Error in Sequential Action: An Evaluation of a Competence Model

Richard P Cooper (R.Cooper@bbk.ac.uk)
Centre for Cognition, Computation and Modelling
School of Psychologial Sciences
Birkbeck, University of London

Abstract

The Wisconsin Card Sorting Test (WCST) is commonly used to assess executive (dys-)function, particularly in neuropsychological patients. Performance on the test typically yields two types of error: perseverative errors, where participants persist in applying an inferred rule despite negative feedback, and set-loss errors, were participants cease applying an inferred rule despite positive feedback. The two types of error are known to dissociate. In this paper we apply an existing model of the WCST – the model of Bishara et al. (2010) – to a novel dataset, focussing specifically on the distribution of the two types of error over the duration of the task. Using Maximum Likelihood Estimation to fit the model to the data, we argue that the model provides a good account of the performance of some participants, but a poor account of individual differences. It is argued that this is because the model is essentially a competence model which fails to incorporate performance factors, and that accounting for the different types of error, and in particular the error distribution during the task, requires incorporating performance factors into the model. Some consequences of this for the broader enterprise of developing normative competence models are discussed.

Keywords: Set-loss errors; Perseverative errors; Wisconsin Card Sorting Test; Maximum Likelihood Estimation; Competence model

Introduction

Action is prone to error. This is particularly true of sequential action, i.e., action that comprises multiple steps which take place over time. Errors in sequential action take a variety of forms (see, e.g., Reason, 1979). In some cases, steps within an action sequence may be repeated after the immediate goal has been achieved (perseverative errors). In other cases, the immediate goal may be lost / neglected, such that individual steps become disconnected from that goal (set-loss errors). Both of these type of error may occur when participants perform the Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948; Heaton, 1981). This test is commonly used within clinical neuropsychology to assess executive, or more specifically dysexecutive, function. The classical pattern of performance indicative of an executive impairment, and a feature of many patients with neural damage to the frontal cortex, is the production of high rates of perseverative errors on the task – errors that reflect continued use of a response principle despite feedback to the contrary. The task is also sometimes used to assess efficacy of executive function in neurologically healthy individuals. For example, Miyake et al. (2000) found in a sample of 134 US college students a statistically significant correlation between the rate of perseverative errors on the WCST and performance measures on several simpler tasks that were held to assess the executive function of mental set shifting.

In the WCST, participants are presented with a series of cards, each showing one, two, three or four identical shapes – triangles, stars, crosses or circles – all coloured either red, green, yellow or blue. There are four target cards visible throughout the task, showing one red triangle, two green stars, three yellow crosses and four blue circles. Participants are asked to match each card from the series as it is presented to one of these four target cards. They are given binary feedback (correct / incorrect) after each card is matched. Their task is to use this feedback to infer the sorting rule being used by the experimenter (which is to match according to colour, number or form of the symbols on the card). Once a participant has demonstrated successful acquisition of the rule, through a run of consecutive correct matches, the experimenter changes the rule without warning. The participant must then infer the new rule. While there are several variations in how the task has been administered, typically participants are asked to sort either 64 or 128 cards, and typically a run of either 6 or 10 consecutive correct matches is required to trigger the experimenter to change the sorting rule.

The WCST yields several dependent measures, but the measures that are most widely discussed concern perseverative errors, i.e., responses that involve applying the previous correct rule, despite negative feedback. These errors have been argued to arise from either failure to integrate negative feedback or inertia in switching away from a previously reinforced rule and in instating a new rule. Either way, they reflect a failure of reactive control. A second type of erroneous response that can occur in the WCST is a set-loss error. These errors occur when a participant appears to have correctly inferred the rule in use by the experimenter (e.g., sort according to colour), as demonstrated by a run of correct responses, but then produces an erroneous response (i.e., a response not consistent with the experimenter's rule) despite having only received positive, reinforcing, feedback over the last series of trials. Set-loss errors are generally far less frequent than perseverative errors. Phenomenologically they appear to result from interference (from alternative possible sorting rule) or attentional drift.

The two forms of error are of general interest because they occur across a range of laboratory tasks (e.g., verbal fluency:

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Reske, Delis, & Paulus, 2011), as well as in everyday behaviour (Reason, 1979). They also dissociate. For example, in studies of WCST performance with healthy aging individuals, the rate of set-loss errors, but not the rate of perseverative errors, correlates with age (Caso & Cooper, 2022). This suggests that the two forms of error arise from different cognitive limitations, or limitations to different cognitive processes involved in the task.

Given the widespread use of WCST as a clinical instrument, it has been the focus many modelling efforts, including early work by Dehaene and Changeux (1991), as well as more recent work (e.g. Barceló, 2021). Of particular interest for the current work is the model of Bishara et al. (2010), which provides a formal parameterised mathematical account of performance of the WCST, and which yields, on a trial-by-trial basis, a probability distribution over response options. Unlike other models of the task, this allows the Bishara et al. (2010) model to be fitted to trial-by-trial participant responses using Maximum Likelihood Estimation (MLE) – an approach to model fitting that is generally considered to be superior to fitting to means of group-level dependent measures (Myung, 2003).

In this paper we report novel (though unremarkable) data from the WCST, and attempt to fit the Bishara et al. (2010) model to that data using MLE. In doing so, we propose a slight adjustment to the model (to account for participant responses that fail to match any rule). While the model fit is generally good, closer examination of errors throughout the task reveals that the model under-estimates perseverative errors on early trials, while also over-estimating set-loss errors on those trials. It also fits the response sequences of participants who produce few errors on the task better than those who produce many errors. We conclude that, while the model provides an adequate account of WCST competence, it is limited as a model of WCST performance.

The Model of Bishara et al. (2010)

Bishara et al. (2010) model performance on the WCST by assuming that participants maintain a set of attentional weights that vary over time — one per dimension (number, colour, form) — and that participants a) use these weights for choosing where to place each stimulus card, and b) update their weights following feedback after placing each card. The basic idea is that feedback allows attention to be focused on the relevant dimension, with positive feedback strengthening the weight of matching dimensions at the expense of mismatching dimensions, and negative feedback weakening the weight of matching dimensions (and thereby strengthening the weight of mismatching dimensions).

For each trial (i.e., each card to be sorted), the model uses the dimensional attention weights to produce a probability distribution over the four possible responses. This contrasts with many process models that generate sequences of discrete responses (such as that of Dehaene & Changeux, 1991), and makes the model ideal for fitting to individual participant responses using Maximum Likelihood Estimation.

Dimensional Attention and Model Initialisation

For notational convenience, we follow Bishara et al. (2010) and represent dimensional attention, a_t, as a 3 by 1 matrix. Without loss of generality, Bishara et al. (2010) assume that the attentional weights (i.e., the three components of a_t) sum to 1.0. Moreover they demonstrate that allowing the initial weights to vary across the dimensions does not significantly improve the model's fit to participant data. Thus, they assume that at trial $t = 0$ the attentional weights are initially equal:

$$a_0 = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix} \tag{1}$$

Action Selection (Selection of the Target Pile)

Given the attentional weights, on presentation of a card on trial t, the probability of placing the card under target card k ($P_{t,k}$), is given by:

$$P_{t,k} = \frac{m'_{t,k} a_t^d}{\sum_{j=1}^{4} m'_{t,j} a_t^d} \tag{2}$$

where m'_t is the transpose of m_t, the *match matrix*, as described below, and d is the *decision consistency*, a parameter that determines how strongly attentional differences should influence behaviour.

The match matrix at time t, m_t, is a 3×4 matrix which, in the formulation of Bishara et al. (2010), is defined as follows:

$$m_{t,k,i} = \begin{cases} 1 & \text{if the presented card matches} \\ & \text{target card } k \text{ on dimension } i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where k ranges over the 4 target cards / columns and i ranges over the 3 attentional dimensions / rows.

Adjusting Attention on the Basis of Feedback

Following feedback, attentional weights are adjusted according to Equation 4 (which defines the feedback signal, s_t, based on reward or punishment) and Equation 5 (which defines the new value of the attention matrix).

$$s_{t,i} = \begin{cases} \frac{m_{t,k,i} a_{t,i}^f}{\sum_{h=1}^{3} (m_{t,k,h} a_{t,h}^f)} & \text{if rewarded} \\ \frac{1 - m_{t,k,i} a_{t,i}^f}{\sum_{h=1}^{3} [(1 - m_{t,k,h}) a_{t,h}^f]} & \text{if punished} \end{cases} \tag{4}$$

$$a_{t+1} = \begin{cases} (1-r) \cdot a_t + r \cdot s & \text{if rewarded} \\ (1-p) \cdot a_t + p \cdot s & \text{if punished} \end{cases} \tag{5}$$

The feedback signal (s_t) defined in Equation 4 reinforces dimensions that match following positive feedback, but punishes dimensions that match following negative feedback. As in Equation 2, the attentional weights are raised to a power (f,

the *attentional focus*) in calculating the strength of feedback signal. Equation 5 calculates the new attentional weights as a weighted average of the existing value and the feedback signal, with the parameters r and p allowing for differential treatment of positive and negative rewards.

Parameters and their Plausible Ranges

The model includes four parameters. Reward sensitivity (r) and punishment sensitivity (p) control how the feedback signal affects the attentional weights (cf. Equation 5). These parameters both range from 0 to 1. When 0, the feedback signal is completely ignored and attentional weights persist. When 1, the attentional weights on the previous trial are ignored and behaviour on each trial is purely a function of feedback following the model's previous response. Attentional focus (f) and decision consistency (d) play similar roles at different points within the model. Attentional focus affects how feedback is weighted to the three dimensions in generating the feedback signal s_t (cf. Equation 4), while decision consistency (d) affects how the attention vector weights are applied when sorting cards (cf. Equation 2). Consistent with this last point, Bishara et al. (2010) found that fixing either f or d at 1 while allowing the other to vary freely did not result in a statistically significant decrease in the model's ability to fit their participants' data.

Fitting the Model to a Novel Dataset

As a first step to understanding the origins of perseverative and set-loss errors and evaluating the model of Bishara et al. (2010), data were collected from 48 participants who completed the WCST. Best fitting parameters were then found for each participant. For both real and simulated participants, the distribution of perseverative and set-loss errors over trials was then explored.

Empirical Study: Method

As part of a larger study approved by the local ethics committee, 48 participants recruited through the Prolific participant service completed an online, computerised version of the long-form (i.e., 128 card) WCST. Participants (aged 18+, and with self-declared fluent English) were asked to match each stimulus card with one of the four standard target cards, and given feedback (correct / incorrect) after each trial according to the standard procedure. The correct sorting rule, which the participant was not informed of but needed to infer and keep track of, was changed after every 10 consecutive correct responses, cycling through the three possibilities (colour, form, number) until all 128 cards were sorted. Stimulus cards were presented in the standard order of Heaton (1981) and for each participant all 128 responses were recorded. The task parameters (128 cards, rule changes only after 10 consecutive correct responses) were chosen so as to ensure ample opportunity for set-loss errors. The WCST took most participants less than five minutes to complete.

DV	Human	Model_a	Model_b
Correct	98.53 (16.31)	100.96 (11.18)	101.17 (11.43)
Categories	7.43 (2.62)	6.72 (2.67)	6.74 (2.88)
Perseverations	20.38 (12.24)	15.24 (7.56)	14.91 (6.20)
Set-Loss	2.06 (1.93)	3.74 (2.94)	3.33 (2.81)
TFC	15.72 (11.90)	15.54 (8.02)	17.74 (16.31)

Table 1: Means (standard deviation) of observed and simulated dependent measures. *Correct* is the number of cards (out of 128) correctly sorted. *Categories* is the number of categories or rule changes completed over the 128 cards. *Perseverations* is the number of classical perseverative errors, as scored according to Heaton (1981). Note that this treats the first response after a rule change as perseverative (if it is an error), even though at this stage the participant will not have received negative feedback. *Set-Loss* is the number of set loss errors, where obtaining a set is determined by a series of consecutive correct responses where at least three in the series are unambiguous. *TFC* is the number of trials to fully complete the first category. $N = 47$ in all cases.

Empirical Study: Results

One participant correctly sorted only 34 out of 128 cards (chance = 32) and did not obtain a run of 10 consecutive correct responses. This participant was excluded from further analysis. Descriptive statistics for standard dependent measures of the remaining 47 participants are shown in the leftmost columns of Table 1.

Figure 1 show the relative frequency of perseverative and set-loss errors produced by participants over the course of the task. Previous studies have tended not to report this level of data, but the figure appears to show that perseverative errors are more common early in the task than later in the task, with a notable spike after the first category is complete (around trial 15), and a lesser spike after the second is complete (around trial 25). This is reasonable given that as participants become acquainted with the way rules change in the task, their tendency to perseverative is likely to decrease. Set-loss errors show a different pattern. While they are rare, they are more frequent (though not significantly so) in the second half of the task than in the first. This is again unsurprising, on the assumption that such errors are due to memory confusions or attentional drift, both of which are likely to increase as the task progresses.

Data Fitting: Method

Maximum Likelihood Estimation Given that the model generates a probability distribution over response options on each trial (i.e., on each card sorted), the model can be fitted to trial-by-trial data by maximising the likelihood of it producing the observed outcome on each trial. To do this, we calculate, for each successive trial, the probability of the model producing the observed response. The observed response is then applied and feedback used to adjust a_t on the basis of the above equations, and the process repeated. The logarithms of the probabilities of all 128 observed responses are summed

(giving the log of the product of the probabilities, which is the log of the probability of the entire observed sequence for the given participant). The result is the log likelihood of the sequence, which is necessarily negative (as the logarithms of all probabilities are negative). By convention, we therefore consider minus log likelihood ($-LL$), where a better fit corresponds to a lower $-LL$. Maximum Likelihood Estimation (MLE) within the context of the Bishara et al. (2010) model involves finding values of model parameters, on a participant-by-participant basis, that maximise likelihood (i.e., minimise minus log likelihood).

Modelling Mismatched Responses One difficulty with using MLE with the model as described by Bishara et al. (2010) is that it requires that participants never attempt to place a card beneath one that matches on no dimensions (e.g., placing two yellow triangles under the four blue circles target card). We refer to these as mismatching responses. Equation 2 to Equation 5 ensure that the probability of selecting such a target location and hence of producing a mismatch response will be exactly zero. Yet such responses were produced occasionally by some of our participants, and this is not uncommon. Attempting to fit the model via MLE to data containing such responses fails, as the log likelihood of such responses is infinitely negative.

This difficulty can be addressed in several ways that are broadly consistent with the model. In particular, as Equation 2 determines the probability of each response based on the match matrix and attention vector, one might simply modify this equation to allow a small but non-zero probability of selecting a mis-matching response. We propose two such potential modifications, corresponding roughly to ε-greedy and Boltzmann (or softmax) action selection in the Reinforcement Learning literature:

Model$_a$: modify the match matrix (Equation 3) so that mismatching entries are small positive values, e.g., by adding a further parameter (call it b, *mismatch baseline*) and replacing Equation 3 with Equation 6:

$$m_{t,k,i} = \begin{cases} 1-b & \text{if the presented card matches} \\ & \text{target card } k \text{ on dimension } i \\ b/3 & \text{otherwise} \end{cases} \quad (6)$$

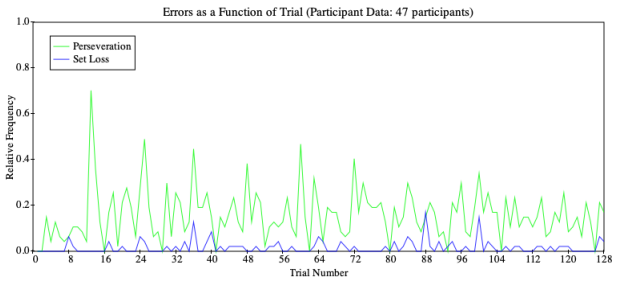

Figure 1: Errors as a function of position in the task for human participants.

The b parameter introduces the possibility of mismatching into the selection process. When b is 0 Equation 6 reduces to Equation 3. When b is 0.75 all entries in m_t are equal, meaning that all four choices will be equi-probable, while when b is 1.0 matching to features will be completely avoided.[1] The value of b may be fixed per participant to the number of mismatching responses divided by the total number of cards sorted. This ensures that b is not a free parameter and furthermore that if a participant never produces mismatching responses b will be zero.

Model$_b$: retain the match matrix of Equation 3 but apply a Boltzmann selection function to the match score. This can be achieved by replacing Equation 2 with Equation 7:

$$P_{t,k} = \frac{e^{\frac{m'_{t,k} a_t}{\tau}}}{\sum_{j=1}^{4} e^{\frac{m'_{t,j} a_t}{\tau}}} \quad (7)$$

As in Equation 2, the denominator in Equation 7 is the sum over all numerators and is therefore a normalising constant which ensures that at any time t, the probabilities of all options k sum to 1. The numerator for each of the four options k, is the weight of that option, which is calculated as e raised to the power w/τ, where w is the product of the transpose of the match matrix and the dimensional attention (as in Equation 2, but with $d = 1$) and τ is the standard Boltzmann temperature parameter. τ serves a function similar to d in Equation 2. When τ is large (much greater than 1) differences in w are compressed and options with relatively low w have a chance of selection. When τ is small (greater than zero but much less than one) choice between options strongly favours the option with the greatest w. Importantly, for all values of τ greater than zero, Equation 7 returns a non-zero probability for each option, even if w is zero. This therefore ensures that the model can, in principle, select non-matching options.

Parameter Estimation Simulated annealing was used to determine values of model parameters that maximise likelihood of the model for each of the 47 valid participants. For each model variant, and each participant's response sequence, initial values were selected for the parameters (r, p, f and d for model$_a$ and r, p, f and τ for model$_b$) based on the parameter values reported by Bishara et al. (2010). Then (for each participant's response sequence and each model variant) the log likelihood of the model with the current parameters was compared with the log likelihood of the model with a perturbed set of parameter values (with each parameter perturbed from the current values by random noise drawn from a normal distribution with mean 0 and standard deviation of 0.1). The parameters producing the best fit were retained, and the process iterated a total of 1000 times. Critically, the model was fitted on a participant-by-participant basis, resulting in different parameter values each participant and hence a

[1] This formulation, with $b/3$ in the *otherwise* clause, ensures that each of the three rows of m_t sum to 1. This is aesthetically pleasing, but not necessary.

Parameter	Model$_a$	Model$_b$
Reward Sensitivity	0.64 (0.29)	0.54 (0.30)
Punishment Sensitivity	0.56 (0.31)	0.49 (0.32)
Attentional Focus	0.71 (0.79)	0.73 (1.03)
Decision Constency	1.29 (1.29)	— —
Mismatch Baseline	0.01 (0.02)	— —
Temperature	— —	0.19 (0.08)
Fit ($-LL$)	44.48 (25.80)	47.47 (28.39)

Table 2: Descriptive statistics (means and standard deviation) of parameter values and model fit for the two models. The statistics are calculated over the 47 virtual participants, with individual parameters as determined by the simulated annealing process described in the text.

"virtual twin" for each human participant.

Data Fitting: Results

Parameters and Dependent Measures Table 1 (rightmost columns) shows the dependent measures for resulting 47 virtual twins for each model variant, while Table 2 shows the resultant descriptive statistics of the parameters and minus log likelihoods. In order to interpret the fit, note that it is the sum of logarithms of probabilities from 128 events. A fit of -44.48 therefore corresponds to a mean contribution of -0.3475 (i.e., $-44.48 \div 128$) to the fit per step, or a mean probability of 0.706 (i.e., $e^{-0.3475}$) as the prediction of each response. In reality, many responses are predicted perfectly, with a handful predicted very poorly, and there is considerable variability in fit across participants.

The values in Table 2 are robust to variation in the random seed used in the simulated annealing process. The fitting process was repeated 12 times for each model with different random seeds and in all cases yielded a fit for model$_a$ in the range of 44.1 to 44.6 and for model$_b$ in the range of 46.4 to 47.8. This suggest that while the Boltzmann approach to action selection can produce a good fit to the data, use of the mismatch baseline produces (on average) a superior fit.

A second point to note concerns the variability in the values of the parameters (and the fit). In all cases the standard deviations are high relative to the mean values and ranges of the parameters. This reflects the considerable variability in the performance of the participants, though it is noteworthy in Table 1 that variability in the standard dependent measures across the virtual populations for both models is similar to the variability in those measures across the human participants.

Related to this second point, across participants, the pairwise correlations between each of the dependent measures in Table 1 and each measure of fit are all significant: see Table 3. This suggests that the degree of fit between the actions of any particular human participant and the model (for either model) depends on how well the participant did on the task. If a participant obtains relatively many categories, with relatively many correct cards and relatively few errors of either type, then it will be possible to fit the model to that participant, but if the participant performs the task poorly, then fitting the model to the participant's responses will be less successful.

DV	Human vs. Model$_a$	Human vs. Model$_b$
Correct	$-.943$	$-.910$
Categories	$-.945$	$-.919$
Perseverations	.841	.790
Set-Loss	.569	.575
TFC	.554	.546

Table 3: Correlations between DVs of Table 1 and model fits. $N = 47$ in all cases. $p < .001$ in all cases.

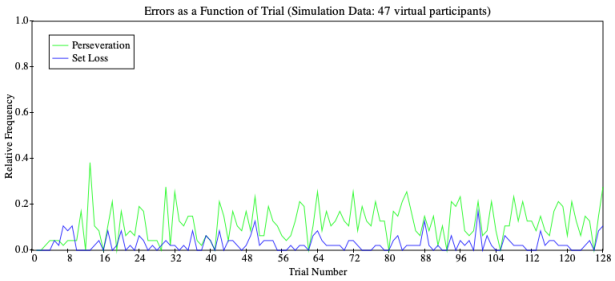

Figure 2: Errors as a function of position in the task for simulated participants (using model$_a$).

A final key point to note concerns the specific values of the parameters. Across participants, surprisingly, the values of r (reward sensitivity) derived from model$_a$ and model$_b$ do not correlate ($r = .179$, $p = .228$). Nor do the values of a (attentional focus; $r = .189$, $p = .203$), though the values of p (punishment sensitivity; $r = 0.814$, $p < .001$) do. Moreover the value of d in model$_a$ correlates with the value of τ in model$_b$ ($r = -.546$, $p < .001$). This latter correlation reflects the fact that increasing d and decreasing τ have similar effects – they both result in the model variant adhering more to dimensional attention and less to random selection. Despite the mixed nature of these correlations between equivalent parameters, the fit of the model variants are almost perfectly correlated ($r = .970$, $p < .001$). This implies that if one model variant is able to fit the responses from a participant, then the other model variant is likely to also be able to fit those responses, but there is considerable variability in how well participant responses can be fit, and if one model variant cannot be made to fit a participant's response sequence, then it is unlikely that the other model variant can be made to fit any better.

Error Profiles Figure 2 shows the simulated rate of errors of each type throughout the task for model$_a$. (Similar results were produced by model$_b$.) Qualitative comparison of this with Figure 1 suggests that the model fails to capture two points of the human data: the tendency for perseverative errors to occur with greater frequency early in the task compared with later in the task, and the opposite tendency for set-loss errors. Quantitative comparison of these apparent trends throughout the task is limited given the relatively small sample size. Nevertheless, the qualitative differences in the two figures are surprising considering that the model is fitted against trial-by-trial data.

Discussion

The preceding simulation studies demonstrate that the Bishara et al. (2010) model of WCST performance, when modified (in two different ways) to address the occasional tendency of participants to place cards in non-matching locations, appears to provide a good account of participant behaviour. By using simulated annealing with Maximum Likelihood Estimation, we were able to fit the model to trial-by-trial data from 47 participants yielding means and standard deviations of key dependent measures that were all close to the observed values (cf. Table 1). However, close examination of the model's performance reveals some potentially concerning issues. Specifically, while both variants of the model are good at modelling performance of the "best" participants, both variants are much less good at modelling the performance of "poorer" participants. At the same time, there appear to be differences in the distribution of errors produced by the model throughout the task when compared with the distribution of errors produced by human participants.

The correlations between the fit of the virtual twins and the dependent measures from the human data (Table 3) are concerning. It follows from this that attempting to account for individual differences in either model is likely to be flawed. For example, fitting the model to a participant with relatively few errors may yield (e.g.,) a high value for punishment sensitivity, while fitting the model to a participant with relatively many errors may yield the opposite, but if (as appears here) the model can be fitted more precisely to the actions of a participant with few errors than a participant with many errors, then one cannot reasonably compare the parameter values obtained in the two cases. Thus, reliable claims cannot be made concerning individual differences in reward or punishment sensitivity, etc., unless those individual differences are in the context of otherwise similar levels of performance of the individuals concerned (and similar degrees of fit of the model to the individuals' behaviours). This also undermines the model's utility in explaining the errors of neurological patients, who typically perform very poorly on the WCST.

Another concern is the apparent qualitative difference in error profiles produced by the human participants and their virtual twins (Figure 1 versus Figure 2). The simulations failed to yield the spike in perseverative errors seen in the human data after obtaining the first rule, but instead yielded a fairly constant rate of perseverative errors across the task. They also yielded considerably more set loss errors (and in fact significantly so), particularly in the early part of the task. This is surprising given that the model was fitted to trial-by-trial data, but perhaps reflects a limitation of the models' parameterisations. Neither model variant attempts to capture performance factors, such as drift in attentional focus throughout the task, or increasing sensitivity to reward/punishment that might occur as the participant becomes familiar with the task. This limits the extent to which MLE can capture the trail-by-trial data. As variation in parameters over time is not captured in the model(s), MLE must smooth those variations, resulting in parameters which smooth the predicted profile of errors. That both variants of the model over-predict set-loss errors suggests that the explanation of such errors (e.g., low decision consistency in model_a or high temperature in model_b) is incomplete. This concern reinforces that of the previous paragraph: the model as it stands may be adequate as a competence model – a model of idealised performance – but it is limited as a performance model (i.e., as a model of real human behaviour; cf. Chomsky, 1965). While the competence/performance gap is a common feature of mathematical models, the current work suggests that the implication of this gap for fitting competence models to human data, particularly on a trial-by-trial basis using MLE, is under-appreciated.

References

Barceló, F. (2021). A predictive processing account of card sorting: Fast proactive and reactive frontoparietal cortical dynamics during inference and learning of perceptual categories. *Journal of Cognitive Neuroscience*, *33*(9), 1636–1656.

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., & Busemeyer, J. R. (2010). Sequential learning models for the Wisconsin Card Sort Task: Assessing processes in substance dependent individuals. *Journal of Mathematical Psychology*, *54*(1), 5–13.

Caso, A., & Cooper, R. P. (2022). Executive functions in aging: An experimental and computational study of the Wisconsin Card Sorting and Brixton Spatial Anticipation Tests. *Experimental Aging Research*, *48*(2), 99–135.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT press.

Dehaene, S., & Changeux, J.-P. (1991). The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, *1*(1), 62–79.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404.

Heaton, R. K. (1981). *A manual for the Wisconsin Card Sorting Test*. Odessa, Florida, 33556: Psychological Assessment Resources, Inc.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Reason, J. T. (1979). Actions not as planned: The price of automatization. In G. Underwood & R. Stevens (Eds.), *Aspects of consciousness* (p. 67 – 89). London: AP.

Reske, M., Delis, D. C., & Paulus, M. P. (2011). Evidence for subtle verbal fluency deficits in occasional stimulant users: quick to play loose with verbal rules. *Journal of psychiatric research*, *45*(3), 361–368.