Distancing and Choice Optimisation

Distancing was hypothesised to increase the likelihood of a participant choosing the optimal option. To explore the data, two different raincloud plots were created to investigate what role, if any, condition and trial block had on selecting the optimal choice (see Fig. 2 and Fig. 3). Visual inspection indicated that there might be a slight increase across conditions, though it may not be significant. A similar trend was noted in when the data was grouped by trial block. The density distributions were approximately normal, suggesting that the assumption of homoscedasticity should hold.

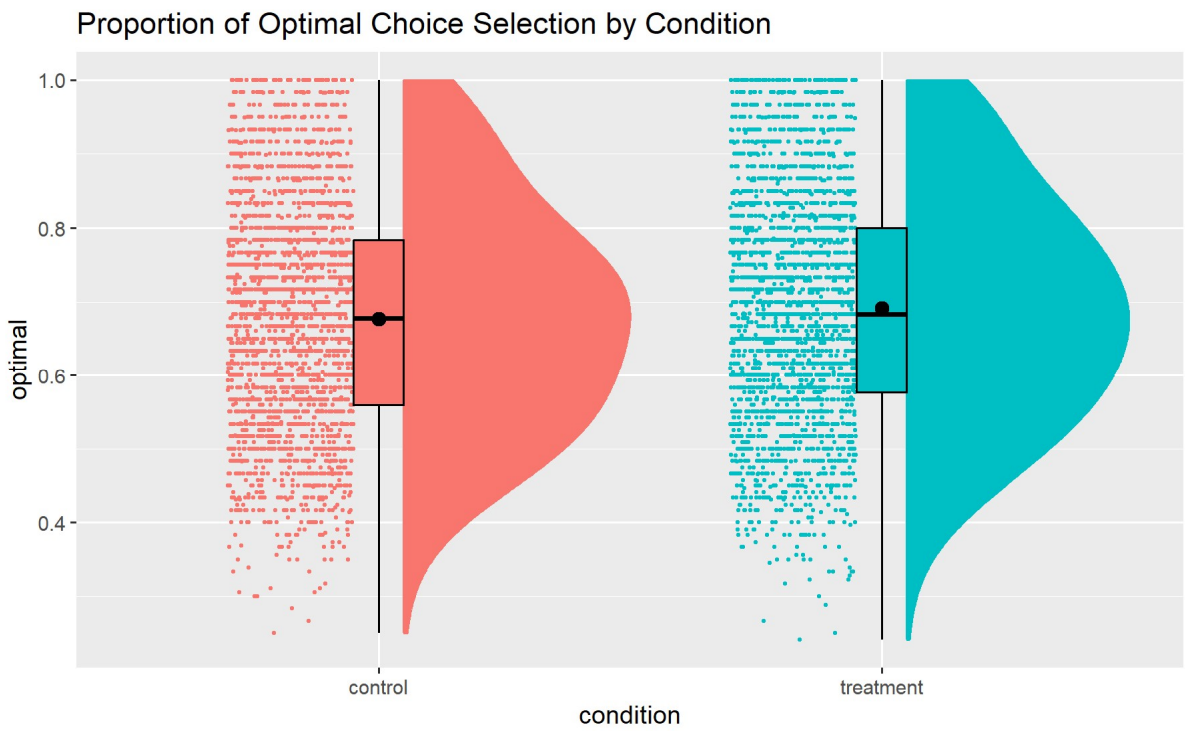

Figure 2. The proportion of optimal choice is separated by condition.

Figure 3. The proportion of optimal choice is separated by condition and by block of trials.

The first model was a simple linear regression model, with the proportion of optimal choice varying with condition. This model found significant results across conditions, $F(1,5278) = 10.96, p = 0.001, \eta^2 = 0.002$. Without the distancing intervention, the proportion of optimal choices selected is $M = 0.684$, $95\%\ CI[0.680, 0.688]$. Participants in the distancing condition selected the optimal choice more frequently, increasing their proportion of total optimal choice selections by $0.014, 95\%\ CI\ [0.005, 0.022]$. An insignificant result from a post hoc Levene test confirmed the assumption of homoscedasticity, $F(1,5278) = 0.7938, p = 0.373$. Inspecting the Q-Q plot of the model indicated that the assumption of normality holds quite well (see Fig. 4).

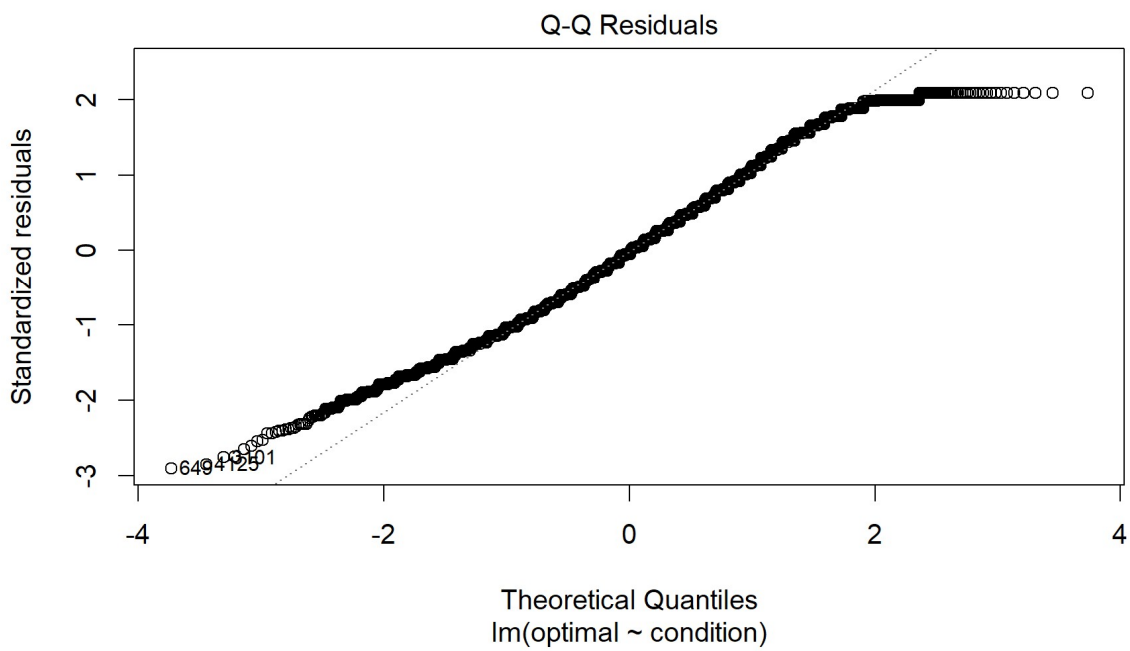

Figure 4. A Q-Q plot for a simple linear regression model of condition on proportion of optimal choice.

Further exploratory analysis was undertaken to unify condition and block in a single model. For this analysis, an effect contrast was used on block to investigate each block separately. A second model was a 2 (condition: control or distancing) by 6 (block: 1-6) ANOVA with optimal choice proportion as the dependent variable. The main effect of condition was significant, $F(1,5268) = 11.511, p = 0.001, \eta_p^2 = 0.002$, as was the main effect of block, $F(5,5268)) = 54.785, p < 0.001, \eta_p^2 = 0.05$. Participants in the distancing intervention increased their proportion of optimal choices by 0.014. In the first two blocks of trials, the proportion of optimal choices decreased by 0.065 in the second block and 0.020 in the third block. The fourth block was not significantly different from the grand mean. The last two blocks significantly increased the proportion of optimal choices, by 0.016 and 0.027 respectively.

While both condition and trial block have a significant effect on optimal choice, the effect of trial block appears to be much larger. It is also abundantly clear that these two variables cannot explain the variance on selecting the optimal choice. Further models could investigate whether affect, trial type, digit span, age, or gender play a role in optimal choice selection.

Distancing and Affect

To assess whether distancing impacted the affect of a participant, the third hypothesis asserts that when a participant was primed with the distancing paradigm, they will report a lower affect when they receive a reward when compared to participants that did not. The initial exploration of the data using raincloud plots (see Fig. 5) indicated that I might expect to find no difference across distancing conditions,

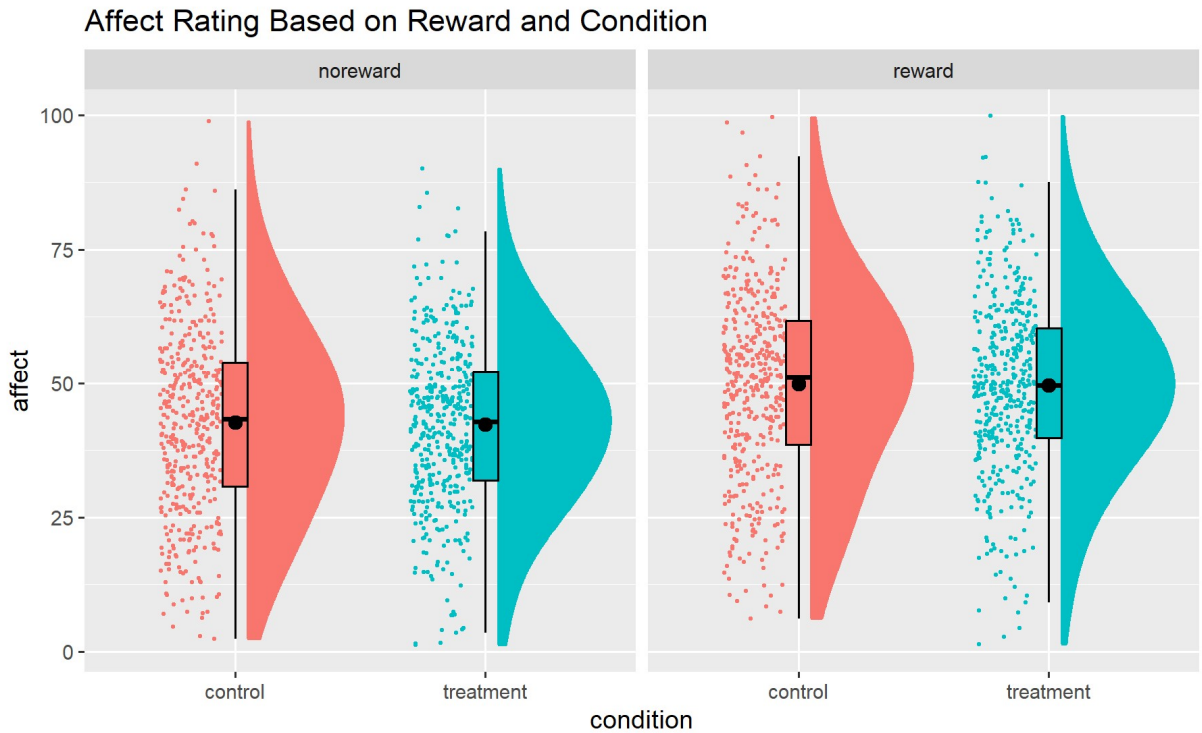

Figure 5. The affect is separated by condition and reward status.

though there still was a difference in affect between participants that received a reward versus those who did not. Visually, the density plots of each subgroup appeared normal, and post hoc Levene's test later supported the assumption of homoscedasticity by returning an insignificant result, $F(878) = 3.693, p = 0.05496$.

The first model was a 2 (reward status: reward or no reward) by 2 (condition: distancing or control) ANOVA with affect as the dependent variable. The ANOVA showed a significant result for the main effect of reward status, $F(1, 1756) = 82.1000, p < 0.001, \eta_p^2 = 0.04$. The reported total affect score after Scheffe adjustment when receiving no reward was smaller, $M = 42.5, 95\%\ CI[41.1, 43.9]$, than total affect when receiving a reward, $M = 49.7, 95\%\ CI\ [48.3, 51.1]$. In comparison, the main effect

of condition and the interaction between condition and reward status had insignificant results, indicating that condition had no direct or mediated effect on affect.

Although the Q-Q plot appeared normal during post hoc analysis, R identified a few data points which might be outliers. Since the Cooks' distances were all less than 1, there was no other evidence to support the existence of outliers and thus no need to alter this model. Further analysis using the `emmeans` package investigated various orthogonal contrast codes to validate the conclusions of the model (Lenth, 2024). As the model suggested, only contrasts which involved looking at differences in reward status yielded any significant results.

A further exploratory model investigated whether variations in working memory influenced affect and had any mediating effect on condition or reward status. This model was a 2 (reward status: reward or no reward) by 2 (condition: distancing or control) by digit span ANCOVA. As in the previous models, the main effect of reward status was significant, $F(1,1752) = 82.2959, p < 0.001, \eta_p^2 = 0.04$. The new main effect, digit span, was also significant, $F(1,1752) = 7.3720, p = 0.007, \eta_p^2 = 0.004$. The rest of the effects and interactions in the model were not significant.

One last exploratory model investigated the effect of including individual variations in affect. This model was a mixed effect ANOVA with a 2 (reward status: reward or no reward) by 2 (condition: distancing or control) by digit span and a random effect to account for differences between participants. Like the previous models, the main effect of reward status was significant, $F(1,876) = 701.55, p < 0.001, \eta_p^2 = 0.44$, while condition had no significant effect. However, the new factor, digit span, was a significant main effect, $F(1,876) = 3.93, p = 0.48, \eta_p^2 = 0.004$. Introducing this factor

created a significant three-way interaction, $F(1,876), = 5.88, p = 0.16, \eta_p^2 = 0.006$. The rest of the results are insignificant, suggesting that digit span mediates a relationship between reward status and condition. Further mediation analysis would be required to ascertain the nature of this interaction.

Transdiagnostic Criteria

The fourth and final investigation assessed whether responses to the various questionnaires could be reduced to three transdiagnostic criteria. In particular, it was hypothesized that 1) the responses to the Self-rating Depression Scale (SDS), Apathy Evaluation Scale (AES), State-Trait Anxiety Scale (STAI), and Dimensional Anhedonia Scale (DARS) questionnaires would correspond to a depression and anxiety criterion; 2) the responses to the Obsessive-Compulsive Inventory Revised (OCI) and Eating Attitudes Test (EAT) questionnaires would correspond to a compulsive behaviour and intrusive thought criterion; and 3) the responses to the Liebowitz Social Anxiety Scale (LSAS) and interpersonal component of the Schizotypal Personality Questionnaire – Brief, Revised, and Updated (SPQ) would correspond to a social withdrawal factor.

To confirm this hypothesis, the `lavaan` package was used to estimate a Structural Equation Model using maximum likelihood (Rosseel, 2012). Factors were created using a simple structure, where non-overlapping answers to the questionnaire battery identified the factors. Each factor fixes the first variable's loading to 1 to determine the scale of the factors. The factor loadings could then be compared by standardizing the loadings. Prior to constructing the model, the responses to the DARS questionnaire were reversed to be consistent with the other questionnaires having an

increase in score indicate a higher likelihood of having the associated symptom.

This model had a significantly better fit than a baseline model with complete independence ($\chi^2(2926) = 47958.240,\ p < .001$). However, the overall model fit was also significant ($\chi^2(2846) = 17401.361,\ p < .001$). Both measures of absolute fit were less than satisfactory, with $SRMR =\ .094$ and $RMSEA\ [90\%\ CI]\ ) = .076\ [.075, .077]$. The comparative fit index, $CFI = 0.677$, and Tucker-Lewis index, $TLI = 0.668$, both indicate that this model is less than satisfactory. Due to the unsatisfactory fit indices, I explored creating a second model.

To construct the second model, a factor was constructed for each questionnaire before loading onto the three diagnostic criteria. Like the first model, each factor fixed the first variable's loading to 1 and used the reversed values of the DARS questionnaire.

Table 3: Standardised loadings of latent variables

Latent Variable		Standardised loading	z-value	p-value
depression_anxiety	SDS	0.957	-	-
	AES	0.744	14.896	0.000
	STAI	0.950	20.965	0.000
	DARS	0.566	9.210	0.000
compulsive_intrusive	EAT	0.544	-	-
	OCI	0.828	9.341	0.000
social_withdrawal	LSAS	0.864	-	-
	SPQ-interpersonal	0.872	19.799	0.000

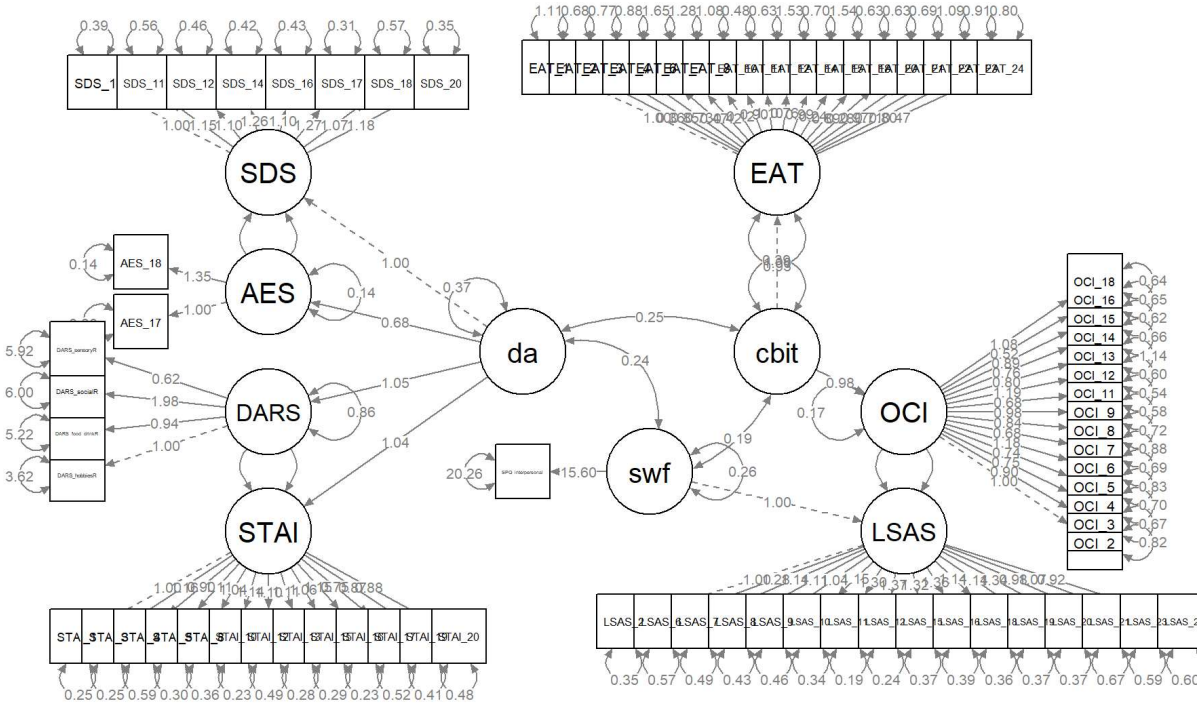

Figure 6. A graph of the SEM model. *da* = depression anxiety, *cbit* = compulsive behaviour and intrusive thought, *swf* = social withdrawal factor.

The details of the resulting model can be found in Table 3 (see Fig. 6). Like the previous model, the baseline fit ($\chi^2(2926) = 47958.240,\ p < .001$) and overall model fit ($\chi^2(2839) = 13565.693,\ p < .001$) returned significant results, providing contradictory indications as to the suitability of the model. Examining the absolute fit, both measures are less than satisfactory ($SRMR = .071,\ RMSEA\ [90\%\ CI] = .066\ [.064, .067]$). Lastly, the comparative fit index, $CFI = 0.762$, and the Tucker-Lewis Index, $TLI = 0.755$, were both less than their cutoff points, indicating an unsatisfactory fit.

Some of these measurements are an improvement compared to the first model. Since the second model is a nested model of the first, the preferred model can be identified using a chi-squared difference test ($\chi^2(7) = 3835.7\ ,\ p < .001$). This test

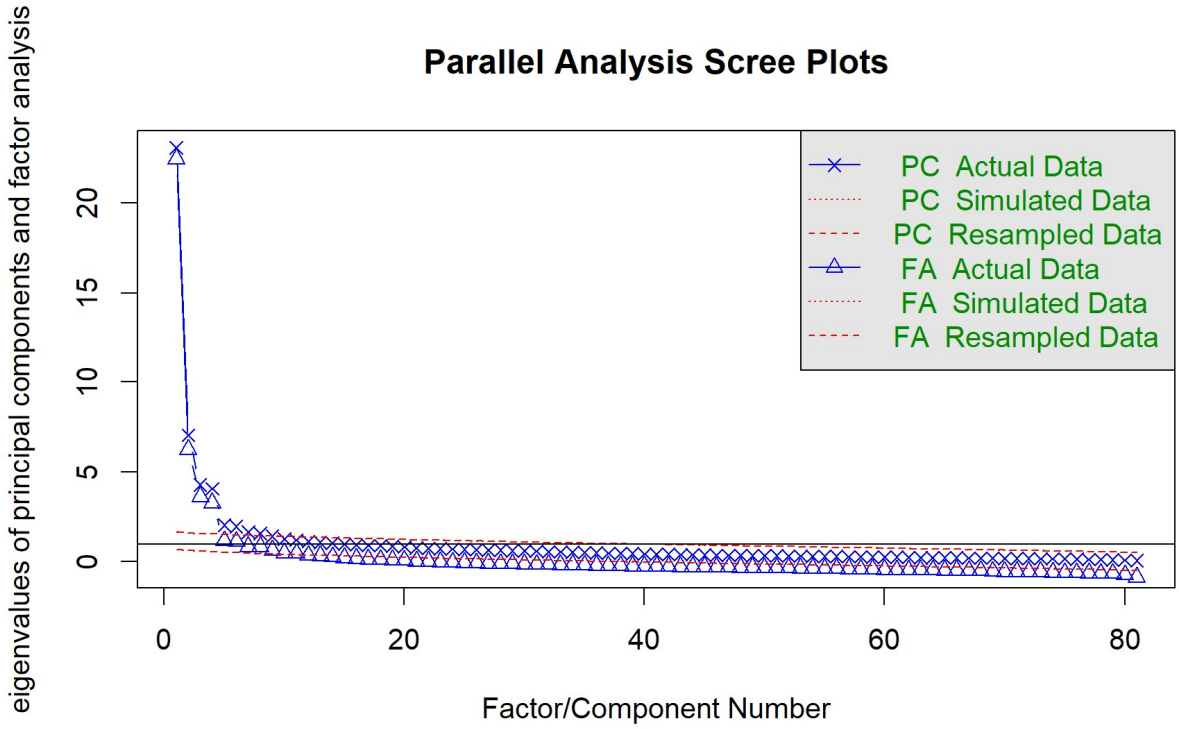

Figure 7. Parallel Analysis Scree Plots for the SEM model pictured in Figure 6.

returned a significant result, suggesting that the second model represents the data better than the first model.

As hypothesized, all the factors loaded significantly onto the three transdiagnostic criteria ($p < .001$). Excluding DARS and EAT, all the remaining factors had standardized loadings > 0.7, suggesting that they load quite strongly onto their respective transdiagnostic criteria. Both DARS and EAT had moderate standardized loadings, between 0.5 and 0.6. The three transdiagnostic criteria were trending toward large covariances as indicated by their standardized loadings falling between 0.57 and 0.78.

To investigate the underlying assumptions in the hypothesis, parallel analysis was used to assess whether the model included extraneous factors and components

(see Fig. 7). The scree plot suggested that there should be 10 factors and 7 components. The model has a similar construction, with each component corresponding with a questionnaire (excluding the SPQ, which stood on its own). The 10 factors suggested by the factor analysis correspond with the seven questionnaires and the three transdiagnostic criteria derived from them. This further supports the proposed model.

Lastly, exploratory factor analysis supported some of the hypothesized connections. Due to the moderate covariance between factors, Promax rotation was used to allow the residuals to vary with one another as well. Assessing the loadings of this analysis using a cutoff of 0.2, a few interesting patterns emerged. `ML3` mostly consisted of the results of the SDS, AES, STAI, and three of the DARS criteria. It also returned a few results from the OCI, EAT, and LSAS questionnaires. These variables are consistent with the depression and anxiety criteria as hypothesized. Since a few questions from other questionnaires loaded onto this factor, positive loadings could reflect an overlap in symptoms assessed by the questionnaires while negative could reflect contrasting symptoms. LSAS variables were most of those loading on `ML1`, but also included `SPQ_interpersonal, DARS_socialR`, and a few responses from other questionnaires. As hypothesized, these variables are largely associated with a social withdrawal factor and support this as a transdiagnostic criteria. `ML2` corresponds largely to responses from the EAT questionnaire while `ML4` corresponds to the OCI questionnaire. The remaining six factors have little relationship to the hypothesis, identifying correlations between various questionnaire answers.

A better model for these transdiagnostic criteria might be obtained by inspecting each question and identifying which criteria it could be associated with. Social withdrawal, depression, anxiety, and compulsive behaviours are frequently seen in the same patients, so it would not be surprising if some of the questions could apply to multiple transdiagnostic affects or if similar questions are asked across questionnaires. Further models could be developed integrating the BIS and remaining components of the SPQ to account for all behaviours assessed by the questionnaire battery.

Conclusion

The multi-armed bandit task was designed to assess reinforcement learning. The analysis of these data validates this idea as participants improve at the task as they advance through the blocks of trials. The aim of the study was to identify the extent of the distancing effect can improve reinforcement learning. A distancing intervention does not prevent reinforcement learning and instead has a small significant effect on the likelihood of choosing the optimal option during the task. Post hoc analysis indicated that distancing didn’t disrupt reinforcement learning. This supports the idea that distancing can improve the ability of a person to learn; however, since the effect is quite small it may not make a suitable intervention. Contrary to hypothesis, distancing had little effect on the participants’ reported affect. From the battery of questionnaires, three correlated transdiagnostic criteria were derived. It is possible that affect is mediated by these transdiagnostic criteria, for which further analysis would be necessary.

Word Count: 2439

Dataset Analysed: data_10

References

Bache, S. M., & Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R*. https://magrittr.tidyverse.org

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://rvlenth.github.io/emmeans/

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (Third edition). Routledge.

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. https://personality-project.org/r/psych/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2024). *afex: Analysis of Factorial Experiments*. https://afex.singmann.science/

Speekenbrink, M. (2022). *Sdamr: Statistics: Data Analysis and Modelling*. https://mspeekenbrink.github.io/sdam-r/

Venables, B. (2023). *codingMatrices: Alternative Factor Coding Matrices for Linear Model Formulae*. https://CRAN.R-project.org/package=codingMatrices

Wickham, H. (2023). *tidyverse: Easily Install and Load the Tidyverse*. https://tidyverse.tidyverse.org