Title

An Investigation of the Prerequisites to Goal-Directed Control: Toward a Revision of the Habit Hypothesis of Obsessive Compulsive Disorder

Author

Ironside, Manon Louise

Publication Date

2023

Peer reviewed|Thesis/dissertation

An Investigation of the Prerequisites to Goal-Directed Control: Toward a Revision of the Habit Hypothesis of Obsessive Compulsive Disorder

By

Manon Louise Ironside

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Clinical Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sheri Johnson, Chair
Professor Anne Collins
Professor Steve Piantadosi
Professor Kiara Timpano

Fall 2023

Abstract

An Investigation of the Prerequisites to Goal-Directed Control: Toward a Revision of the Habit Hypothesis of Obsessive Compulsive Disorder

by

Manon L Ironside

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Sheri Johnson, Chair

Obsessive Compulsive Disorder (OCD) is characterized by intrusive and unwanted thoughts followed by specific patterns of behavior or mental acts that function to 'neutralize' those thoughts. These behaviors and mental acts, referred to as compulsions, share qualities with habits, responses triggered by a cue and independent of outcome values. In recent years, a number of studies have linked OCD to habit learning and failures in goal-directed control. It has been suggested that OCD, and disorders of compulsivity in general, are characterized by a bias toward learning habits and/or over-relying on habitual control at the expense of goal-directed control over behavior. The aim of the present study was to address two remaining gaps in our understanding of the relationship between OCD and failures in goal-directed control: first, to distinguish habit-related response styles from non-habit-related deficits in goal-directed control by modifying a commonly used slips-of-action test, and second, to examine more specific relationships between task performance and core motivational dimensions of OCD proposed to drive compulsive behavior. To address these goals, a sample of U.S.-based adult participants with symptoms of OCD were recruited from a mixture of undergraduate and community sources. One hundred fourteen participants completed an instrumental learning task with a subsequent slips-of-action test, and 82 of these also completed clinical self-report measures and a diagnostic clinical interview. The results support an alternative possibility to the habit hypothesis: that OCD is more strongly associated with failures in execution of goal-directed action than habit-related insensitivity to outcome devaluation. Further, results show that a commonly used task-based measure of habit-driven failures in goal-directed control in fact reflects multiple learning-related processes unrelated to habitual control, and what can appear as an increase in habitual control as measured by insensitivity to outcome devaluation on a slips-of-action test is strongly influenced by failures of learning implicit and explicit representations of causal relationships between stimuli, actions, and outcomes. These results implicate a need to investigate the intermediate cognitive processes between initial instrumental learning and execution of goal-directed control as they relate to OCD and related disorders.

Acknowledgements

I'd like to acknowledge all the research participants and therapy clients I've worked with over the course of this PhD as part of my training who have shared their struggles and strengths in managing mental illness. While this dissertation does not reflect any individual's experience, their stories of lived experience have deepened my understanding and appreciation for the essence of mental illness, including obsessive compulsive disorder, beyond any didactic input.

I feel lucky that my PhD experience was not an isolating one, and this was partly due to small-group seminars of graduate students and professors similarly passionate about psychology and cognitive neuroscience. I especially acknowledge professors who challenged me to reason through complex method sections, including Anne Collins, Rich Ivry, and Joni Wallis. Tom Griffith's 2017 course in probabilistic models of cognition was the first time I'd ever heard of Markov decision processes and considered quantitative models of sequential decision problems, and the modest portion of that course that I absorbed shifted the direction of my PhD and the way I think about goal-directed decision making. Steve Hinshaw's leadership of my first-year clinical proseminar inspired me to begin the PhD with purposeful intent, and to critically consider the broader paradigm of mental illness.

My primary mentor, Sheri Johnson, maintained a full scope of the forest at times when I was stuck in the trees. She generously let me learn from mistakes without chastising and bolstered my confidence in periods of self-doubt. I am especially grateful for her endless support, honest feedback, and scientific optimism.

I approached Anne Collins in my first year of the PhD with eagerness to understand but no technical knowledge of computational modeling of reinforcement learning, and she opened doors to levels of scientific inquiry I had never imagined I'd be capable of accessing.

Kiara Timpano's research support and clinical supervision was crucial in the recruitment and clinical diagnostic interviewing of research participants for this study. Her expertise on obsessive compulsive disorder enriched this study's development and deepened my clinical knowledge.

Yixin Chen, former UC Berkeley undergraduate, provided instrumental back-end development support of the web-based application used in my dissertation experiment in its early stages. Madeline Kushner and Lea Savoy provided invaluable support with study recruitment in their roles as research coordinators at the University of Miami.

I am deeply appreciative to fellow past and present UC Berkeley graduate students in psychology. To Maria Eckstein, Fred Callaway, and Rachit Dubey for kindling a passion for understanding reinforcement learning and decision making early in my graduate school experience. To my labmates Ben Swerdlow, Jen Pearlstein, Devon Sandel, and Matt Elliott for their camaraderie and thoughtfulness, and their generous input on research ideas.

My ability to complete a PhD was bolstered by a web of support outside of academia. I acknowledge in particular my parents, brothers, and fiancée as critical pillars of support.

And lastly, I acknowledge fellow scientists and friends, Christina Merrick and Maedbh King. Their friendships over the course of this PhD have meant the world to me.

An Investigation of the Prerequisites to Goal-Directed Control: Toward a Revision of the Habit Hypothesis of Obsessive Compulsive Disorder

Obsessive-compulsive disorder (OCD) is a disorder characterized by failures of goal-directed control. People with OCD typically experience intrusive and unwanted thoughts to a much greater extent than the general population and engage in specific patterns of behavior in an attempt to neutralize those thoughts. For example, someone with OCD might experience vivid, graphic thoughts of violently harming a loved one, despite having no desire to cause harm. These thoughts are experienced as very distressing. This person may then intentionally bring to mind 'good' thoughts, for example, in an attempt to neutralize the unpleasant obsession, or check on their loved one frequently to make sure they have not in fact caused them harm. These compulsive mental acts or overt behaviors ultimately exacerbate the frequency and distress of obsessions (Rachman, 1997). Without adequate treatment, OCD symptoms tend to endure over time and cause increasing functional impairment (Abramowitz, 2006).

The predominant OCD model of the past four decades is based on a cognitive-behavioral framework, and posits that obsessive thoughts - and specifically, over-interpretations of the significance of these thoughts - drive compulsions (Rachman et al., 1980). More recently, a competing theory has proposed that compulsions in OCD are produced via an overreliance on habitual control, possibly driven by deficits in goal-directed control that correspond with deficits in the brain's frontostriatal circuits (Gillan, 2021). While newer, this latter habitual control theory has gained considerable traction in the past decade. Within this framework, an explanation for obsessions has not yet been fully developed, but one possible role posited is that obsessions themselves represent automated, stimulus-driven thought patterns (Graybiel & Rauch, 2000). In a separate line of clinical research on OCD, there has been increased interest in the functional role of compulsions as they relate to persistent feelings of incompleteness versus harm-avoidance (Pietrefesa & Coles, 2008; Rasmussen & Eisen, 1992; Summerfeldt, 2004).

Currently missing from the habitual control theory is an explanation of why individuals with OCD show behavioral patterns consistent with an overreliance on habit in lab-based tasks, and how such patterns relate to clinical phenomenology more specifically than a general proposed relationship with compulsive behavior. Because the majority of experimental support for the habitual control theory of OCD stems from a single lab-based task (known as the Fruit Task), it stands to reason that gaining a deeper understanding of the cognitive processes driving performance on this task through experimental design and task manipulations may yield a more precise explanation of impacted goal-directed control in OCD.

What is Habitual Control, and How is it Measured?

Dual system models suggest the presence of two action-oriented systems that work in parallel: a goal-directed system, through which actions are guided by outcome expectation, and a habitual system, through which actions are triggered automatically by a stimulus, independently of expected outcome. It is largely accepted that most behaviors deemed 'habits' initially develop via the goal-directed system through instrumental learning. Over time, if a given action reliably leads to a desired outcome, the habit system 'takes over': actions are guided by the presence of a cue, irrespective of changes in outcome values. This transition in control systems over time is often adaptive. If an action is consistently effective in yielding a positive outcome, it becomes unnecessary to exert the additional computational effort necessary to calculate the expected value of an outcome anew each time a decision point arises, and more efficient to trigger action from

the presence of a stimulus. Habitual control becomes problematic only when the value of the outcome shifts to become detrimental to one's goals, and the action persists inflexibly.

It has been notoriously difficult to induce habit behaviors in human lab-based studies in the way they have been defined and well-studied through animal research, and thus difficult to measure clear habit-driven failures in goal-directed control in humans (de Wit et al., 2018). To show that impaired goal-directed behavior is driven by habit (and not some other process), it is essential to show that insensitivity to devalued outcomes occurs as a function of the amount of behavioral repetition during learning (Adams & Dickinson, 1981). Very few studies have succeeded in showing this relationship with human samples (Watson & Wit, 2018; though see Tricomi et al., 2009 for an example of successful habit induction in humans via overtraining and selective food satiation). There are a number of reasons for why this may be the case. It can be infeasible or unethical to subject human participants to the amount of overtraining necessary to induce habit-responding in response to devaluation, or it may be that goal-directed control is primed to outweigh prepotent habit responses in most lab-based environments. One workaround has been to use tasks that measure the relative balance between habit and goal-directed control, rather than directly measuring the influence of habitual control. One such task, often termed "the Fruit Task" or "Fabulous Fruit Task" due to its original use of fruit images as stimuli and outcomes, has provided much of the behavioral support for the habitual control theory of OCD.

The Fruit Task as a Measure of Action-Outcome Response Mapping Conflict

The basic design of the task first used to demonstrate a direct link between OCD and relative reliance on habit over goal-directed control was developed by de Wit & colleagues (2007). It was initially designed to test outcome-response theory, the idea that in reinforcement learning, humans and other animals map relationships between cues, actions, and outcomes in a stimulus-outcome-response structure. In the task, participants begin by learning through trial-and-error which of two keys to press in response to a stimulus to effect a rewarding outcome. After this learning phase, participants complete an instructed outcome devaluation test in which they are instructed to respond with the key that previously led to a presented outcome without being able to reference the original stimulus that led to that outcome. According to outcome-response theory, if the stimulus linked to one response is also associated with the outcome of a different response, this will produce response conflict in the ability to map actions to outcomes as well as insensitivity to outcome devaluation. The authors found results to support this theory in both humans and rats: when pairs of stimuli and outcomes were incongruent as opposed to congruent, participants made more errors mapping actions to outcomes (de Wit et al., 2007).

OCD-related Impairments in Goal-Directed Control on the Fruit Task

Gillan & colleagues (2011) modified the original Fruit Task to include a novel slips-of-action test in which participants had to respond selectively to stimuli in a time-pressured context according to new outcome values (some remained valuable, others were devalued). The slips-of-action test is structured like a go/no-go task; participants must respond quickly to still-valued stimuli and withhold responses to newly devalued stimuli. In this test, 'slips' of action refer to responses to stimuli linked to newly devalued outcomes for which the goal-directed action would be a withheld response. Researchers compared performance on the slips-of-action test between controls and participants with OCD and found that the OCD group demonstrated more 'slips' of action, especially when stimuli were part of an incongruent pair (Gillan et al., 2011; see Figure 1 in this manuscript for details on task interface). In this study, the OCD group also showed

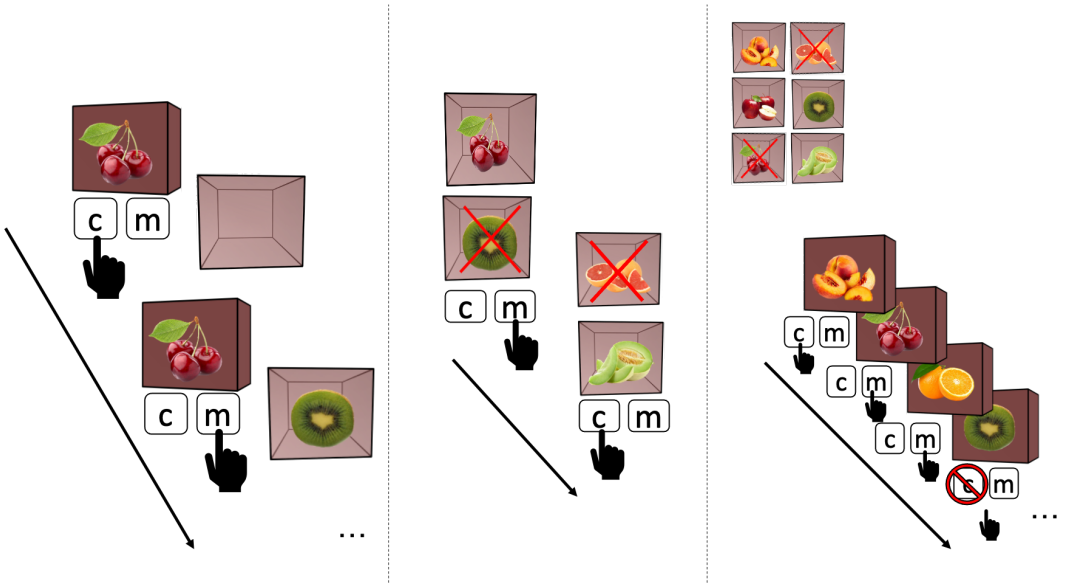


Figure 1. User interface for the 3 main phases of the Fruit Task in its original form (the explicit knowledge survey is not included in the figure). **Left**: During instrumental discrimination training, participants learned which of two keys to press in response to each fruit stimulus in order to gain a desired outcome, represented as a fruit inside of the box. **Center**: During an outcome devaluation test, participants press the key linked to the still-rewarding outcome. No feedback is provided during this phase. **Right**: For each block of the slips-of-action phase, participants are informed that 2 of the 6 fruit outcomes are no longer valuable. They are then presented with stimuli in rapid succession and had to press the correct key for stimuli linked to still-valuable outcomes, and withhold response to stimuli linked to devalued outcomes. No feedback is provided during this phase.

impaired explicit knowledge of outcomes linked to stimuli, as well as impaired action-outcome response mapping, especially for incongruent discrimination pairs. In other words, when presented with the same outcomes from learned stimulus-response-outcome sequences following the initial learning phase, participants with OCD were less likely to accurately respond with the action that had led to that outcome.

Other studies have since used the slips-of-action test using a design consistent to the Fruit Task to show purported habit-driven deficits in goal-directed control in the context of various forms of psychopathology, including substance use disorder (Ersche et al., 2016), alcohol use disorder (Sjoerds et al., 2016), Parkinson's Disease (de Wit et al., 2011), and nonclinical levels of OCD symptoms (Snorrason et al., 2016); though not all studies have supported a link between disorders involving compulsive behavior and decreased sensitivity to devaluation on the Fruit Task (e.g. null results comparing performance in adults with and without anorexia nervosa; Godier et al., 2016). Structural neuroimaging studies have shown links between brain regions known to be involved in habit vs goal-directed control and performance on the Fruit task: for example, 'slips' of action on the Fruit Task corresponded with increased structural connectivity between motor cortex and striatal regions (Delorme et al., 2016; de Wit et al., 2012) and conversely, fewer 'slips' of action (indicating greater goal-directed control) were associated with greater white matter tract strength linking the caudate to prefrontal cortical regions (de Wit et al., 2012).

Group-Level Performance Indicators Obscure Individual Learning Differences

While multiple of the studies mentioned in the previous paragraph have reported relationships between OCD or OCD-related symptoms and performance on the slips-of-action test, none directly considered the individual-level impact of learning (measured in phases of the task leading up to slips of action) on slips-of-action test performance, nor have they considered methods of identifying (and potentially removing from analyses) non-habit-related failures in goal-directed control during the slips-of-action test. The emergence of habit behavior critically depends on (over)learning stimulus-response contingencies to the point of automaticity. If these contingencies are not adequately learned during initial training, it is not possible for habit to override goal-directed control. In such cases, it would be inappropriate to interpret failures of goal-directed control as habit-driven responses in subsequent testing, even if those responses happened to coincide with the correct stimulus-response association.

The Fruit Task in full consists of 4 components: 1) an instrumental discrimination phase in which participants learn stimulus-response-outcome associations, 2) an instructed devaluation phase in which participants are shown only outcomes and must provide the correct response that led to that outcome, 3) a multiple choice survey to assess explicit knowledge of stimulus-response-outcome associations, and 4) a slips-of-action test in which participants are challenged to withhold previously learned responses to certain stimuli that have newly devalued outcomes while continuing to respond to still-valuable stimuli as stimuli are presented in rapid sequence with no feedback.

In its original form, the slips-of-action test defines goal-directed control as correctly withheld responses to stimuli with newly devalued outcomes. To demonstrate habit-driven failures in goal-directed control, difference scores are usually calculated between the number (or percentage) of responses to still-valued outcomes and the number (or percentage) of responses to devalued outcomes; when calculated per-person, this metric has been termed the "devaluation sensitivity index," or DSI (e.g., Snorrason et al., 2016; Yousuf et al., 2019; Watson et al., 2022). Higher DSI values are interpreted as showing increased sensitivity to devaluation and/or greater goal-directed control, and lower values as increased habitual control. While the DSI partially addresses the issue of differential base rates of responding, it does not distinguish participants who show overall low (but similar) rates of responding to both valued and devalued stimuli from participants who showed a fully habit-driven profile of responding – in both cases, participants could earn a percentage DSI score of zero, but the score would reflect different behavioral mechanisms driving responses. If participants make 'slips' of action that reflect habit-driven responses, it would be expected to see relatively high 'hit' rates for still-valued stimuli; if participants show relatively low hit rates for still valued stimuli, it may reflect other failures in goal-directed control or poor learning of the initial stimulus-response contingencies.

A Preliminary Experiment: Fruit Task Performance in an mTurk Sample

To enhance understanding of the fruit task, I conducted a preliminary experiment with a nonclinical sample of 135 adults recruited through Amazon Mechanical Turk as part of a larger study, details of which can be found on Open Science Framework (https://osf.io/x72y9/). Self-report indices of psychopathology were administered to these participants, including the Obsessive Compulsive Disorder-Revised (OCI-R), a well-validated measure of OCD symptom severity (Foa et al., 2002). The goals of this experiment were threefold: 1) to replicate previous findings associating OCD symptoms with habit-driven failures in goal-directed control with a larger sample size, 2) to examine relationships between learning metrics on earlier phases of the Fruit Task with performance during the slips-of-action test, and 3) to address the issue of equal

DSI values signifying distinct behavioral control mechanisms by taking a novel signal detection approach to identify and exclude participants with below-chance levels of stimulus discriminability. The specific learning metrics of interest included explicit knowledge of stimulus-outcome relationships and action-outcome response mapping performance on the instructed outcome devaluation phase. This latter measure represents the ability to correctly respond to an outcome (in the absence of the original stimulus) with the action that previously led to that outcome during instrumental learning.

Parsing the Slips-of-Action Test with a Signal Detection Approach

To better distinguish cognitive processes leading to low DSI values, I used a signal detection approach to derive estimates of discriminability and response bias on the slips-of-action test (see Appendix B for detailed methods). Figure B1 demonstrates how DSI alone can obscure meaningful differences in hit rates for still valuable stimuli. When using d' to set thresholds for inclusion, lower DSIs at higher hit rates are included, while equivalent DSIs at lower hit rates are excluded. Figure B2 shows a substantial overlap in DSI scores between participants who were and were not identified as responding with chance levels of discriminability on the slips-of-action test.

The Fruit Task was implemented as described in Gillan & colleagues (2011) with two main modifications: the length of the training phase was increased from 96 to a minimum of 168 trials with implementation of a learning criterion to ensure that all participants learned stimulus-response relationships to near automaticity, and the number of instructed outcome devaluation trials was doubled to better distinguish performance on this measure. Details of modifications from the original Gillan & colleagues (2011) task design can be found in Appendix B.

Results

Higher discriminability scores based on slips-of-action test performance corresponded with higher overall levels of responding on still-valuable trials (r_s = -.49, $p < .0001$). Examination of participants with high false alarm rates showed that the discriminability threshold allowed inclusion of fully habit-driven responders with devaluation scores of zero while excluding responders with higher DSIs but increased error responses or indiscriminate non-response styles.

At a group level, participants showed characteristic learning curves and increases in response speed over time during the instrumental learning task. Having complete vs incomplete explicit knowledge of stimulus-outcome relationships had a significant effect on instructed outcome devaluation performance ($W = 2800$, $p < .0001$, $r = .35$) as well as devaluation sensitivity on the slips-of-action test ($t(131.4) = 7.98$, $p < .0001$, Cohen's $D = 1.3$, $CI_{95} = 1.03$ - 1.59). When participants with chance levels of discriminability were removed from analyses, an even stronger effect of explicit knowledge on devaluation sensitivity emerged ($t(74 = 8.5$, $p <$ .0001, Cohen's $D = 1.7$, $CI_{95} = 1.17$ - 2.19). Action-outcome response mapping similarly had a relatively strong effect on devaluation sensitivity during the slips-of-action test (r_s = .48, $p <$ .0001, $CI_{95} = .33$ - .60); this relationship also strengthened with removal of participants with chance levels of discriminability (r_s= .64, $p < .0001$, $CI_{95} = .51$ - .74).

As expected, analyses indicated that OCI-R scores were unrelated to explicit knowledge. That is, there was no difference in OCI-R scores between participants who showed complete vs. incomplete explicit knowledge of stimulus-outcome relationships as measured with a multiple choice questionnaire ($t(128) = -1.54$, $p = .13$, Cohen's D = -.29); however, participants with higher OCD symptom severity scores showed poorer action-outcome response mapping performance (r_s = -0.18, $p = .04$); the strength of this relationship did not change when excluding

participants with below chance discriminability (r_s = -.21, p = .04). Contrary to predictions, there was no significant relationship between devaluation sensitivity and OCD symptom severity before (r_s = -.13, p = .13) or after excluding participants with chance level discriminability (r_s = -.14, p = .18) .

Discussion

While we did not find an expected relationship between OCD symptom severity and devaluation sensitivity during the slips-of-action test, we did find an association between action-outcome response mapping performance and OCD symptom severity, and a strong positive relationship between action-outcome response mapping and devaluation sensitivity. These results suggest that OCD symptom severity may be equally or more relevant to action-outcome response mapping than habit-driven failures in goal-directed control. Consistent with previous research on learning in OCD, we did not find evidence to support problems with explicit stimulus-action-outcome knowledge in participants with increased OCD symptom severity. Our signal detection approach to excluding participants based on chance levels of discriminability provided a theory-driven method of removing data contributing error variance to slips-of-action test performance. Previous studies suggested that DSI scores at or below zero represented chance performance during the slips-of-action test; however, this benchmark conflated random responders from participants with fully habit-driven response profiles. Our method allowed us to differentiate indiscriminate responding from habit-driven responders and exclude the former, but not the latter.

While the habit hypothesis of OCD offers some compelling links between OCD symptoms, particularly compulsions, and the habitual control system, it is still not clear whether there is a direct relationship between overreliance on habit and compulsions in OCD. One problematic aspect of the slips-of-action test is the ambiguity of whether a 'slip' of action truly demonstrates a habit-driven failure in goal-directed control, or the consequence of some other aspect of learning. A 'slip' of action in the Fruit Task could represent memory failure of the outcome associated with the presented stimulus, and/or memory of whether or not it was devalued during that particular test block. For example, Gillan et al. (2011) found that participants with OCD showed more 'slips of action' than controls, but also (consistent with what we found in experiment 1) performed worse during the instructed outcome devaluation phase. Poor performance during this phase cannot represent overriding habitual responding due to absence of the cue, or stimulus. It is possible that action-outcome response mapping problems, which appear related to OCD diagnostic status and symptom severity, have resulted in apparent habit-like responding during later stages of the task. If OCD is not primarily a disorder of habit-driven failures in goal-directed control, differences in learning acquisition may explain the apparent shift toward habitual control on lab-based tasks.

OCD: Disorder of Learning?

Several findings suggest that performance differences on learning tasks in OCD may be driven by the more cognitive process of information-gathering and value updating rather than behavioral differences, like perseverative or otherwise automatic response patterns. For example, in one study, participants with high levels of compulsivity symptoms sampled more information relative to participants with low levels of compulsivity symptoms before making a decision in a sequential sampling task when there was no penalty for additional sampling choices (Hauser et al., 2017). In a separate study involving a paradigm that required learning from feedback, patients with OCD over-corrected their responses based on feedback, leading to decreased performance accuracy (Vaghi et al., 2017). Interestingly, in spite of this behavior mismatch, patients with

OCD indicated equivalent confidence in their responses to controls, which reflected the reality of task contingencies (e.g. decreased confidence following a change in underlying contingencies, and increasing confidence as task stability increased). This suggests that explicit awareness of environmental contingencies may be intact in OCD, and that deficits in goal-directed performance may be linked to mapping responses to feedback.

This conceptualization is consistent with a computational model of OCD proposed by Fradkin & colleagues (2020), in which *state transition uncertainty* was postulated as a core cognitive feature of OCD, resulting in poor goal-directed control of behavior. State transitions refer to the mapping between one's current state (a state includes all goal-relevant features of one's present environment in space and time) and what one's future state will be, given a particular action. Someone with increased state transition uncertainty will find it difficult to predict changes resulting from actions taken and may not trust outcomes as a reliable source of information to predict future consequences. For example, imagine leaving your house (state 1), locking the door (action), and, while standing on the doorstep, knowing that the door is locked (state 2). In the context of OCD, the certainty that locking the door (action) led to the door being locked (state 2) may deteriorate in a short amount of time, causing repeated returns to the door to check whether it is, in fact, locked. The link between state, action, and new state is encoded with a high degree of uncertainty. Rather than indicating poor episodic memory, this particular uncertainty may be better characterized as deficits in meta-memory, or memory confidence (e.g. Radomsky & Alcolado, 2010), as studies have found intact memory accuracy but decreased confidence about memories in adults with OCD (Boschen & Vuksanovic, 2007; Radomsky et al., 2014).

A number of recent studies focusing on the relationship between OCD or compulsivity and decreased goal-directed control have proposed a key role of maintaining an accurate representation of task structure to successfully execute goal-directed control. Seow & colleagues (2021) that symptoms of intrusive thoughts and compulsivity related to decreased model-based planning on the 2-step task as well as decreased suppression of parietal-occipital alpha power (as measured with electroencephalography) following uncommon state transitions. They interpreted this decreased suppression as decreased sensitivity to task structure and possibly a failure to accurately represent state transition probabilities. A study specifically examining the role of state transition learning on model-based control using computational modeling found that participants with higher compulsivity showed impairments in state transition learning, which corresponded with updating state transition values too quickly in stable learning environments and too slowly in volatile learning environments (e.g., shifting contingencies), resulting in suboptimal task performance (Sharp, Dolan, & Eldar, 2023).

State transition uncertainty is consistent with action-outcome response mapping deficits: if a particular outcome is not encoded as a reliable consequence of a given stimulus and response, we would expect to see deficits in the instructed outcome devaluation test, which requires backward mapping from outcome to response. We would not expect state transition uncertainty to affect stimulus-response learning, just as someone with OCD who has checking compulsions would not be expected to struggle with automatically locking the door on leaving the house. Similarly, we would expect problems to arise during interactive experience but not with respect to accruing explicit knowledge of environmental contingencies (for example, someone with the OCD checking compulsion described above could still accurately explain that the action of locking a door leads to the outcome of a locked door).

Toward a more specific link between goal-directed response deficits and clinical models of compulsions

If action-outcome response mapping in OCD indeed corresponds with a cognitive profile of excessive uncertainty about state transitions, it stands to reason that people with OCD who primarily experience a persistent sense of incompleteness in their actions (as opposed to anticipatory anxiety regarding a particular feared outcome) may show greater deficits in goal-directed control that depends on interactive accrual of information. In the proposed study, I focus on building a more precise link between failures in goal-directed control and the clinical dimension of incompleteness, a feature present in varying degrees among people diagnosed with OCD and explained in greater detail below. Further, though results from Experiment 1 appear to support goal-directed deficits via response-mapping problems that correspond with higher levels of OCD symptoms, it is still possible that OCD is also characterized by habit-driven failures in goal-directed control. To address this issue, other measures of habit-driven influences on behavior below and adjust the slips-of-action test accordingly.

Motivation Model of OCD: Core Dimensions of Incompleteness and Harm-Avoidance

Compulsions can differ both in form and in function. While patients with OCD have been classically grouped based on the form of their compulsions (e.g. washers, checkers), the differing functions of compulsions have garnered increasing interest in recent years. In particular, some compulsions in OCD are performed in order to avoid a feared outcome, while others are completed to decrease a 'not-just-right' feeling, untied to any specific imagined outcome. Two compulsions that appear identical in behavioral manifestation may serve different functions. A person may wash their hand compulsively to avoid potential contamination (harm-avoidance), or a person may do the same act repeatedly because they continued to experience a 'not-just-right' sensation after each of the first few times they washed and dried their hands (incompleteness). Importantly, in this second scenario, no articulable fear of contamination is present. Incompleteness and harm-avoidance as core dimensions of OCD were first proposed by Rasmussen and Eisen (1992) in the context of a motivational model of the disorder. Since then, there has been much interest in the functional roles of these dimensions, especially incompleteness. While harm avoidance corresponds with models of anxiety disorders, incompleteness is not a feature of anxiety disorders. Further, it appears that incompleteness is more strongly related to overall OCD symptom severity than harm avoidance (Ecker & Gönner, 2008; Sibrava et al., 2016). Indirect evidence also supports this relationship; for example, it has long been established that patients unable to articulate clear consequences of their fear have shown poorer treatment response to exposure and response-prevention (ERP), the gold-standard treatment for OCD (Foa et al., 1999). Harm-avoidance and incompleteness have been proposed to exist orthogonally rather than two ends of a spectrum, though studies examining factor structure of these constructs have found moderate correlations between them, even as factor analysis has consistently supported 2-factor models (Pietrefesa & Coles, 2008; Summerfeldt et al., 2014). The most commonly used measure designed to specifically assess both harm-avoidance and incompleteness is the Obsessive Compulsions Core Dimensions Questionnaire (OC-CDQ; Summerfeldt et al., 2001), which has been empirically validated in clinical and nonclinical samples (Pietrefesa & Coles, 2008; Summerfeldt et al., 2014).

It has been proposed that incompleteness is driven by problems integrating sensory feedback to guide subsequent behavior (Summerfeldt, 2004). This explanation is consistent with the behavioral study findings in OCD samples discussed above. Next, I will describe how

measurement of habit-driven failures in goal-directed control was fine-tuned in this study in order to show more clearly whether errors on the slips-of-action test are truly habit-driven, or caused by failures in other goal-directed mechanisms.

Reaction time as a more precise measurement of habit-driven influences on goal-directed control

In Experiment 1, our signal detection analysis showed that the same DSI score could reflect differential influences of habitual control and therefore lacked the precision necessary – especially at low values – to clearly demonstrate the degree of habitual control over behavior. In recent years, some researchers have suggested a focus on reaction time as opposed to overt choice to measure habit responses with increased precision. One study showed that shortening the time available for response preparation revealed habit-like responding (old patterns of response to a newly devalued stimulus) whereas they found no evidence for habit-like behavior when participants were given more than ~600ms to respond following presentation of the stimulus (Hardwick et al., 2019). This finding highlighted a major problem with using response selection alone as a measure of habitual control, especially when responses are self-timed and allowed too much time – in the case of their task, more than 600ms – to respond. It is possible that the goal-directed system is usually able to override habit even if a habit response is indeed prepared, and the self-timed nature of many behavioral tasks could mask the existence of prepared habit responses. Indeed, Luque & colleagues (2020) showed results to support this possibility: when using overt response selections to devalued stimuli as a metric of habit, they found no effects of overtraining, which would be expected if the task successfully induced habitual responding. However, they did find an overtraining effect when they used a response time metric of habit. They proposed that habit can be conceptualized as the cost of time to correctly switch a response to a newly devalued stimulus, relative to baseline response time to the same stimulus when still valuable. They reasoned that a longer response time to devalued stimuli (in the event of a correct response) indicated greater habit-driven interference.

In addition to supporting reaction time measures of habit, these findings prompt a reconsideration of the language used around habits. For example, if it is true that habit responses are very often prepared, and it is possible to 'uncover' them by manipulating response timing demands, we should be cautious in inferring that people who show failures of goal-directed control have a 'tendency toward' or 'proclivity for' developing habits. It may be the case – and indeed Hardwick & colleagues' (2019) results suggest – that most people have a proclivity toward habit, and those who show habit-like response deficits have a primary deficit with the overriding goal-directed system, which may result from multiple possible mechanisms.

Increasing identifiability of goal-directed behavior during the slips-of-action test

Using a signal detection approach in our first study provided a data-driven method of excluding participants with chance levels of stimulus discrimination without removing participants with low DSIs due to habit-driven failures in goal-directed control. However, this method set a low bar for exclusion, and still did not fully discriminate behavioral mechanisms leading to high non-response rates across valuable and devalued stimuli. It is possible that some participants took an intentionally cautious approach and only paid attention to devalued stimuli, correctly withholding responses to these stimuli but also withholding responses to still valuable stimuli. It is also possible that participants found the slips-of-action test too difficult and withheld responses out of confusion. Non-responses may indeed signal goal-directed control, but they might also indicate uncertainty or inattention.

In the present study, participants were instructed to press a third key (space bar) in response to stimuli linked to devalued outcomes rather than withholding a response. The inclusion of an overt goal-directed response to trials with devalued stimuli allowed us to a) capture reaction-time information related to the balance of habit/goal-directed control and b) obtain a more specific measure of goal-directed control, less likely to reflect task-irrelevant processes.

Goals and Hypotheses

This study included three major aims. First, to gain a more nuanced understanding of the failures in goal-directed control seen in OCD that have been demonstrated relatively consistently across studies; Second, to replicate a specific relationship between habit-driven failures in goal-directed control and OCD symptom severity relative to dimensions of psychopathology unrelated to compulsive behavior; and third, to test the prediction that core OCD motivational dimensions of incompleteness and harm-avoidance have a differential impact on learning action-outcome relationships which in turn moderates the outcomes of goal-directed control.

Results from my first experiment suggested that a deficit in action-outcome response mapping may be driving what appeared to be habit-driven errors in the slips-of-action test; in this second study, I addressed methodological confounds and recruited a symptomatic sample to further clarify the link between stimulus-outcome learning and slips-of-action performance in general and in relation to OCD symptoms. I hypothesized that action-outcome response mapping performance, but not explicit knowledge of action-outcome relationships, would have a strong influence on goal-directed control, the execution of which depends on accurate implicit representation of action-outcome relationships. However, I did not expect action-outcome response mapping to relate to the purportedly more sensitive response time difference measure of habit, as this more precise measure should be further differentiated from goal-directed control, and it is not expected that components of learning (e.g. action-outcome response mapping) would impact the degree of overriding habitual control.

In this study, I conceptualized differential effects of these task-related parameters to two distinct core dimensions of OCD, incompleteness and harm-avoidance. I expected these two core dimensions to relate distinctly to failures of goal-directed control rather than habitual control. That is, I expected self-reported incompleteness to account for variance in action-outcome response deficits, whereas I expected no such relationship between self-reported harm-avoidance and action-outcome response mapping. Given the hypothesis of this paper that goal-directed control, but not habitual control, is impacted in OCD, I did not expect an effect of incompleteness, harm-avoidance, or general OCD symptom severity on the response time difference measure of habit-driven failures in goal-directed control. Further, I expected a nonsignificant role of OCD core dimensions and overall OCD symptom severity on devaluation sensitivity once accounting for the effect of action-outcome response mapping performance. Finally, a stronger relationship between OCD symptom severity and goal sensitivity relative to devaluation sensitivity was hypothesized. This relationship was predicted to remain significant after accounting for the influence of action-outcome response mapping performance.

Method

All procedures for this study were approved by the institutional review boards at both universities before data collection commenced. The experimental design, detailed task description, and

multiple regression data analysis plan were pre registered prior to data collection (https://osf.io/82ypd).

Participants

Participants were recruited from multiple community sources with oversampling for clinically significant symptoms of OCD. All participants completed the study virtually, allowing for broad recruitment across the United States, with main recruitment hubs in major metropolitan areas of the 1) southeast and 2) west coast of the United States (see Table 1 for detailed demographic information). Exclusionary criteria included age younger than 18 or older than 60, diagnosis of schizophrenia, history of head injury or neurological illness, severe substance use disorder or alcohol use disorder within the past 6 months, and uncontrolled endocrine or neurological disease that may interfere with task completion or confound diagnosis. We also included a single inclusion criterion to ensure a range of clinically significant OCD symptoms, defined as presence of any OCD-related symptom at time of study completion. This criterion was determined based on scores greater than 11 on the 12-item version of the Obsessive Compulsive Inventory-Revised (OCI-12) for participants sampled outside of OCD-specific participant pools (e.g., undergraduate research participants). Participants sampled within OCD-specific populations (International OCD Foundation website, OCD support groups, participants who met diagnostic criteria for OCD in lab-based structured clinical interviews within the past year) were not required to meet additional OCD-related criteria for inclusion. Undergraduate research participants were compensated with research credits toward course requirements at their respective undergraduate institutions. Community-based participants were compensated $10 for Fruit Task completion and an additional $20 for completion of the clinical phone interview and clinical self-report surveys.

A total of 192 participants were recruited from a combination of sources comprising 1) undergraduate students opting to complete research studies in exchange for course credit, 2) community members local to either study site who had endorsed clinically significant symptoms of OCD in past lab-based structured clinical interviews and who had indicated interest in being re-contacted for future studies, 3) community members participating in OCD support groups local to the San Francisco Bay Area in California, 4) United States-based members of the public accessing the list of research studies posted on the International OCD Foundation website, and 5) community members accessing the r/OCD reddit sub-thread.

Out of 192 participants originally identified as eligible based on inclusion/exclusion criteria, 38 were excluded from analyses following data collection for the following reason: It was discovered that many of the participants from the r/OCD reddit subthread did not have any OCD symptoms and were participating from a similar international location. Additionally, some of these participants were identified as having completed the task more than once when they were contacted over the phone for the clinical interview portion of the study and the experimenter recognized the vocal idiosyncrasies of the participant. Due to multiple issues with data integrity, the experimenter made the conservative choice to exclude all participants recruited via Reddit from analyses. 2) **Two participants endorsed symptoms consistent with current severe substance or alcohol use disorders during the clinical interview.** 3) Five participants did not complete the post-task survey used to verify status as an invited research participant, their study participation was considered incomplete. In total, 154 participants completed the Fruit Task and

Table 1

Sociodemographic Characteristics of Participants.

	Clinical Analyses Sample (n = 82)		Task Analyses Sample (n = 114)	
	n	%	*n*	%
Gender				
Female	57	70	77	68
Male	22	27	33	29
Non-binary	2	2.4	3	2.6
Declined to respond	1	1.2	1	0.8
Ethnicity				
Hispanic/Latino/a/x	17	21	19	17
Non-Hispanic/Latino/a/x	64	78	93	82
Declined to respond	1	1.2	2	1.8
Race				
White	56	68	74	65
Asian	18	22	26	23
Black or African American	5	6.1	9	7.9
Multiracial	1	1.2	3	2.6
Declined to respond	2	2.4	2	1.8
Highest Education Level				
High School graduate	10	12	16	14
Some college	22	27	44	39
2-year college degree	3	3.7	3	2.6
4-year college degree	40	49	32	28
Professional degree	2	2.4	2	1.8
Master's degree	12	15	12	11
Doctorate degree	3	3.7	3	2.6
Declined to respond	0	0	2	1.8
Income				
Less than $10,000	11	13	11	9.6
$10k - 29,999	7	9	9	7.9
$30k - 49,999	13	16	13	11
$50k - $69,999	4	4.9	4	3.5
$70k - $99,999	14	17	18	16
$100k - $149,999	9	11	11	9.6
Over $150,000	9	11	20	18
Declined to respond	15	18	28	25
Recruitment source				
Undergraduate research participation programs	29	35	59	52
Community	53	65	55	48

Note. Median age of clinical measures completion sample = 21 years old, Median age of task completion sample = 24 years old. Age of participants was skewed positive in both samples.

were considered eligible for inclusion. Of these 154, 114 successfully met the learning criterion on the Fruit Task (60 of whom were undergraduate research participants, 54 of whom were community-based adults, learning criteria exclusion rates did not differ by undergraduate vs. community samples, $\chi^2(1) = .62$, $p = .43$). Eighty-four out of 114 eligible participants elected to complete session 2, including the structured clinical interview and clinical survey measures. Two participants were subsequently excluded due to moderate-severe alcohol or substance use disorders within the past 6 months. This left a final sample of $n = 114$ for task-based analyses and $n = 82$ for analyses including diagnostic information and clinical survey measures. The timing of exclusions and full study flow are shown in Figure A1. Frequencies of OCD diagnoses and comorbidities for undergraduate and community samples are shown in Figure A2.

Procedures

All participants completed written informed consent procedures before completing study measures. All data collection was completed remotely and divided into two separate sessions that occurred approximately 1 week apart. In session 1, participants completed a 30-minute online task (the Fruit Task) and demographic information was collected via an online survey. Participants who successfully passed a pre-defined learning criterion on the task were invited to participate in the second session, which involved a 30-minute structured clinical interview over the phone and additional symptom severity survey measures completed online. Participants completed the symptom severity questionnaire measure while still on the phone with the experimenter in case clarifying questions emerged during completion.

Task

The Fruit Task was designed according to the description in Gillan & colleagues (2011) (see Appendix B for details on minor modifications from the original design) and with the following additional adaptations:

Instrumental discrimination training. Previously, we used a learning criterion of >87% correct responses across all stimuli for at least 2 blocks following the first 4 instrumental learning blocks. Because we did not count accuracy separately by stimulus, it would have been possible for participants to respond incorrectly to 3 out of 4 presentations in a block for the same stimulus and still meet the learning criterion if all other responses were correct. This could have led to uneven learning which would have impacted our interpretation of devaluation performance by pair type (congruent, incongruent, standard) for a given participant. In experiment 2, the learning criterion included >87% accuracy AND >=75% accuracy for each individual stimulus for 2 additional blocks (each block comprises 24 trials) following the initial 72 training trials. Therefore, the minimum number of training trials was 120 and the maximum eligible number of trials was 288.

Instructed outcome devaluation test. In the previous task version, we included explicit instructions with images for this phase, but we did not include practice trials. It is possible that some cases of poor performance on this phase of the task were due to a misunderstanding of task instructions rather than a deficit with response-outcome mappings. Though we can partially solve this confound by checking performance on standard pairs (in which correct response depends entirely on response-outcome mappings), we hoped to increase performance accuracy by implementing practice trials in which all the necessary information was present, so that responses reflected understanding of the task instructions without reliance on memory. Practice trials for this phase included images of fruit not used during the training phase. Participants repeated

practice devaluation trials with two standard pairs and two incongruent pairs until they responded correctly at least four times in a row to each pair.

Slips of Action Test. In this study, participants were asked to press the spacebar in response to stimuli linked to devalued outcomes rather than withholding responses. This marks a difference from prior versions of the task in which goal-directed responses were characterized by response omission rather than commission. Other modifications to this phase included adjusting the pseudo-randomization of stimulus presentation to ensure that the same stimulus was never repeated in consecutive trials, and that the two types of devalued outcomes in each block were paired accordingly for each participant: 1) incongruent and standard, 2) congruent and standard, and 3) congruent and incongruent. The specific discrimination pair within each pair of devalued outcomes was selected to ensure different responses between the stimuli (e.g. if the two devalued stimuli were from congruent and standard pairs and the congruent outcome was linked with the "c" key, the devalued outcome from the standard pair would be the one linked with the "m" key). The main purpose of these modifications was to ensure equal difficulty across all participants and blocks of the Slips of Action test while maintaining as much randomization of stimulus presentation and specific images associated with each discrimination pair as possible. Finally, based on pilot testing of this version that showed participants struggled to respond within the timeframe between stimuli, the stimulus presentation time during this phase was increased from 1 to 1.25 seconds.

The Fruit Task was programmed in JavaScript with use of De Leeuw's (2015) JsPsych library according to the task description in Gillan et al. (2011). After local testing, the task was set up as a Node.js web application and hosted for free on Heroku[1]. Data was collected without personal identifying information and sent to a password-protected MongoDB Atlas cloud database. A pseudo-random 10-digit number beginning with a pre-designated identifier (based on whether the participant passed or failed to meet the learning criterion) was produced at completion of the experiment for each participant, and they were instructed to copy and save this value for later use before the webpage automatically redirected to a Qualtrics survey where they manually entered this number along with demographic information in order to receive compensation for Fruit Task completion and continue in the study. This process allowed us to ensure that a) all entries to the Qualtrics survey were matched with a unique task completion identifier and b) allowed the experimenter to proceed with study flow (e.g. compensating the participant based on task completion and advising them of full study completion if they did not pass learning criterion, or otherwise inviting the participant to participate in the clinical interview session) without close inspection of the data prior to full completion of data collection.

Measures

Clinical Measures

Mini International Neuropsychiatric Interview version 7.0. (MINI). The MINI (Sheehan, 2014) is a commonly used brief structured clinical interview for mental health diagnoses. The MINI was used in this study to establish diagnostic history of participants with regards to obsessive compulsive disorder, mood disorders, anxiety disorders, psychosis, substance and alcohol use disorders, posttraumatic stress disorder, and eating disorders. In addition to its use in establishing presence or absence of current OCD and comorbid diagnoses,

[1] At the time this experiment was conducted, Heroku provided a free tier to host a limited number of web applications for personal use. As of November 2022, Heroku no longer provides any free plans.

the MINI was also used for exclusionary diagnoses of severe substance or alcohol use disorders and schizophrenia. All MINI interviews were completed by two trained clinical interviewers in clinical psychology PhD programs who consulted weekly on complex diagnostic presentations with a licensed psychologist and expert in OCD diagnosis and treatment (Kiara Timpano). The MINI achieves good inter-rater and test-retest reliability and correspondence with more comprehensive, validated measures of clinical diagnosis. Diagnostic consensus was reached for all participants prior to examining self-report or task-related data.

Obsessive-Compulsive Trait Core Dimensions Questionnaire (TCDQ). The TCDQ was developed to measure the core motivational dimensions of harm-avoidance and incompleteness as they relate to OCD symptoms, the primary OCD-related variables of interest for the current study (Ecker & Gönner, 2008; Summerfeldt et al., 2014 for validation with confirmatory factor analysis). The TCDQ has been commonly used and validated in both psychiatric and non-psychiatric samples. This scale includes 20 items, each rated on a 5-point Likert scale from 0 (never applies to me) to 4 (always applies to me). In this sample, both dimensions of the TCDQ had high internal consistency (harm-avoidance $\alpha = 0.92$, incompleteness $\alpha = 0.89$), and the two dimensions were moderately correlated ($r(80) = 0.54$, $p < .0001$).

Not-Just-Right-Experiences Questionnaire From the Outcome Assessment Information Set (NJRE-OASIS). The NJRE-OASIS includes 10 yes-or-no items assessing the presence of common not-just-right experiences over the past month, and seven additional items assessing frequency and severity of not-just-right experiences (Coles & Ravid, 2016). The first 10 questions can be summed (yes = 1, no = 0) to derive a metric of the breadth of not-just-right experiences endorsed over the past month. Only participants who indicated that they had experienced at least one NJRE in the past month were asked to answer o the seven frequency and severity scale items, which were rated on a likert scale ranging from 1 = "not at all" to 7 = "extremely" and summed to derive a metric of NJRE intensity. In this sample, the seven items comprising the NJRE intensity metric had high internal consistency ($\alpha = 0.94$). We will use the total number of NJREs and NJRE intensity metric to examine convergent validity of the incompleteness measure from the OC-TCDQ.

Dimensional Obsessive Compulsive Scale (DOCS). The DOCS is a widely used scale designed to measure severity of OCD symptoms with items that parallel DSM-5 diagnostic criteria. It includes 20 total items, comprised of 5 symptom severity items for each of the 4 most common OCD symptom dimensions (contamination, responsibility for harm, unacceptable thoughts, and symmetry). It has strong internal consistency, test-retest reliability, and convergent, construct, and divergent validity (Abramovitz et al., 2010).

Hoarding Rating Scale (HRS). This brief self-report questionnaire includes questions about clutter, having a hard time getting rid of things/throwing things away, and distress regarding hoarding behaviors (Tolin et al., 2014). Though hoarding disorder is considered a separate disorder from OCD as of DSM-5 (American Psychological Association, 2013), hoarding symptoms commonly co-occur with OCD (e.g., Matthews et al., 2014). Like OCD, hoarding disorder is defined in part by compulsive behavior, the dimension proposed to have a specific relationship with failures of goal-directed control. We will examine the relationship of hoarding symptom severity with measures of learning and habit vs goal-directed control. The HRS had high internal consistency within our sample ($\alpha = 0.88$).

Depression, Anxiety, and Stress Scale (DASS). The DASS is a widely-used measure including well-validated and reliable subscales for depression, anxiety, and stress (Brown et al., 1997). Mood and anxiety disorders are the most common comorbid psychiatric conditions among adults with OCD (Sharma et al., 2021). The depression and anxiety subscales from the DASS will be included as covariates in multiple regressions to assess the specificity of OCD-related effects. Both depression and anxiety scales had high internal consistency within our sample (DASS depression $\alpha = 0.93$, DASS anxiety $\alpha = 0.85$).

Task-related measures

Pre- and Post-Task Ratings of Low Mood and Intrusive Distress. Participants were asked to self-report levels of low mood and intrusive distress both before and after completing the task. Low mood was measured on a single item, with ascending value options defined as "Very slightly or not at all", "A little", "Moderately", "Quite a bit", "Extremely". Intrusive distress was similarly measured with a single item, and options included: "None", "Not too disturbing", "Disturbing, but still manageable", "Very disturbing", "Near constant and disabling distress". These values were recoded from 0-4, with greater values indicating lower mood and greater intrusive distress, respectively. Some levels of low mood and intrusive distress were expected in our sample because our inclusion criteria required clinically significant OCD-related symptoms. In this study, these ratings were used to assess whether in-the-moment levels of low mood or intrusive distress a) impacted participants' ability to successfully meet the learning criterion or b) increased or decreased as a function of task completion.

Habit and Goal-Directed Choice Measures. Individual level habit-driven response conflict was measured by a devaluation sensitivity index (DSI), computed as the difference between percentage of previously learned responses to still-valuable standard stimuli and percentage of previously learned responses to devalued standard stimuli on the slips-of-action test. Goal sensitivity was similarly computed but by examining the difference in rates of goal-directed (spacebar) responses to devalued versus still-valuable stimuli. Percentages were used rather than raw values to account for base rates of responding. Differences were computed such that higher values of both DSI and goal sensitivity referred to participants with less habit-driven interference and greater goal-directed control, respectively. Both DSI and goal sensitivity were calculated based on responses to standard discrimination pairs only to avoid stimulus-outcome confounds present with congruent and incongruent pairs (de Wit et al., 2012).

Habit-Driven Response Time Conflict. This metric was calculated by subtracting baseline response times (response times to stimuli during blocks in which they were still valuable) from goal-directed response times for the same stimuli during blocks in which they were devalued. The response time switch cost, as proposed by Luque and colleagues (2020), is thought to capture covert habit-driven conflict evidenced by increased response times even when goal-directed control successfully overrides prepotent habit responses in time to make the correct overt choice. According to this idea, higher values of this variable indicate greater habit-driven conflict. Response time conflict was computed for standard discrimination pairs only.

Action-Outcome Response Mapping Accuracy. Action-outcome response mapping accuracy was calculated as the total number of correct responses on standard discrimination pairs during the Instructed Outcome Devaluation phase. Performance on congruent pairs were not included in this summary score because stimulus-response and response-outcome mapping were confounded on these trials, and accuracy may reflect either or both of these learned associations.

Incongruent discrimination pairs were designed to induce response conflict during this phase; performance on these pairs was examined separately.

Explicit Knowledge of Stimulus-Response-Outcome Relationships. Knowledge of stimulus-response-outcome relationships following instrumental discrimination training was assessed using a multiple choice survey, in which participants viewed each stimulus sequentially and were asked to 1) select the correct response to that stimulus ("m" or "c") and 2) select which of six outcomes, presented visually exactly as they appeared during training, followed that particular stimulus, given the correct response. Explicit knowledge of stimulus-response associations and stimulus-outcome associations were examined separately in analyses. In cases where explicit knowledge was compared with action-outcome response mapping, goal sensitivity, or devaluation sensitivity, knowledge of standard pairs only was included to maintain consistency with other task metrics.

Data Analysis

All data analysis was conducted in R (Version 4.2.1; R Core Team 2022). Before conducting regression analyses, all task-related and symptom-related variables were examined to assess normality, skew, and kurtosis. Statistical assumptions for linear regressions, ANOVAs, and t-tests were inspected for each model. For regressions, linearity, normality of residuals, and independence of errors were assessed through visual examination of QQ and residual plots, and both outliers and high leverage of individual points were considered using Cook's distance with a criterion of $4/n - p - 1$ as suggested in Bruce & Bruce (2017). For independent samples t-tests, groups were tested separately for extreme outliers (>3 standard deviations from the mean), normality via the Shapiro Wilks test, and homogeneity of variance using Levene's test. In cases where groups showed unequal variances but other assumptions were satisfied, Welch's t-test was used. Bonferroni-adjusted *p* values are reported in tests of simple effects or pairwise comparisons following ANOVA tests. Mauchly's test was used to assess sphericity for ANOVAs, and the Greenhouse-Geiser correction was applied in cases where sphericity was violated. Friedman tests were used in place of repeated measures ANOVAs in the case of non-normally distributed variables. In cases not involving a comparison between stimulus discrimination types, all averaged task-related metrics refer to performance on standard pairs only. We used the R-packages *broom* [Version 1.0.5; Robinson, Hayes, & Couch (2023)], *car* [Version 3.1-2; Fox, Weisberg, & Price (2022)], *coin* [Version 1.4-2; Hothorn et al., 2006], *effsize* [Version 0.8.1; Torchiano 2020], *GGally* [Version 2.1.2; Schloerke et al 2021], *ggplot2* [Version 3.4.3; Wickham (2016)], *quantreg* [Version 5.94; Koenker 2022], *rstatix* [Version 0.7.2; Kassambra 2023], *sjPlot* [Version 2.8.11; Lüdecke (2022)], *viridis* [Version 0.6.1; Garnier et al. (2021a); Garnier et al. (2021b)], and *viridisLite* [Version 0.4.0; Garnier et al. (2021b)] for analyses and to produce figures in this paper.

Pre-registered Analyses

Four multiple regression models and two bivariate regressions were pre-registered prior to data collection (https://osf.io/82ypd). Key outcomes of interest included: 1) Individual response time (RT) differences between correct responses to devalued stimuli and correct responses to the same still-valued stimuli (in different blocks) during the slips-of-action test, operationalized as habit-driven conflict in goal-directed responses and 2) Devaluation sensitivity index (DSI), or the difference between number of responses to devalued stimuli and the number of responses to still-valued stimuli calculated per participant. Relatively higher DSI values indicate greater sensitivity to devaluation, and fewer slips of action relative to correct responses to still valuable stimuli. To

test the influences of OCD-related symptoms and response-outcome mapping on habit-driven response conflict, comparison of the following models was proposed:

$$DSI = b_0 + b_{1(R-O\ performance)} + b_{2(incompleteness)} + b_{3(harm-avoidance)} + b_{4(incompleteness\ x\ R-O\ performance)}$$

$$DSI = b_0 + b_{1(R-O\ performance)} + b_{2(DOCS\ score)}$$

Two additional multiple regression models using the response time conflict measure of habit as the outcome variable were also included with the same set of predictors. It was hypothesized that action-outcome response mapping performance and self-reported incompleteness would have significant influence on devaluation sensitivity, but no effect on habit-driven response conflict as measured with the response time difference metric. A significant positive relationship between incompleteness, but not harm-avoidance, and action-outcome response mapping performance on the instructed outcome devaluation test was hypothesized.

Sample Size Determination Based on A Priori Power Analysis. A power analysis conducted using G Power (Version 3.1.9.6, Faul et al., 2009) based on the multiple regression analyses determined that a sample size of 119 would be sufficient to detect a medium effect size of $f^2 = .15$ at alpha = .05 for 3 independent variables (action-outcome response mapping, harm-avoidance, and incompleteness) including one interaction term (action-outcome response mapping x incompleteness). Based on piloting and the previous experiment implementing a learning criterion, it was anticipated that about 75% of recruited participants would successfully meet the learning criterion and therefore 160 should be recruited. Due to the possibility of participant attrition between completion of the Fruit Task and the clinical interview session (which took place on a separate date), over 160 adults were invited to participate.

As described previously, a number of participants were excluded for the unexpected circumstance of fraudulent study completion; therefore, our sample size was smaller than planned. However, many studies implementing the Fruit Task have included sample sizes of 15-30 participants per diagnostic group when comparing performance between participants with and without OCD, and as small as 14 to examine correlations with individual neural measures and the same behavioral outcome measures used in this study. Therefore, our sample size of $n = 114$ to examine task effects and $n = 82$ to examine clinical relationships with task variables, though not statistically ideal, was comparably or better poised to detect effects relative to prior research using the Fruit task.

Preliminary Analyses

Learning. Average learning and response time over blocks of instrumental discrimination training were inspected separately by discrimination type. The median (and modal) number of trials to criterion was 7 blocks, or 168 trials (range = 120-288 trials). Distributions of response-outcome mapping accuracy on the instructed outcome devaluation test and explicit knowledge of stimulus-action and stimulus-outcome relationships were inspected separately for each stimulus discrimination type.

Habit-Driven Response Time Conflict. Individual averaged goal-directed response times were highly variable, and most averaged habit response times fell within one standard deviation of that participant's goal-directed response times (see Figure A6). A split-half reliability test of reaction time to still valuable stimuli (for which there were more data points, and theoretically should have been more reliable given automaticity of habit responses) showed overall low spearman correlations (average r_s = .11, *SD* = .30, range = -.89-.80). The low reliability of response times per participant during the slips-of-action test put into question the meaningfulness of the habit-driven response time conflict metric within this task design and sample. Response time conflict scores showed a nonsignificant relationship with DSI scores (r_s = .18, p = .15) and with goal sensitivity scores (r_s = .14, p = .27). These preliminary checks suggested poor reliability and validity of the habit-driven response time conflict, and therefore this index was not included in further analyses.

Devaluation Sensitivity and Goal Sensitivity. Distributions of slips-of-action performance variables were non-normally distributed and showed different patterns by stimulus discrimination type (DSI: $\chi^2(2) = 72.3$, $p < .0001$; Goal sensitivity: $\chi^2(2) = 81.6$, $p < .0001$, see Figure A4. Both devaluation and goal sensitivity were characterized by bimodal distributions, in which most participants' scores were normally distributed at lower values while a sizable minority of scores were grouped at the positive tail of the distribution with a smaller, but pronounced, local maximum.

DOCS scores. Due to experimenter error, two of the 20 DOCS items (both related to frequency and severity of unacceptable thoughts) were not included in the clinical symptom survey for the first 30 participants. Because the DOCS variable of interest was overall score rather than symptom subsets, I estimated a modeled DOCS sum from the data available using the following method: 1) Using complete DOCS data from the 52 participants who responded to all 20 DOCS items, I conducted a linear regression to estimate total DOCS score from the 18 items included on all surveys. Total scores summed from 18 items strongly predicted total scores including all 20 items ($F(1,50) = 4296$, $p < .0001$, beta = 1.11, $r^2 = .99$). Modeled total scores were then estimated for all 82 participants who completed the full study using the intercept and beta estimate derived from the linear regression, with estimated residual error terms for each participant sampled from a normal distribution with a mean of 0 and standard deviation of 1.

Core Motivational Dimensions of OCD. As expected, self-reported harm-avoidance and incompleteness were moderately correlated with each other (Pearson's r = .54, $p < .0001$) and both strongly correlated with overall OCD symptom severity as measured with the DOCS (harm-avoidance: r_s = .77, $p < .0001$; incompleteness: r_s = .64, $p < .0001$). Contrary to predictions, intensity scores of the Not-Just-Right-Experiences scale did not show a differentiable relationship with the core OCD dimensions; NJRE intensity scores accounted for an equivalent amount of variance in incompleteness (Pearson's r = .62, $p < .0001$, CI_{95} = .46-.74) and harm avoidance (r = .60, $p < .0001$, CI_{95} = .43-.73). Due to this lack of discriminant validity, it was decided to use DOCS symptom severity scores as the primary OCD-related measure in analyses.

Quantile Regression Analyses

The originally proposed pre-registered multiple regression analyses assumed that key variables of interest would be relatively normally distributed; however, multiple variables of interest were distributed bimodally or heavily skewed (e.g. action-outcome response mapping, explicit knowledge of actions and outcomes, devaluation sensitivity, goal sensitivity). Quantile regression is a logical alternative to ordinary least squares regression in the case of our bimodally

distributed and heavily skewed data because rather than estimating relationships based on the mean (which does not carry the same significance with bimodally distributed or skewed variables), it allows conditional estimation of the influence of predictor variables at specific quantiles of the outcome distribution (Wenz 2019, Konstantopoulos, Li, & van der Ploeg, 2019). Further, quantile regression does not require constant residual variance of the outcome variable and is robust to outliers. One important caveat is that coefficients of quantile regression cannot be used to infer population estimates, but it can still be useful in describing effects as observed within the sample (Konstantopoulos, Li, & van der Ploeg, 2019).

Because there were no a priori hypotheses about relationships between variables at specific quantiles, the entire distribution of key outcome variables were modeled on quantile intervals of 0.1. Reliability of model coefficients significant at the $p < .05$ level was assessed using a bootstrap approach as advised in Konstantopoulos, Li, & van der Ploeg (2019) to estimate standard error of beta estimates for quantile regression. This procedure computed standard error of model coefficients by taking the standard deviation of 1000 model estimates derived through simulations, each of which modeled the y ~ x relationship with random samples (with replacement) of size *n* from the data. The bootstrapped standard error was then used to compute 95% confidence intervals for the original coefficient estimate.

Exploratory Analyses

Exploratory analyses were performed to assess whether non-responses on devalued trials of the slips-of-action test may signify early initiation of goal-directed control unable to be fully carried out within the trial time limit. Non-responses during the slips-of-action test are of interest because previous task versions have instructed participants to withhold responses in response to stimuli with devalued outcomes, and researchers have interpreted lack of response to devalued stimuli to reflect goal-directed control. This reasoning is based on an assumption that habit-driven responses should occur quickly and relatively automatically, and lack of an automatic response should infer the converse of habitual control assuming a dual process theory of behavior, i.e. goal-directed control. Our task design required overt responses to signal goal-directed choice, but a closer examination of non-responses to devalued vs still-valuable trials and their associations with other measures of learning and goal-directed control could clarify whether non-responses likely signify meaningful goal-directed intention or random lapses in attention/memory failures.

Preliminary analyses compared anxiety, depression, and OCD-related scores for participants who did and did not meet diagnostic criteria for an OCD diagnosis on measures of anxiety and depression in addition to OCD-related measures (see Table 2).

Group-based analyses were performed to assess effects of OCD diagnostic status on measures of learning and habit vs goal-directed control. Bivariate correlations between symptom measures and learning/decision making variables were also examined to assess the specificity of relationships with compulsivity-related symptom dimensions (OCD, hoarding) relative to symptom dimensions not expected to correlate with goal-directed control (anxiety, depression).

Table 1

Symptom Measure Scores and Difference Statistics for Participants Grouped by OCD Diagnostic Status.

Measure	Score Range	OCD + *n* = 36		OCD - *n* = 46		*t*	*DF*	*p*	Cohen's *D*
		M	*SD*	*M*	*SD*				
DOCS	0-60	33	14	14	9.4	6.8	57.6	< .00001	1.5
TCDQ Incompleteness	10-50	35	7.4	28	6.7	4.7	80	< .0001	1.0
TCDQ Harm Avoidance	10-50	35	7.1	24	7.7	6.7	80	< .00001	1.5
NJRE: intensity	0-49	32	11	20	11	4.7	80	< .00001	1.1
Hoarding Rating Scale	0-40	11	8.9	9.8	8.4	.71	80	.48	-.16
DASS Depression	0-21	9.3	6.1	4.4	3.9	4.0	61.9	.0002	.91
DASS Anxiety	0-21	6.5	4.9	3.9	3.6	2.7	62.9	.01	.60

Note. OCD + and OCD - denote participants with and without a current diagnosis of OCD, respectively. DOCS = Dimensional Obsessive Compulsive Scale. TCDQ = Obsessive Compulsive Trait Core Dimensions questionnaire. DASS = Depression, Anxiety, and Stress Scale. NJRE: intensity = Not Just Right Experiences measure of intensity. DOCS, DASS Depression, and DASS Anxiety measures had unequal variances.

Results

Learning-Related Task Effects

Pre vs Post-Task Mood Ratings

Participants who successfully met the learning criterion did not differ from those who did not meet criterion on pre-task measures of low mood or intrusive distress. The median pre-task low mood rating in participants who successfully passed the learning criterion was 1 (range = 0-4) compared to 2 for participants who did not pass learning criterion; a Wilcoxon test showed that this difference was not significant ($p = .60$). The difference in intrusive distress ratings between those who did and did not meet the learning criterion was similarly non-significant ($p = .94$, median learners = 1; median non-learners = 2). On the other hand, there was a significant difference between pre- and post-task ratings of low mood among participants who successfully met the learning criterion. Post-task low mood ratings were significantly lower (indicating mood improvement) than pre-task low mood ratings (Wilcoxon paired signed rank $p = .002$, $r = .31$, Hodges-Lehmann estimate of Median difference = 1.0). Pre vs post intrusive distress ratings showed a similar trend of improvement with a mean difference of -.17, but this difference did not reach the significance threshold ($p = .07$).

Instrumental Discrimination

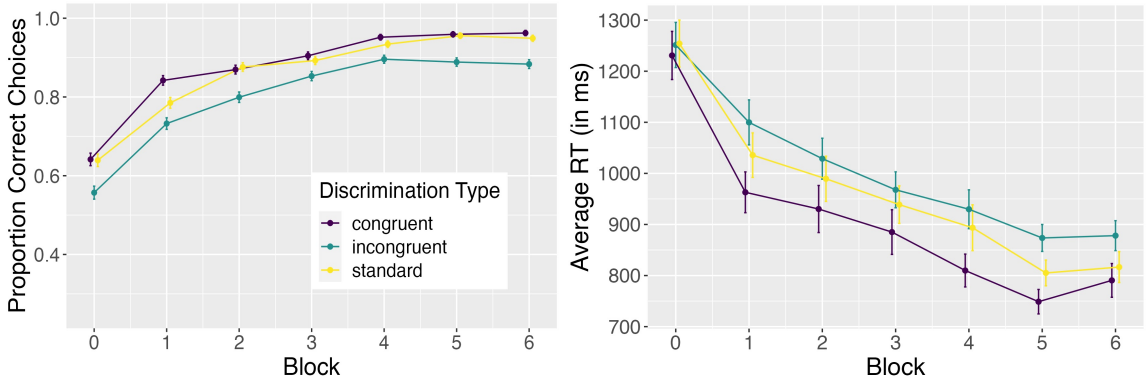
Visual inspection of the learning curve (averaged over participants) across blocks of instrumental discrimination showed that learning began to asymptote between blocks 4 and 5 (see Figure 2). To assess for differences in learning rate by discrimination types during initial learning, a two-way repeated measures ANOVA was computed with time (block) and discrimination type (congruent, incongruent, standard) as predictors of the sum of correct choices over blocks 1-3. Both block ($F(1.72, 194) = 177$, $p < .0001$, $/^2_G = .20$) and discrimination type ($F(1.9, 213$, $p < .0001$, $/^2_G = .04$) had significant effects on accuracy, with no significant block by discrimination type interaction. Post-hoc pairwise comparisons showed that participants had consistently lower accuracy for incongruent relative to congruent stimulus pairs during the first three blocks of learning, and performed better on standard relative to incongruent pairs in blocks one and three (see Table A3 for complete results). Parallel analyses were conducted to examine response time. Discrimination type ($F(1.9, 204.3 = 14.7$, $p < .0001$, $\eta^2_G = .01$), block ($F(1.4, 147 = 43.7$, $p < .0001$, $\eta^2_G = .09$), and their interaction ($F(3.3, 359 = 2.56$, $p = .049$, $\eta^2_G = .003$) all had significant effects on response time during the first three blocks of learning. Pairwise comparisons showed no difference in response times between discrimination types during the first block, and faster response times to congruent relative to both incongruent and standard discrimination pairs in the two subsequent blocks.

Instructed Outcome Devaluation

Action-outcome response mapping performance was non-normally distributed for each of the discrimination types (congruent, incongruent, standard), and each type showed distinct distributions. Accuracy on congruent trials, on which participants could rely on stimulus-response relationships, was skewed negative with median accuracy of 100%. Accuracy on standard and incongruent trials was bimodally distributed, with median accuracy of 75% for standard trials and 50% for incongruent trials. Notably, accuracy on incongruent trials showed a multimodal distribution pattern, with 31% of participants providing incorrect responses on all 8 incongruent trials, 18% of participants providing correct responses on all 8 trials, and 14% of participants performing at chance level with 4/8 correct.

Figure 2

Learning and Response Time Curves During Instrumental Discrimination Training.

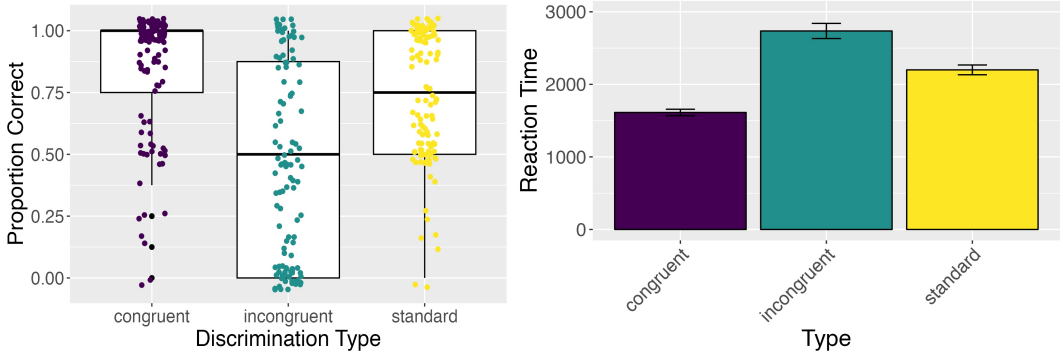


Note. Average accuracy (left) and response times in milliseconds (right) are shown over the first seven blocks of instrumental discrimination training for participants who successfully met the learning criterion. Response times over 5 seconds were excluded before averaging. Error bars denote standard error.

A repeated measures ANOVA was computed to assess effects of discrimination type (congruent, incongruent, standard) on response time during instructed outcome devaluation, considered as a proxy measure of trial difficulty. For this analysis, only trials with correct responses were included, response times over 5 seconds were excluded, and response time was log transformed, as raw values violated assumptions of normality. As expected, participants showed significant differences in response time by discrimination type ($F(2, 146) = 21.5$ $p < .0001$, $/^2_G = .07$), with post-hoc comparisons showing that participants took longer to respond correctly on incongruent relative to standard ($t(73) = 2.5$, $p = .04$) and congruent ($t(73) = 6.42$, $p < .0001$) trials, and took longer to respond correctly on standard relative to congruent ($t = 4.11$, $p < .001$) trials (see Figure 3).

Figure 3

Instructed Outcome Devaluation Accuracy Distributions and Response Time Averages by Discrimination Type.



Note. Reaction time (right) denotes response times on correct trials only. Response time was unlimited on instructed outcome devaluation trials; this figure does not include trials with response times longer than 5 seconds.

Explicit Knowledge

The majority (88.6%) of participants accurately identified all correct stimulus-response relationships while only 32.5% of participants accurately identified all six stimulus-outcome relationships. Table ~~A4~~ A3 shows frequencies of responses across levels of accuracy (0, 1, or 2 correct responses) by discrimination type. Distribution of explicit knowledge on stimulus-outcome relationships was nearly identical across congruent and incongruent stimulus pair types, with slightly increased accuracy for standard stimulus-outcome pairs (e.g., 54% of participants correctly identified 2/2 standard stimulus-outcome relationships relative to 48% and 45% for incongruent and congruent relationships, respectively).

Devaluation and Goal Sensitivity on the Slips-of-Action Test

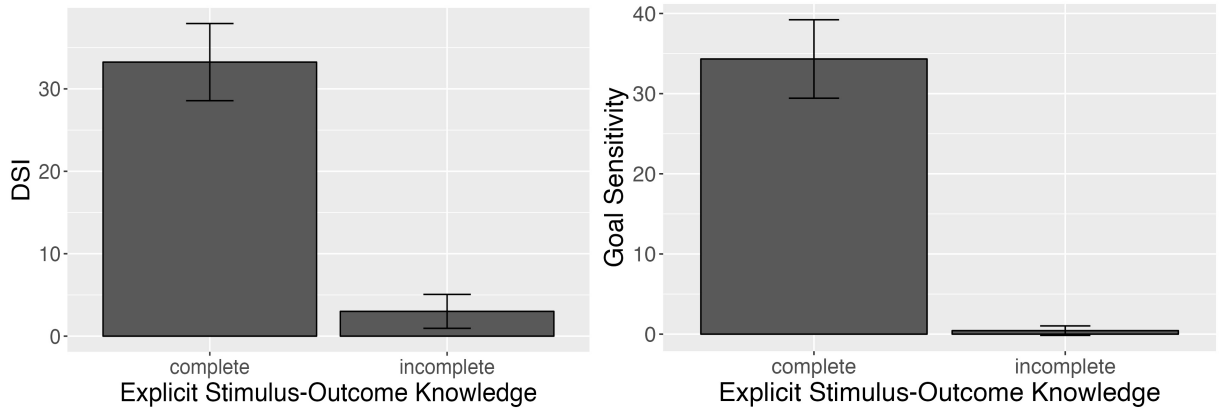
Non-Responses On Devalued Trials of the Slips-of-Action Test Do Not Correspond With Measures of Goal-Directed Control. In previous studies using the Fruit Task, goal-directed control was measured through covert (response withholding) rather than overt measures (e.g. pressing a key). Given the difficulty of this task, indicated by relatively low numbers of goal-directed responses and the presence of non-responses, I examined whether non-responses in our sample might have signaled an initiation of goal-directed cognition in which participants ran out of time to complete the action before presentation of the next stimulus. To explore this question, I first calculated individual differences in rates of non-response to devalued relative to still-valuable stimuli, to standardize non-responses in cases where goal-directed action was warranted against participant-specific base rates of non-responding. A linear model that regressed the total number of correct goal-directed actions against this difference in non-response variable showed that greater rates of non-response to devalued compared to still valuable stimuli did not predict overall numbers of correct goal-directed choices ($F(1,112) = 2.992$, $p = .087$). However, one might argue that an initiation of goal-directed cognition may not be reflected in overt choices during the slips-of-action test. Therefore, I also examined increases in non-response rates to devalued trials as a function of response-outcome mapping performance, which had a robust positive relationship with goal-directed control. Because of the bimodal distribution of action-outcome response mapping performance, a median split was used to group participants into high vs low performers. A t-test comparing non-response rate differences in participants with low vs high action-outcome response mapping performance showed a significant effect: participants with better action-outcome response mapping performance had relatively higher non-response rates to *still-valuable* as opposed to devalued trials ($t(112) = 2.44$, $p = .02$), the opposite relationship expected if non-responses were to signal initiation of goal-directed control n devalued trials. Taken together, these findings suggest that non-responses during the slips-of-action test did not indicate covert goal-directed control.

Both Implicit and Explicit Knowledge of Action-Outcome Relationships Influence Slips-of-Action Test Performance. A Wilcoxon rank sum test comparing devaluation sensitivity scores between participants who did and did not demonstrate complete explicit knowledge of stimulus-outcome relationships for standard discrimination pairs showed that explicit knowledge had a large effect on DSI ($W = 2213$, $p < .0001$, $r = .45$). Stimulus-outcome explicit knowledge

also had a significant and large effect on goal sensitivity (W = 2282, p < .0001, r = .50) as shown in Figure 4. Figure A5 includes the full distribution of DSI and goal sensitivity scores stratified by stimulus-outcome knowledge. Of note, 42% of participants with complete stimulus-outcome explicit knowledge had higher devaluation sensitivity scores than the highest scoring participant with incomplete knowledge of stimulus-outcome relationships.

Figure 4.

Slips-of-Action Test Performance Across Levels of Explicit Stimulus-Outcome Knowledge.



Note. DSI = devaluation sensitivity index.

OCD-Related Effects on Learning

Participants with and without OCD diagnoses did not differ on the number of instrumental discrimination blocks required to meet the learning criterion ($\chi^2(7) = 3.8$, $p = .80$), indicating no discernible difference in the speed of instrumental learning. A two-way mixed ANOVA was conducted to examine effects of group (OCD diagnosis vs no OCD diagnosis) and stimulus type (congruent, incongruent, standard) on action-outcome response mapping accuracy. This analysis showed a significant main effect of stimulus type ($F(2,160) = 31.53$, $p < .0001$, $/^2_G = .16$), but no significant main effect of diagnostic status group and no interaction effect. Table A3 shows post-hoc pairwise comparisons in accuracy by discrimination type. Bivariate correlations similarly showed no relationship between OCD symptom severity (DOCS) and action-outcome response mapping performance ($r_s = .06$, $p = .62$). With respect to explicit knowledge of stimulus-outcome relationships, participants with vs without OCD diagnoses did not differ on accuracy for standard discrimination pairs ($\chi^2(2) = .89$, $p = .64$).

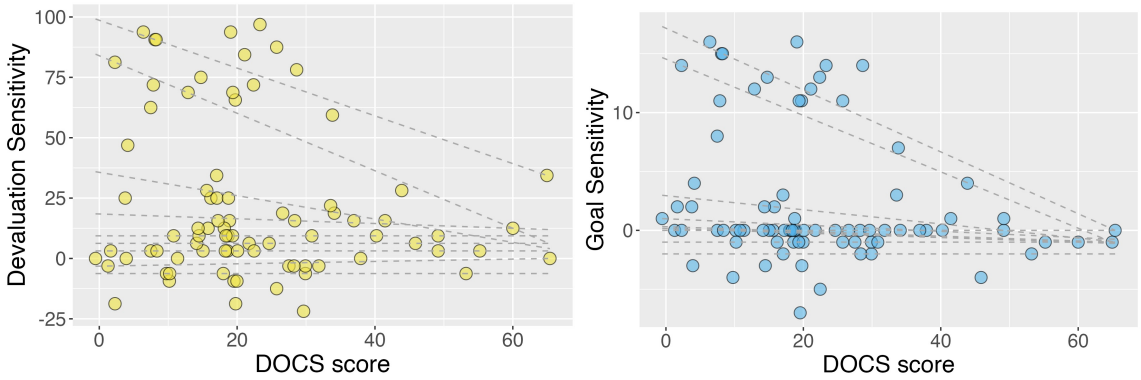
Bivariate correlations between OCD symptom severity and slips-of-action test performance showed that DOCS scores were significantly associated with goal sensitivity (r_s = -.24, $p = .03$) but not with devaluation sensitivity (r_s = -.08, $p = .48$).

Quantile regressions were performed to test the influence of OCD symptom severity at quantiles of devaluation sensitivity and goal sensitivity, ranging from 0.1 to 0.9 in intervals of 0.1. Increasing OCD symptom severity predicted a decrease in devaluation sensitivity at the highest quantiles of devaluation sensitivity ([beta](tau = 0.9) = -.99, $p = .02$, bootstrapped CI_{95} = -1.7 to -.32; [beta](tau = 0.8) = -1.2, $p = .007$, bootstrapped CI_{95} = -2.0 to -.36) and goal

sensitivity ([beta(tau = 0.9) = -.26, p = .0003, bootstrapped CI_{95} = -.36 to -.16; [beta](tau = 0.8) = -.24, p = .003, bootstrapped CI_{95} = -.40 to -.08. None of the beta estimates for devaluation sensitivity or goal sensitivity quantiles less than 0.8 were significantly different than zero, indicating no conditional relationship between OCD symptoms and slips-of-action test performance at medium and low levels of task performance (see Figure 5).

Figure 5

Quantile Regression Estimates of OCD Symptom Severity Effect on Habit vs Goal-Directed Control During the Slips-of-Action Test.



Note. Gray dotted lines denote quantile regression lines ranging from 0.1 to 0.9 of the outcome variable at intervals of 0.1. DOCS = Dimensional Obsessive Compulsive Scale.

Effects of OCD Diagnostic Status and Symptoms on Learning

A Wilcoxon rank sum test showed a significant difference in goal sensitivity by OCD diagnostic status (W = 1043, p = .04, r = .23, Hodges-Lehman estimate of median difference = 3.13) with decreased goal sensitivity in participants with OCD, but no significant difference in devaluation sensitivity between diagnostic status groups (W = 985, p = .14, r = .16, see Figure 6), mirroring correlational findings with the DOCS.

Discussion

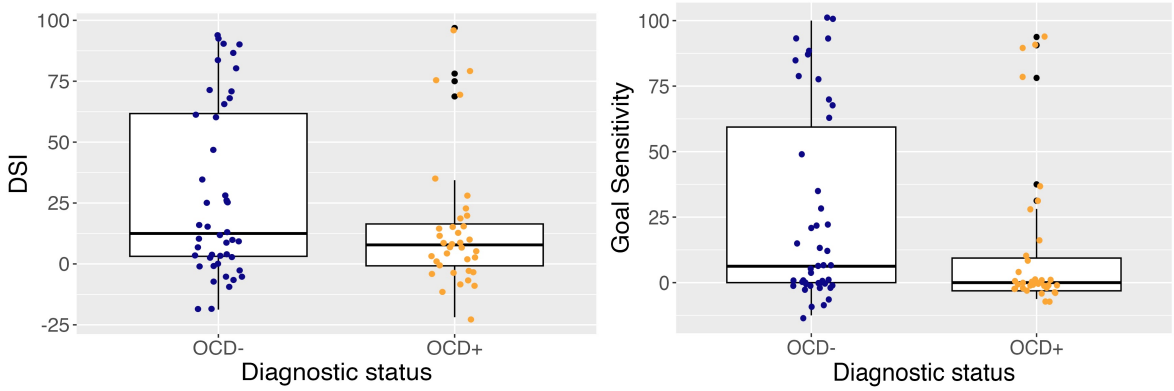
This study links two distinct bodies of research - clinical research on the dimensional structure of OCD, and research in the neurocognitive domain aiming to capture lab-based measures of habit and goal-directed control. This study aimed to clarify whether a commonly used metric of habit-driven failures in goal-directed control truly captures an overreliance on habit, or instead reflects other aspects of goal-directed control. The study sample included participants experiencing clinically significant symptoms of OCD, among whom we expected to see deficits in goal-directed control.

One of the goals of this study was to consider learning-related cognitive factors that could contribute to slips-of-action test performance. Multiple modifications to the task were implemented to better understand facets of learning that predicted DSI scores, many of which would not reflect habitual control. When controlling for initial stimulus-response learning by implementing a strict learning criterion for all participants, task effects emerged showing that

both explicit knowledge of stimulus-outcome relationships and action-outcome response mapping performance had strong influences on both key metrics of slips-of-action test performance. Participants with incomplete explicit knowledge of stimulus-outcome relationships and participants who had lower accuracy mapping responses to outcomes had a lower ceiling of devaluation sensitivity and goal sensitivity scores; in other words, it appeared that a lack of explicit and implicit knowledge of stimulus-action-outcome relationships limited the ability to execute goal-directed control during the slips-of-action test.

Figure 6

Goal Sensitivity But Not Devaluation Sensitivity Differed by OCD Diagnostic Status



Note. OCD + and OCD - denote participants with and without a current diagnosis of OCD, respectively.

The possibility that OCD dimensions of harm avoidance and incompleteness have a differential impact on learning-related outcomes or the balance of habitual vs goal-directed control remains an open question. Within our sample, these subscales were highly correlated with one another and with overall OCD symptom severity, and we were not able to establish discriminant validity due to both subscales correlating to an equivalent degree with a separate measure of self-reported incompleteness.

Turning to the links between task parameters and overall OCD symptom severity, we found a significant relationship between goal sensitivity on the slips-of-action test and OCD symptom severity, but this relationship was not specific to OCD symptoms, as depression and anxiety symptom severity also correlated with goal-sensitivity. Of the four symptom dimensions assessed, only depression showed a significant negative association with devaluation sensitivity. A quantile regression analysis showed that the negative association between OCD symptom severity and slips-of-action test performance was only evident at higher levels of overall performance on the slips-of-action test for both goal sensitivity and devaluation sensitivity measures. Finally, an assessment of group differences between participants with and without a current OCD diagnosis showed an effect of OCD diagnostic status on goal sensitivity, but not devaluation sensitivity, suggesting that OCD may be characterized by failures in goal-directed control unrelated to habit-driven interference.

Novel Design Approaches to the Measurement of Habitual vs Goal-Directed Control

This study implemented two novel approaches to discriminate habitual control from non-habit related failures in goal-directed control: examination of response time switch cost as a more precise measure of habit-driven response conflict, and use of an overt goal-directed choice option rather than a covert response inhibition metric. This study also implemented a learning criterion during instrumental discrimination to ensure that all participants reached close to 100% accuracy and sustained this accuracy over multiple learning blocks. Use of a learning criterion has not been reported in prior studies using the Fruit Task. The learning criterion guaranteed that a participant's degree of implicit or explicit knowledge of stimulus-response-outcome relationships, as well as the extent of habitual vs goal-directed control exerted on the slips-of-action-test, could not be explained by faulty stimulus-response learning.

Response Time Switch Cost Was Not Associated with Habit-Driven Conflict on the Fruit Task

Luque & colleagues (2020) proposed that response time switch cost may be a more reliable measure of habit than overt response selections during devaluation tests. However, we found average response times of goal-directed responses to be unreliable estimates due primarily to the low number of goal-directed responses per participant. Moreover, when we examined relationships between response time switch cost and other measures of habit vs goal-directed control, we found no significant relationships and were therefore unable to establish validity of this measure within our study. The response time switch cost measure had been previously suggested on the basis that humans are too easily able to override prepotent habitual responses, leading to a ceiling effect on choice accuracy and few cases of overt habit-driven responding to include in analyses. The Fruit Task, on the other hand, has nearly the opposite problem: The slips of action phase is so cognitively taxing that on average, participants make more erroneous habit-driven responses than correct goal-directed responses. Therefore, within this task, response time switch cost may not accurately reflect habit-driven conflict in goal-directed responding.

Average Learning Curves May Obscure Large Subsets of Non-Learners

Previous uses of the Fruit task have not, to my knowledge, implemented a learning criterion during the instrumental discrimination phase; these studies have relied instead on learning curves demonstrating improvement in accuracy over blocks when averaged over all participants (e.g. de Wit et al., 2012; Worbe et al., 2015; Bogdanov et al., 2018) or have not directly addressed accuracy during learning, only referring to lack of significant differences in learning rate between patient and control groups (e.g. Godier et al., 2016). Implementation of a learning criterion and subsequent examination of learning curves for those who did and did not meet this criterion revealed that averaged metrics of performance absconded discretely separate processes underlying performance across phases. Rather than exhibiting slower learning, participants who did not reach near perfect accuracy across stimulus types by the end of 300 trials were not, on average, learning much at all (see Figure A3). This figure shows that the learning curve of the full sample of task completers (n = 153) showed a typically shaped learning curve with average asymptotic performance around 85% correct for standard and congruent pairs and just under 80% for incongruent pairs by the end of 7 blocks of training, demonstrating similar or better accuracy than published studies employing this task (e.g. Gillan et al., 2011). This performance appears to indicate evidence of successful learning. However, an examination of participants within this group who did not pass our high threshold learning criterion (87% correct for each discrimination pair for at least 2 consecutive blocks with a maximum of 300 total trials)

revealed absence of any learning curve – a nearly flat line hovering around 60% representing a quarter of the sample, while three quarters of the sample showed the characteristic learning curve with a higher asymptote (around 95% accuracy for standard and congruent pairs, and just under 90% for incongruent pairs). In other words, the majority of the sample pulled the learning curve into its characteristic slope, while a quarter of the sample showed a discrete pattern of non-learning behavior. The possibility that this may have occurred in other studies was not addressed in any of the papers reviewed; all past studies reviewed validated learning by comparing averaged or group-level indicators of performance against chance levels of performance. Studies that validate task-based learning using average learning curves but derive participant-level learning or choice metrics for analyses – such as devaluation sensitivity – should be interpreted with caution, as a failure to learn initial stimulus-response relationships to the point of near automaticity negates the possibility of habit-driven choice interference and places strong limits on a participant's capacity to exert goal-directed control.

Participants Showed Non-Linear Patterns of Stimulus-Outcome Learning

Despite being prompted to pay attention to which actions led to fruit outcomes (as opposed to an empty box) in response to each stimulus, only about half of participants correctly identified outcomes for each discrimination pair on a multiple choice test with images of outcomes as choice options. Differences in accuracy by discrimination type by the end of the learning phase did not appear to translate into explicit knowledge of stimulus-outcome relationships, as there were no significant differences in the number of correct stimulus-outcome identifications by discrimination type on the multiple choice survey.

Action-outcome response mapping performance was distributed differently by discrimination type, and performance was non-normally distributed across discrimination types. Accuracy was bimodally distributed for standard pairs with most likely accuracy at 50% or 100%, and skewed negative for congruent pairs, with the bulk of the distribution near 100% accuracy. In the case of incongruent discrimination pairs, a clear multimodal distribution emerged: participants were most likely to demonstrate 0% or 100% response-mapping accuracy relative to values in between. The high number of participants responding incorrectly to all incongruent trials (chance performance would be 50%) suggested a flipped representation of action-outcome associations rather than a misunderstanding of task rules, forgetting of initial action-outcome relationships, or random response patterns. The poor performance on incongruent trials is consistent with outcome response theory and prior findings implementing this task design (de Wit et al., 2007). Inaccuracies in action-outcome response mapping even after reaching close to 100% accuracy during instrumental discrimination training are consistent with habit literature showing that as stimulus-response associations strengthen, response-outcome associations weaken (Bouton, 2021). However, this does not fully explain why only about half of our sample retained explicit knowledge of stimulus-outcome relationships even after being prompted to remember these relationships and being shown images of stimuli, actions, and outcomes during this test to aid recall.

Multiple Learning Processes Lead to the Capacity for Goal-Directed Control of Behavior

Strong relationships were observed between stimulus-outcome explicit knowledge and slips-of-action test performance, demonstrating the importance of maintaining an accurate representation of these associations in order to exert subsequent goal-directed control. Incomplete explicit knowledge of stimulus-outcome relationships before the slips-of-action test phase

appeared to have a nearly deterministic effect on the upper limit of slips-of-action performance, even though complete explicit knowledge did not guarantee better devaluation sensitivity or goal sensitivity. For example, participants with complete knowledge spanned the full range of devaluation sensitivity scores. However, *no* participants with incomplete knowledge had a devaluation sensitivity score greater than 34 (corresponding with the 77th percentile of DSI scores), where the maximum score was 100 and scores of zero indicated fully habitual control. It is unclear whether lower DSI scores among participants with incomplete explicit knowledge truly indicate lower devaluation sensitivity due to overriding habitual control, or whether other learning or memory mechanisms caused a failure to maintain representations of stimulus-outcome relationships. In this latter case, apparent low devaluation sensitivity would be driven by the impossibility of exerting goal-directed control relative to maintained stimulus-response relationships. It is worth noting that participants with complete explicit knowledge of stimulus-outcome relationships did not simply represent the upper tails of the DSI or goal sensitivity distributions; they appeared to comprise separate distributions of these variables with local maxima. It is possible, therefore, that these metrics of habit vs goal-directed control are capturing multiple processes at once, in part distinguished by features of initial learning.

The contribution of initial instrumental learning to the balance of habit vs goal-directed control is generally taken for granted once learners demonstrate high accuracy on stimulus-response tests, and emphasis is concentrated on choices and outcomes that occur *after* (over)learning. The strong relationships uncovered between behavioral and explicit awareness of stimulus-action-outcome associations and subsequent performance on a task requiring goal-directed control reveals that aspects of the instrumental learning phase – but not necessarily speed or accuracy of stimulus-response acquisition – play a key role determining subsequent goal-directed control capabilities. Of import, whereas poor representation of stimulus-outcome relationships appeared to preclude higher levels of goal-directed control, strong performance on action-outcome response mapping and explicit knowledge tests after learning did not *guarantee* increased goal-directed behavioral control. This suggests that multiple cognitive mechanisms contribute to the balance of habit vs goal-directed control, including, but not limited, to explicit representation and behavioral mapping of causal relationships acquired through instrumental learning.

An Attempt to Consider Specific Dimensions of OCD

The TCDQ measures of incompleteness and harm-avoidance were included in this study in an attempt to better specify the relationship between compulsivity and failures in goal-directed control. It was hypothesized that incompleteness, but not harm-avoidance, would relate to action-outcome response mapping deficits due to parallels between the experience of not-just-right experiences and action-outcome uncertainty. This question was not addressed through this study due to a lack of discriminant validity to support a distinction between incompleteness vs harm-avoidance. Further, neither incompleteness nor harm-avoidance were significantly related to any of the measures of learning or habitual vs. goal-directed control, in contrast to other symptom measures. One possibility for this lack of relationship is that the TCDQ as implemented was designed to measure incompleteness and harm-avoidance as traits that could apply to the general population, rather than only to adults diagnosed with OCD. Indeed, distributions of these variables were normal rather than skewed positive like all other symptom measures included in this study. Given that state-based symptom measures of depression, anxiety, and OCD - but

neither TCDQ dimension - correlated with goal sensitivity on the slips-of-action test, it is possible that in-the-moment symptom severity, rather than stylistic behavioral or cognitive patterns that may relate to psychopathology, is more influential in the exertion of goal-directed control.

OCD Diagnostic Status and Symptom Severity Related to Decreased Goal Sensitivity But Not Devaluation Sensitivity

In line with studies suggesting that OCD relates to deficits in goal-directed control (e.g., Gillan & Robbins, 2014; Seow et al., 2021, Sharp, Dolan, & Eldar, 2023), we found an association between OCD –whether measured by symptom severity or diagnostic status– and goal sensitivity, indicating impairments in goal-directed control. In contrast to prior accounts suggesting that OCD-related failures in goal-directed control are habit-driven, we did not find a relationship of OCD symptom severity or diagnostic status with devaluation sensitivity on the slips-of-action test. It is unclear why this is the case, but may reflect some of the additional controls we implemented in this experiment, such as using a learning criterion to ensure near-perfect accuracy for every participant during instrumental discrimination and the larger sample size relative to many of the studies showing this effect with the Fruit Task. It is also worth noting that many of the studies not employing the Fruit Task to motivate the habit hypothesis of OCD have considered model-free control on the 2-step reinforcement learning task first introduced by Daw & colleagues (2011) as a metric of habitual control. Though commonly conflated, habits and model-free control are not equivalent. Evidence supporting this notion includes that model-free choices on the 2-step task have been shown to decrease with extensive training on the 2-step task (Economides et al., 2015); the opposite would be expected of habit-driven choices. Further casting doubt on the interpretability of model-free control on the 2-step task, one study found that providing explanations for the occurrence of common vs rare transitions to new states led to participants making fully model-based decisions (Feher da Silva & Hare, 2020), suggesting that misunderstandings of task structure can lead to apparent model-free control (or hybrid model-free and model-based control) while in theory, decision control mechanisms should be independent from an understanding of task premises. These results are consistent with our finding that explicit knowledge of stimulus-outcome relationships limited participants' abilities to exert goal-directed control during the slips-of-action test. It is indeed possible that habits develop through model-free learning, in which actions are guided by the incremental value of immediate outcomes, but implementation of model-free control in a given context does not itself represent habitual behavior. Therefore, the extent to which participants employ model-free control during lab-based tasks does not represent a reliance on habitual over goal-directed behavior.

Unlike bivariate correlations, a quantile regression analysis showed a significant relationship between OCD symptom severity and devaluation sensitivity - but only at the highest quantiles of slips-of-action test performance. The same pattern was demonstrated when goal sensitivity was regressed against OCD symptom severity. The quantile regression results appear to mirror a pattern shown in Figure A5, in which both devaluation sensitivity and goal sensitivity distributions are characterized by bimodal distributions, distinguished in part by implicit and explicit learning performance. At a broad level, these results suggest that different cognitive/decision processes are occurring to explain variance at lower vs higher levels of performance. A logical question following this observation is whether a stronger bivariate relationship between slips-of-action performance measures and OCD symptom severity would

emerge among participants with complete stimulus-outcome explicit knowledge, or among participants with high action-outcome response mapping accuracy. Unfortunately, this study was underpowered to adequately investigate this possibility - significant bivariate correlations did emerge when subsetting the sample accordingly, however, confidence intervals were large and therefore correlation coefficients were not significantly stronger than those found in the original bivariate correlations of the full sample.

In contrast to past studies suggesting a specific relationship between failures in goal-directed control and compulsivity relative to other symptoms of psychopathology (e.g., Gillan et al., 2016), we found that levels of general anxiety and depression in addition to OCD symptom severity significantly correlated with multiple measures of goal-directed control. Of these, depression was the only symptom measure that also correlated with action-outcome response mapping. Further, hoarding symptoms did not correlate with any measures of learning or goal-directed control, despite being the symptom dimension (other than OCD symptoms) most closely linked to compulsivity. These relationships (and lack of relationships) contrast expectations based on the habit hypothesis of OCD, suggesting that compulsive behavior represents a general pattern of overreliance on the habitual system at the cost of successfully exercising goal-directed control, or that individuals with compulsive disorders (which would include both hoarding and OCD) show a bias toward habitual control of behavior.

Contrary to hypotheses, we did not find any relationship between action-outcome response mapping and OCD diagnostic status or symptom severity. This lack of relationship may have been due in part to problems with the measure itself: the performance distribution on the instructed outcome devaluation phase was multimodal and we were limited to non-parametric tests with few trials on which to base our indices. Given that the direction of the correlation between OCD symptom severity and action-outcome response mapping - though not significant - was in the expected direction, it is possible that a true effect exists and could be detected with a larger sample. Action-outcome response mapping measures may also show increased variability using probabilistic rather than deterministic task contingencies; different task designs may be better posed to examine this aspect of learning in OCD.

Limitations and Future Directions

We have shown here that the complexity of the Fruit Task, invoking nonlinear learning processes prior to outcome devaluation, muddles the interpretation of slips-of-action test performance as measured by devaluation sensitivity. However, this complexity may be a strength as well as a weakness, depending on the focus of investigation. Many lab-based experiments attempting to index habit-driven responses have encountered ceiling effects with overt behavioral responses, prompting creative solutions to establish covert metrics of habit-driven responding (e.g., Luque et al., 2020; Hardwick et al., 2019). The Fruit Task is difficult enough that overt behavioral responses show meaningful differences in performance. To increase the interpretability of these responses, future studies might incorporate *only* standard discrimination pairs and better prompt the maintenance of explicit knowledge during learning (e.g. with repeated questionnaires between blocks of instrumental discrimination). This method could also be useful to show whether explicit knowledge of stimulus-outcome relationships decreases with increased instrumental discrimination training, as successful performance relies less on maintained outcome representation. The Fruit Task was originally designed to highlight a feature of outcome response

theory; namely, that introducing incongruent discrimination pairs (in which the stimulus of one pair is the outcome of the other, and vice versa) should induce response conflict during the instructed outcome devaluation phase relative to congruent or standard discrimination pairs due to the stimulus-outcome-response representational structure (de Wit et al., 2007). Outcome response theory is not directly relevant to habit learning, and it is therefore unclear why the original design including congruent, incongruent, and standard discrimination pairs has not been modified from the original in the majority of studies conducted to assess habit vs goal-directed control using the slips-of-action test. Not only is the inclusion of congruent and incongruent discrimination pairs irrelevant to hypotheses regarding the slips-of-action test, it also interferes with the interpretation of how individuals balance habit and goal-directed control, given evidence showing distinct learning acquisition processes between discrimination types. While some published studies have explicitly stated that they include only standard pairs in slips-of-action test analyses (e.g., de Wit et al., 2012; de Wit et al., 2018), many have demonstrated inclusion of standard, incongruent, and congruent pairs but have not reported on differences in performance by discrimination type, nor have all explicitly included hypotheses about differential performance or how different distributions were handled in analyses.

Interpretation of devaluation sensitivity as indicating an individual's balance between habitual vs goal-directed control during the slips-of-action test relies on the dual-process model of decision making. Our results, showing that apparent habitual responses as well as goal-directed responses can be at least partially explained by implicit and explicit knowledge of task-related causal relationships, reveal problems with interpreting the dual process model too literally and without consideration of the hierarchy of goal-directed control. While a given situation may yield a single, narrowly defined habitual option, the possible actions under goal directed control are much broader and include levels of hierarchy. For example, consider the action of opening a social media app without thinking (habitual action) when one spots their smartphone at arm's length (stimulus) and experiencing a rewarding outcome of immediately viewing a photo of smiling friends. Eventually, performing the same action in response to a stimulus becomes less rewarding, as unwelcome advertisements inundate the social media feed, or feelings of shame accompany this action as one becomes aware of wasting time and perceives a loss of self-control with this automatic action. Again, the habitual action is clearly defined (opening the social media app) but goal-directed actions are less precisely defined, differ based on the context in which the stimulus arises (e.g., in a work meeting or at the waiting room of a doctor's office?) and may involve a re-consideration of the chosen policy (e.g. goal-directed control by reflection on the probability of possible outcomes, or goal-directed control by installing an app that prevents access to social media apps) adding layers of complexity to goal-directed control. Not all goal-directed choices are equally effective, and while it is relatively straightforward to compare degree of habitual control, goal-directed control is unlikely to scale on a single dimension. If indeed habitual control is less relevant to OCD than non-habit-related failures of goal-directed control, future studies may involve defining goal-directed control more precisely than the inverse of habitual control and clarifying which aspects (selection processes during learning? State transition uncertainty? Policy selection?) are most relevant to OCD-related compulsive behavior.

Conclusion

The two experiments discussed show key contributions of learning on goal-directed control, and call into question the interpretability of devaluation sensitivity on the slips-of-action

test, a popular measure of habit-driven failures in goal-directed control in human subject research. We were unable to address all OCD-related hypotheses due to questionable convergent and discriminant validity of a key measure meant to differentiate harm-avoidance and incompleteness, proposed to be core motivational dimensions driving compulsive behavior. Contrary to prior work finding specific associations of habit-driven failures in goal-directed control with compulsivity relative to other psychopathology dimensions, we found that decreased devaluation sensitivity was related to depression, but not OCD, symptom severity. Further, we found a nonspecific relationship of goal sensitivity with the severity of depressive, anxiety, and OCD symptoms, in that each symptom dimension related to poorer goal sensitivity on the slips-of-action test. An exploratory quantile regression analysis suggested that a relationship may exist between OCD symptom severity and devaluation sensitivity only at the highest ~20% of devaluation sensitivity; however, due to the sample size and exploratory nature of the quantile regression analysis, this finding should be interpreted with caution. To improve our understanding and identification of the specific processes influencing goal-directed control in humans, future studies might investigate neural or behavioral mechanisms that influence earlier processes between initial instrumental learning and the need to exert goal-directed control, including the ability to maintain a representation of stimulus-outcome and action-outcome relationships following instrumental learning.

References

Abramowitz, J. S. (2006). The psychological treatment of Obsessive—Compulsive disorder. *Can. J. Psychiatry*, *51*(7), 407–416.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*(2), 109–121.

Bogdanov, M., Timmermann, J. E., Gläscher, J., Hummel, F. C., & Schwabe, L. (2018). Causal role of the inferolateral prefrontal cortex in balancing goal-directed and habitual control of behavior. *Sci. Rep.*, *8*(1), 1–11.

Boschen, M. J., & Vuksanovic, D. (2007). Deteriorating memory confidence, responsibility perceptions and repeated checking: comparisons in OCD and control samples. *Behaviour Research and Therapy*, *45*(9), 2098–2109.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media, Inc.

Delorme, C., Salvador, A., Valabrègue, R., Roze, E., Palminteri, S., Vidailhet, M., … Worbe, Y. (2016). Enhanced habit formation in gilles de la tourette syndrome. *Brain*, *139*(Pt 2), 605–615.

Ecker, W., & Gönner, S. (2008). Incompleteness and harm avoidance in OCD symptom dimensions. *Behav. Res. Ther.*, *46*(8), 895–904.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., & Dolan, R. J. (2015). Model-Based Reasoning in Humans Becomes Automatic with Training. *PLoS Computational Biology*, *11*(9), e1004463.

Ersche, K. D., Gillan, C. M., Jones, P. S., Williams, G. B., Ward, L. H. E., Luijten, M., … Robbins, T. W. (2016). Carrots and sticks fail to change behavior in cocaine addiction. *Science*, *352*(6292), 1468–1471.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Foa, E. B., Abramowitz, J. S., Franklin, M. E., & Kozak, M. J. (1999). Feared consequences, fixity of belief, and treatment outcome in patients with obsessive-compulsive disorder. *Behav. Ther.*, *30*(4), 717–724.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive inventory: Development and validation of a short version. *Psychol. Assess.*, *14*(4), 485–496.

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053–1066.

Fox J, Weisberg S, Price B (2022). _carData: Companion to Applied Regression Data Sets_. R package version 3.0-5, <https://CRAN.R-project.org/package=carData>.

Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P., & Huppert, J. D. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychol. Rev.*, *127*(5), 672–699.

Garnier, Simon, Ross, Noam, Rudis, Robert, … Cédric. (2021a). *viridis - colorblind-friendly color maps for r*. https://doi.org/10.5281/zenodo.4679424

Garnier, Simon, Ross, Noam, Rudis, Robert, … Cédric. (2021b). *viridis - colorblind-friendly color maps for r*. https://doi.org/10.5281/zenodo.4679424

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & Wit, S. de. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am. J. Psychiatry*, *168*(7), 718–726.

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological Psychiatry*, *75*(8), 631–638.

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive-compulsive disorder. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1655). https://doi.org/10.1098/rstb.2013.0475

Gillan, C. M., Apergis-Schoute, A. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2015). Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. *The American Journal of Psychiatry*, *172*(3), 284–293.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*. https://doi.org/10.7554/eLife.11305

Gillan, C. M. (2021). Recent developments in the habit hypothesis of OCD and compulsive disorders. In *Current topics in behavioral neurosciences* (pp. 1–21). Berlin, Heidelberg: Springer Berlin Heidelberg.

Godier, L. R., de Wit, S., Pinto, A., Steinglass, J. E., Greene, A. L., Scaife, J., Gillan, C. M., Walsh, B. T., Simpson, H.-B., & Park, R. J. (2016). An investigation of habit learning in Anorexia Nervosa. *Psychiatry Research*, *244*, 214–222.

Graybiel, A. M., & Rauch, S. L. (2000). Toward a neurobiology of obsessive-compulsive disorder. *Neuron*, *28*(2), 343–347.

Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time-dependent competition between goal-directed and habitual response preparation. *Nature Human Behaviour*, *3*(12), 1252–1262.

Hauser, T. U., Moutoussis, M., Dayan, P., & Dolan, R. J. (2017). Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. *Transl. Psychiatry*, *7*(12), 1–10.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego system for conditional inference." _The American Statistician_, *60*(3), 257-263. doi:10.1198/000313006X118430

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a class of permutation tests: The coin package." _Journal of Statistical Software_, *28*(8), 1-23. doi:10.18637/jss.v028.i08

Kassambara A (2023). _rstatix: Pipe-Friendly Framework for Basic Statistical Tests_. R package version 0.7.2, <https://CRAN.R-project.org/package=rstatix>.

Koenker R (2022). _quantreg: Quantile Regression_. R package version 5.94, <https://CRAN.R-project.org/package=quantreg>.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods*, *47*(1), 1–12.

Lindeløv, J. K. (2011). Calculating d', beta, c and ad' in python and PHP. *Neuroscience, Stats, and Coding*. Retrieved from https://lindeloev.net/calculating-d-in-python-and-php/
Lüdecke D (2022). _sjPlot: Data Visualization for Statistics in Social Science_. R package version 2.8.11, <https://CRAN.R-project.org/package=sjPlot>.
Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. E. (2020). Measuring habit formation through goal-directed response switching. *J. Exp. Psychol. Gen.*, *149*(8), 1449–1459.
Mathews, C. A., Delucchi, K., Cath, D. C., Willemsen, G., & Boomsma, D. I. (2014). Partitioning the etiology of hoarding and obsessive-compulsive symptoms. *Psychological Medicine*, *44*(13), 2867–2876.
Pietrefesa, A. S., & Coles, M. E. (2008). Moving beyond an exclusive focus on harm avoidance in obsessive compulsive disorder: Considering the role of incompleteness. *Behav. Ther.*, *39*(3), 224–231.
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rachman, S. (1997). A cognitive theory of obsessions. *Behav. Res. Ther.*, *35*(9), 793–802.
Rachman, S., Rachman, S. J., & Hodgson, R. J. (1980). *Obsessions and compulsions*. Prentice-Hall. Retrieved from https://books.google.com/books?id=z2seAQAAIAAJ
Radomsky, A. S., & Alcolado, G. M. (2010). Don't even think about checking: mental checking causes memory distrust. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(4), 345–351.
Radomsky, A. S., Dugas, M. J., Alcolado, G. M., & Lavoie, S. L. (2014). When more is less: doubt, repetition, memory, metamemory, and compulsive checking in OCD. *Behaviour Research and Therapy*, *59*, 30–39.
Rasmussen, S. A., & Eisen, J. L. (1992). The epidemiology and clinical features of obsessive compulsive disorder. *Psychiatr. Clin. North Am.*, *15*(4), 743–758.
Robinson D, Hayes A, Couch S (2023). _broom: Convert Statistical Objects into Tidy Tibbles_. R package version 1.0.5, <https://CRAN.R-project.org/package=broom>
Sheehan, D. V. (2014). The mini-international neuropsychiatric interview, version 7.0 for DSM-5 (MINI 7.0). *Jacksonville, FL: Medical Outcomes Systems*.
Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J (2021). _GGally: Extension to 'ggplot2'_. R package version 2.1.2, <https://CRAN.R-project.org/package=GGally>.
Sharp, P. B., Dolan, R. J., & Eldar, E. (2023). Disrupted state transition learning as a computational marker of compulsivity. *Psychological Medicine*, *53*(5), 2095–2105.
Sjoerds, Z., Dietrich, A., Deserno, L., de Wit, S., Villringer, A., Heinze, H.-J., Schlagenhauf, F., & Horstmann, A. (2016). Slips of Action and Sequential Decisions: A Cross-Validation Study of Tasks Assessing Habitual and Goal-Directed Action Control. *Frontiers in Behavioral Neuroscience*, *10*, 234.
Sibrava, N. J., Boisseau, C. L., Eisen, J. L., Mancebo, M. C., & Rasmussen, S. A. (2016). An empirical investigation of incompleteness in a large clinical sample of obsessive compulsive disorder. *J. Anxiety Disord.*, *42*, 45–51.
Snorrason, I., Lee, H. J., Wit, S. de, & Woods, D. W. (2016). Are nonclinical obsessive-compulsive symptoms associated with bias toward habits? *Psychiatry Res.*, *241*, 221–223.

Summerfeldt, L. J., Kloosterman, P. H., Parker, J. D. A., Antony, M. M., & Swinson, R. P. (2001). Assessing and Validating the Obsessive-Compulsive-Related Construct of Incompleteness. Poster presented at the 62nd Annual Convention of the Canadian Psychological Association, Ste-Foy, Quebec.

Summerfeldt, L. J. (2004). Understanding and treating incompleteness in obsessive-compulsive disorder. *J. Clin. Psychol.*, *60*(11), 1155–1168.

Summerfeldt, L. J., Kloosterman, P. H., Antony, M. M., & Swinson, R. P. (2014). Examining an obsessive-compulsive core dimensions model: Structural validity of harm avoidance and incompleteness. *J. Obsessive Compuls. Relat. Disord.*, *3*(2), 83–94.

Tolin, D.F., Frost, R.O., & Steketee, G. (2010). A brief interview for assessing compulsive hoarding: The Hoarding Rating Scale-Interview. Psychiatry Research, 178, 147-152.

Torchiano M (2020). _effsize: Efficient Effect Size Computation_. doi:10.5281/zenodo.1480624 <https://doi.org/10.5281/zenodo.1480624>, R package version 0.8.1, <https://CRAN.R-project.org/package=effsize>.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.*, *29*(11), 2225–2232.

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity reveals a novel dissociation between action and confidence. *Neuron*, *96*(2), 348–354.e4.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., … Bullmore, E. T. (2014). Disorders of compulsivity: A common bias towards learning habits. *Mol. Psychiatry*, *20*(3), 345–352.

Watson, P., & Wit, S. de. (2018). Current limits of experimental research into habits and future directions. *Current Opinion in Behavioral Sciences*, *20*, 33–39.

Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L., & Turner, K. (2022). Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neuroscience and Biobehavioral Reviews*, *142*, 104869.

Wenz, S. E. (2019). What Quantile Regression Does and Doesn't Do: A Commentary on Petscher and Logan (2014) [Review of *What Quantile Regression Does and Doesn't Do: A Commentary on Petscher and Logan (2014)*]. *Child Development*, *90*(4), 1442–1452.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot

de Wit, S., Barker, R. A., Dickinson, A. D., & Cools, R. (2011). Habitual versus goal-directed action control in parkinson disease. *J. Cogn. Neurosci.*, *23*(5), 1218–1229.

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., … Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *J. Exp. Psychol. Gen.*, *147*(7), 1043–1065.

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *J. Exp. Psychol. Anim. Behav. Process.*, *33*(1), 1–11.

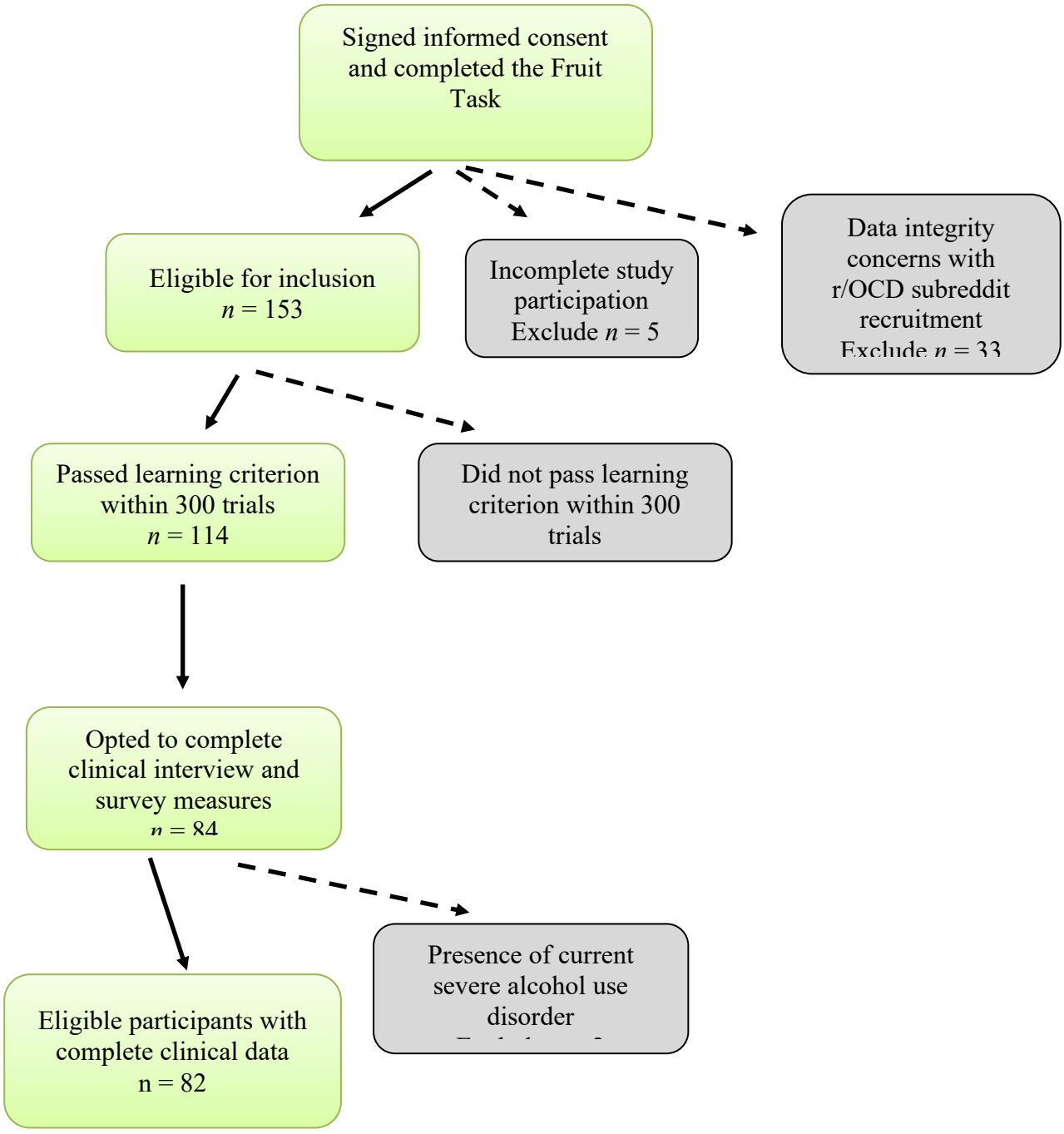
de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., Vijver, I. van de, & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *J. Neurosci.*, *32*(35), 12066–12075.

Wootton, B. M., Diefenbach, G. J., Bragdon, L. B., Steketee, G., Frost, R. O., & Tolin, D. F. (2015). A contemporary psychometric evaluation of the obsessive compulsive Inventory—Revised (OCI-R). *Psychol. Assess.*, *27*(3), 874–882.

Appendix A – Supplementary Tables and Figures

Figure 1

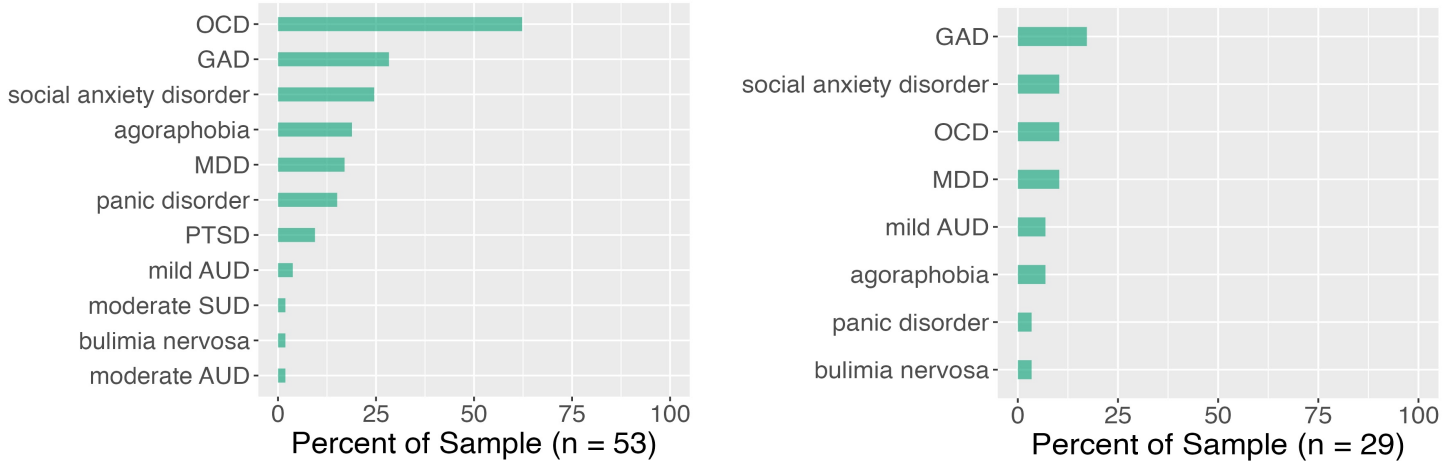
Participant Inclusion Flowchart



Note. Dotted lines leading to gray boxes indicate points of participant exclusion. Rows from top to bottom indicate ascending chronology.

Figure 2

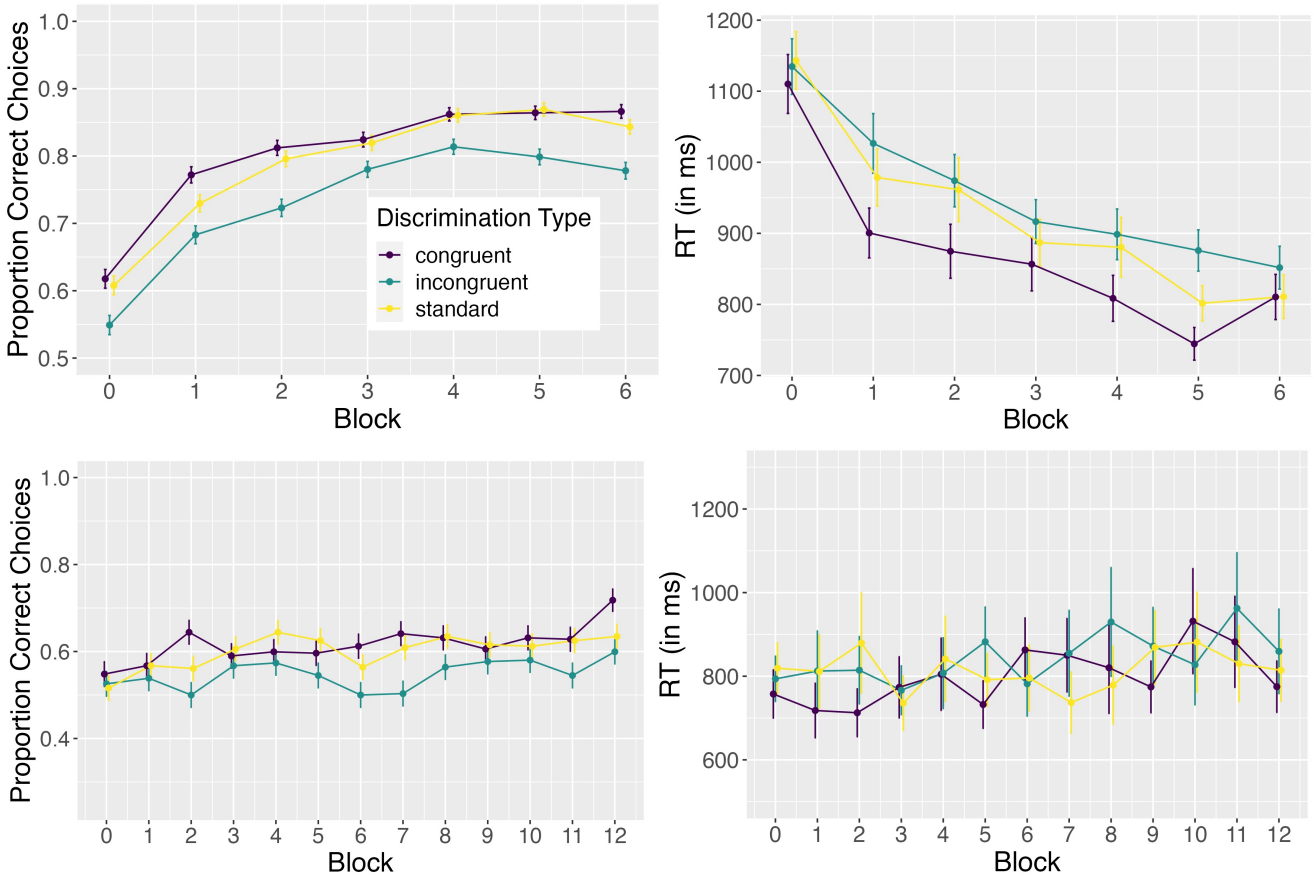
Clinical Characteristics of Participants Recruited from the Community (Left) and Undergraduate Research Participation Programs (Right)



Note. Diagnoses as assessed with the Mini International Neuropsychiatric Interview are listed on the left of each figure and ordered by most to least common within the respective sample. Diagnoses are not shown if no participants within that sample met criteria. OCD = obsessive compulsive disorder, GAD = generalized anxiety disorder, MDD = major depressive disorder, PTSD = posttraumatic stress disorder, AUD = alcohol use disorder, SUD = substance use disorder.

Figure 3

Average Learning and Response Time Curves During Instrumental Discrimination Conceal a Subset of Non-Learning Participants



Note. Top row shows average accuracy (top left) and response time (top right) curves for all participants who completed the fruit task (n = 153). Bottom row shows the subset (*n* = 39) of participants who did not meet the learning criterion within 300 instrumental discrimination trials. Error bars show standard error.

42

Figure 4

Devaluation Sensitivity and Goal Sensitivity Distributions Differed by Discrimination Type

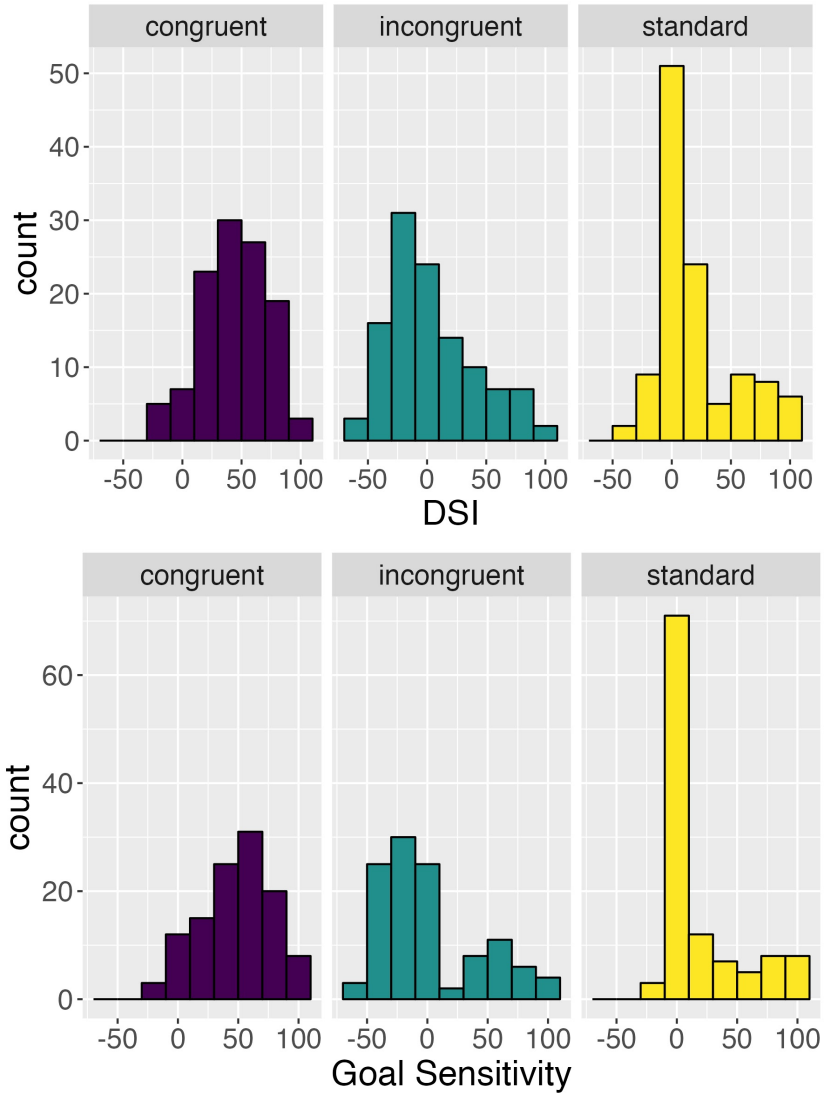
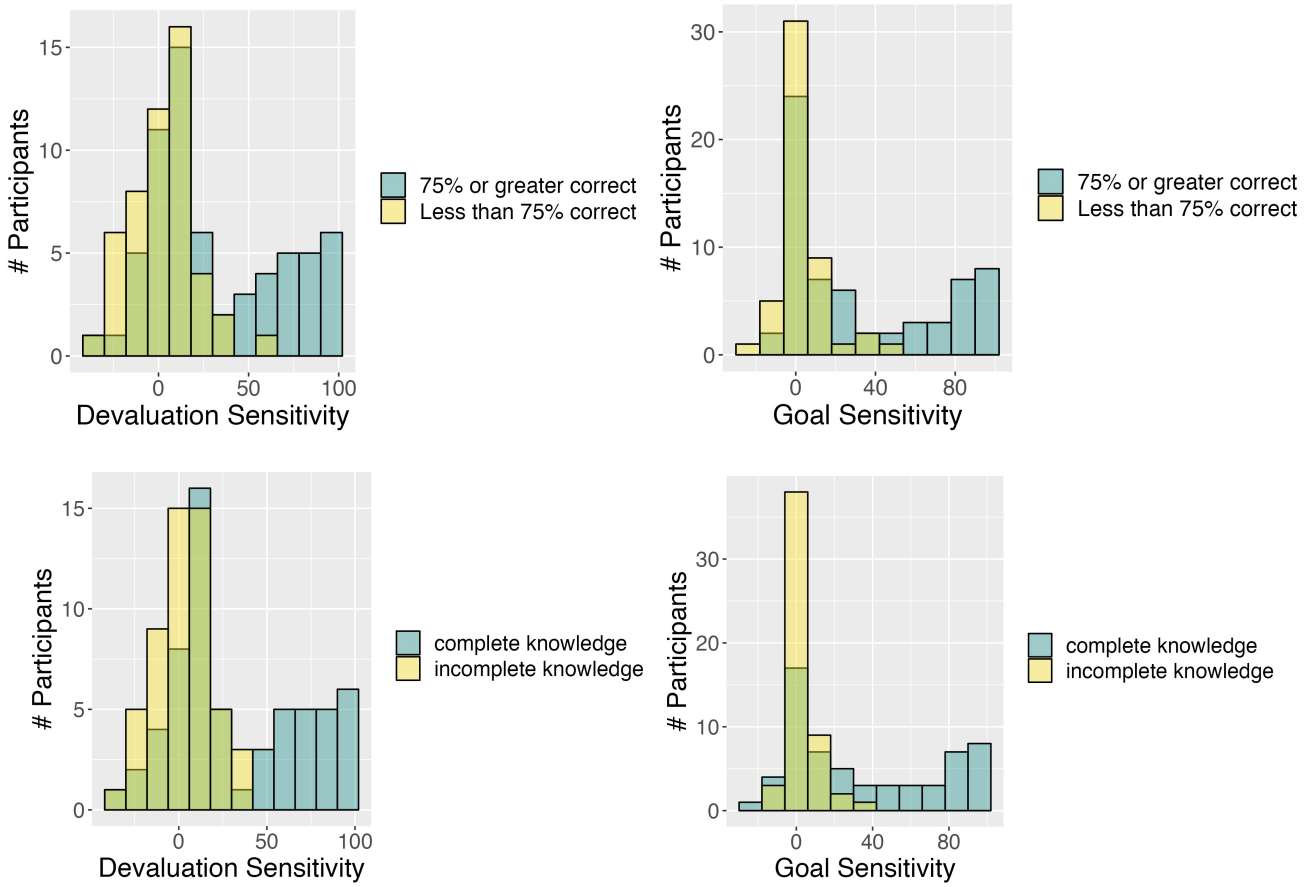


Figure 5

Slips-of-Action Test Performance Stratified by Individual Measures of Stimulus-Outcome Representation



Note. Action-outcome response mapping performance during instructed outcome devaluation (top row) is grouped based on a median split (Median accuracy = 75% correct). Bottom row shows participants grouped by complete vs incomplete explicit knowledge of stimulus-outcome relationships based on a multiple choice test.

Table 1

Sum of Correct Responses During Initial Learning by Discrimination Type

	Discrimination Type								
Block	Congruent		Incongruent		Standard		Comparison	*t*(113)	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*			
	5.13	1.64	4.46	1.70	5.10	1.62	Congruent > Incongruent	3.15	.006
1							Congruent = Standard	0.18	.86
							Standard > Incongruent	3.53	.002
2	6.74	1.49	5.85	1.61	6.27	1.71	Congruent > Incongruent	5.49	< .0001
							Congruent > Standard	2.64	.028
							Standard = Incongruent	2.39	.055
3	6.94	1.52	6.39	1.70	6.99	1.34	Congruent > Incongruent	3.00	.01
							Congruent = Standard	0.72	1.0
							Standard > Incongruent	4.19	.0002

Note. Means describe the average sum of correct responses within each block. Blocks included 8 trials per discrimination type.

Table 2

Bivariate Correlations Between Task-Related Variables and Symptom Measures

	1	2	3	4	5	6	7	8	9
1. Action-outcome response mapping									
2. Devaluation sensitivity	.44***								
3. Goal sensitivity	.41***	.79***							
4. TCDQ Incompleteness	-.06	-.07	-.18						
5. TCDQ harm-avoidance	.06	-.05	-.15	.54***					
6. NJRE Intensity	-.09	-.11	-.22	.62***	.60***				
7. DOCS	.06	-.08	-.24*	.64***	.77***	.68***			
8. Hoarding Rating Scale	.04	.08	-.03	.28*	.42***	.36**	.38***		
9. DASS Depression	-.01	-.25*	-.37***	.52***	.68***	.49***	.60***	.44***	
10. DASS Anxiety	.02	-.12	-.26*	.56***	.56***	.47***	.59***	.35**	.67***

Note. * = p < .05, ** = *p* < .01, *** = *p* < .001. TCDQ = Trait Core Dimensions Questionnaire; NJRE = Not-Just-Right Experiences, DOCS = Dimensional Obsessive Compulsive Scale; DASS = Depression, Anxiety, and Stress Scale. NJRE intensity was only reported by 77 participants who endorsed at least one not-just-right experience in the past month. All other variables include responses from all 82 participants who completed clinical measures. Pearson correlation coefficients are reported for relationships between NJRE intensity, TCDQ incompleteness, and TCDQ harm avoidance; Spearman's correlation coefficients are reported for relationships involving all other variables due to non-normal distributions.

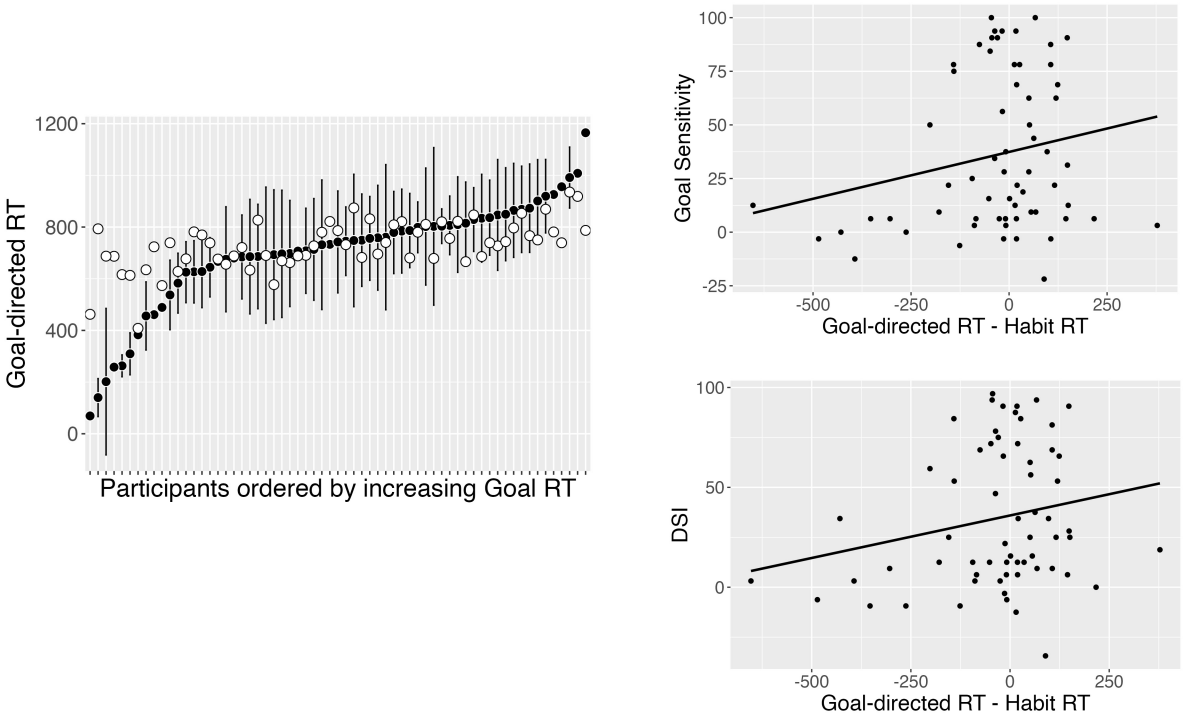
Table 3

Explicit Knowledge of Stimulus-Outcome Relationships

	0 of 2 Correct	1 of 2 Correct	2 of 2 Correct
Discrimination Type	% of sample (n = 114)		
Congruent	37	18	45
Incongruent	35	17	48
Standard	26	20	54

Figure 6

Individual Habit vs Goal-Directed Response Did Not Demonstrate Reliability or Validity as a Measure of Habit-Driven Response Conflict



Note. Left: Black filled dots show average goal-directed response times per participant, ordered by increasing response time. White filled dots show average habit response times. Error bars denote standard deviations of goal-directed response times (black dots with no error bars indicate participants with a single goal-directed choice). Right: Response time difference scores (in milliseconds) did not have significant relationships with either goal sensitivity (top) or devaluation sensitivity (bottom).

Appendix B. Experiment 1 Supplementary Method.

Participants

Participants were recruited using Amazon's Mechanical Turk (MTurk). Participants completed a battery of self-report measures of psychopathology before completing the Fruit Task via an external link to a Heroku-based web-application. Of the 187 participants who completed the self-report measures, 135 also completed the Fruit Task. The remaining 52 did not meet the pre-designated learning criterion in the training phase of the Fruit Task within 300 trials. Age and gender demographics of participants are shown in Table 1.

Table 1

Age and gender demographics for Fruit Task completers

Gender	n	Mean Age	*SD* Age	Education (years)	*SD* Education
female	71	38	12	15	2
male	58	39	11	15	2
other	1	29	NA	16	NA

Note: 5 participants did not report age, gender, or education information.

Procedure

After completing other study-related surveys on mTurk, participants were provided an external link to complete the Fruit Task, which was hosted as a web-based application on Heroku. Choice data was collected without personal identifying information and sent to a password-protected MongoDB Atlas cloud database. A random 10-digit number was produced at completion of the experiment for each participant, and they manually entered this number in MTurk to match task data with their Turker ID and receive payment.

The Fruit Task

The Fruit Task was programmed in JavaScript with use of Leeuw (2015)'s JsPsych library according to the description in Gillan & colleagues (2011). After local testing, the task was set up as a Node.js web application and hosted on Heroku.

Instrumental discrimination training. In this phase, participants learned by trial-and-error which key to press (right-hand key, "m" or left-hand key, "c") in response to a stimulus (a single fruit on the outside of a box) in order to reveal a rewarding outcome (a single fruit on the inside of the box). Participants completed 4 blocks of 24 trials. They then completed additional blocks until learning criterion was reached, pre-defined as at least 2 additional blocks with >87% accuracy. In other words, participants had to respond correctly on at least 21 of 24 trials in these blocks. Participants who did not reach criterion within 300 trials were disqualified from completing the rest of the task. Participants were incentivized to respond quickly during this phase by a point system that rewarded faster responses with more points: 5 points for correct responses within 0-1 second; 4 points for 1-1.5 seconds; 3 points for 1.5-2 seconds; 2 points for 2-2.5 seconds; and 1

point for > 2.5 seconds. All participants were explicitly instructed to pay attention to the outcomes of their responses and advised that they would be quizzed on which fruits on the outside of boxes led to which fruits on the insides of boxes.

Instructed outcome devaluation test. Following the training phase, participants completed a devaluation phase in which they were presented with two fruit outcomes at once, one of which was marked with a red "X" and termed "spoiled." These outcomes were grouped according to discrimination type (standard, incongruent, or congruent). Participants were instructed to press the key that led to the still-valued outcome. They were encouraged to take their time and do their best to respond accurately. The two outcomes were oriented vertically to avoid potential spatial-motor confounds with the left vs right-hand "c" and "m" keys, and the position of each outcome was counterbalanced across trials. Each individual stimulus was devalued on 4 trials for a total of 24 trials. The order of presentation was randomized. No feedback was provided during this phase.

Explicit action and outcome knowledge survey. After the instructed outcome devaluation phase, participants completed a survey in which they were shown an image of each stimulus and asked to indicate a) the correct key response ("m" or "c") to that stimulus and b) the outcome associated with that stimulus (all outcome images were shown and participants had to select only one per stimulus).

Slips-of-Action test. The final phase of the Fruit Task involved 6 test blocks, each with 24 trials. Each block began with a 15-second screen that showed all 6 fruit outcomes in a 3x2 grid (see Figure). Two of these outcomes were marked with a red "X" (as in the instructed outcome devaluation phase). Participants were instructed that the fruits inside the boxes marked with an "X" were spoiled, and that in this phase, they should respond to stimuli leading to still-valuable fruit outcomes with the correct key but should withhold responding to fruit stimuli that led to spoiled fruit outcomes. After this 15-second screen, stimuli appeared in quick succession, one at a time, as in the training phase. Participants were instructed that they could earn points by responding quickly to still-valuable fruit stimuli but would lose points if they responded to stimuli that were linked to spoiled fruit outcomes. They could avoid losing points by withholding responses to fruits leading to devalued outcomes. No feedback was delivered during the slips-of-action test. Each stimulus was shown for 1 second, followed by a jittered intertrial interval lasting 1-2 seconds, and then replaced with a new stimulus.

Pseudo-randomization for the Slips of Action Test

The combinations of devalued outcomes during each block of the slips-of-action test were pseudo-randomized for the following reason: If the 2 devalued outcomes per block were chosen randomly, certain blocks might end up easier or more difficult than others; for example, if both devalued outcomes came from the congruent pair, or if both devalued outcomes involved the same key response. In addition to creating within-subject confounds between blocks, this could also lead to differences in the difficulty of the slips-of-action phase between participants, removing our ability to reasonably compare performance across our sample. In light of these concerns, we applied pseudo-randomization as follows: The two devalued outcomes per block were selected in pairs so that each of the two devalued outcomes was associated with a distinct key press, and so that the two devalued outcomes represented one of the following stimulus discrimination combinations: 1) congruent & standard, 2) incongruent & congruent, or 3) incongruent & standard. Each participant thus completed 2 blocks of each stimulus discrimination combination during the 6-block Slips-of-Action phase. Though the six

combinations of devalued outcomes were pre-selected for each participant at the start of the task, they were assigned randomly to each block, so that their order of appearance was randomized.

Table 2

Key modifications to the original Fruit Task used in Gillan et al. (2011)

Task Phase	Gillan & colleagues (2011) version	Our version for Experiment 1
Instrumental Discrimination Training	72 trials	Minimum of 168 trials (96 trials + 3 blocks x 24 trials with correct > 90%)
Instructed Outcome Devaluation	12 trials (4 trials of each discrimination pair; outcome was devalued twice)	24 trials (8 trials of each discrimination pair; each outcome was devalued 4 times)
	10-second screen at the start of each test block showing 2 devalued outcomes	15-second screen at the start of each test block showing 2 devalued outcomes
Slips of Action	1 second presentation, 1 second intertrial interval	1 second presentation, ~1.5 seconds jittered intertrial interval

Signal Detection Equations

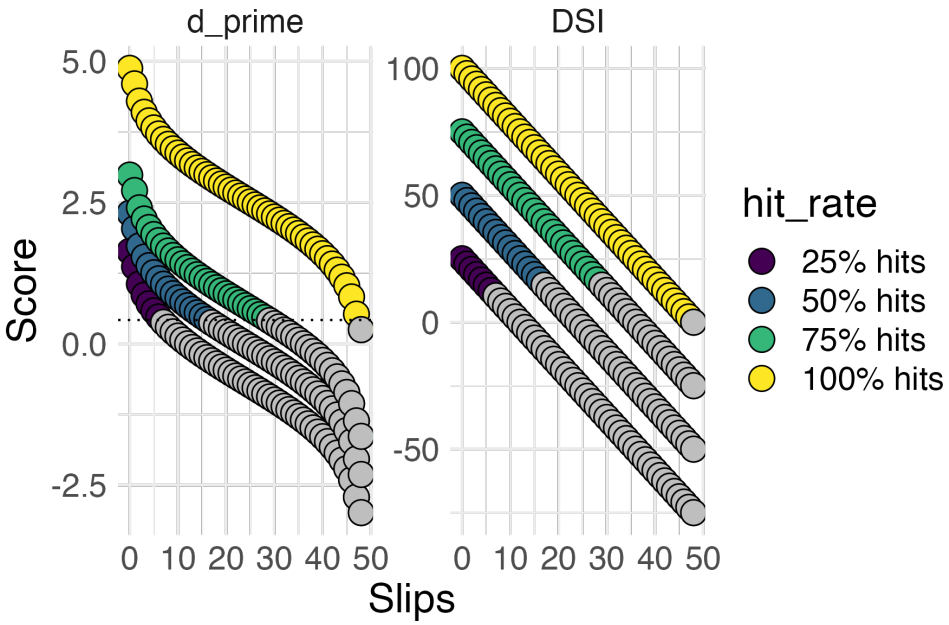
$$D' = Z\left(\frac{Hits}{Hits + Misses}\right) - Z\left(\frac{False\ Alarms}{False\ Alarms + Correct\ Rejections}\right)$$

$$Response\ Bias = \frac{Z\left(\frac{Hits}{Hits + Misses}\right) + Z\left(\frac{False\ Alarms}{False\ Alarms + Correct\ Rejections}\right)}{2}$$

Code for computing discriminability and response bias was modified from Lindeløv (2011) for R. To compute a chance-level threshold of d', 1000 discriminability values were computed based on randomly sampled hit, miss, false alarms, and correct rejection values from a binomial distribution. The d' value at the 95 percentile of this sample, corresponding to d' of 0.42, was considered to be the upper limit of 'chance.' Values above this point in the experimental data were included, and values below excluded from further analysis. This led us to remove 35 participants from analysis, leaving 100 participants in our final sample.

Figure 1

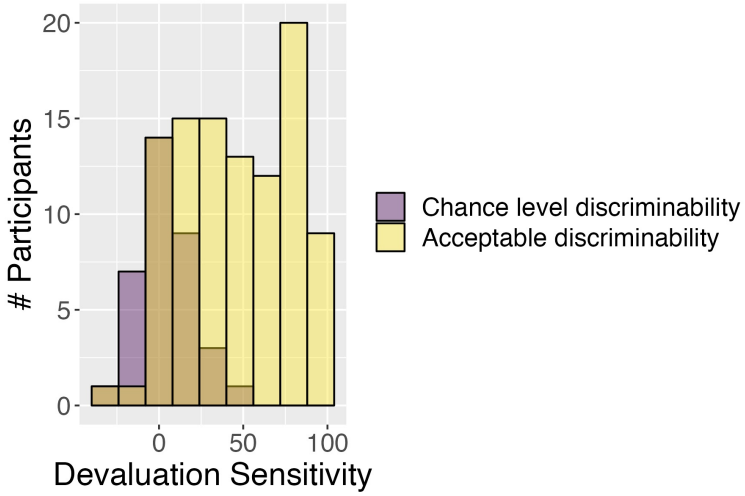
Modeled Discriminability (d') and Devaluation Sensitivity (DSI) at Varying Slips of Action and Hit Rates.



Note. D' and DSI were calculated for all possible values of 'slips' of action (0-48) across different response rates to still-valuable stimuli. Grayed out dots represent below-chance levels of d'. The horizontal dotted line on the left figure shows d' values representing chance level of d' (< 0.42).

Figure 2

Participants with Chance Level Discriminability Scores on the Slips-of-Action Test Spanned A Wide Range of Devaluation Sensitivity Scores



Note. Chance level discriminability included values below d' = 0.42.