Leveraging Psychometrics of Rational Inattention to Estimate Individual Differences in the Capacity for Cognitive Control

Huang Ham
hamhuang@sas.upenn.edu
University of Pennsylvania

Ivan Grahek
ivan_grahek@brown.edu
Brown University

Laura A. Bustamante
lauraab@princeton.edu
Princeton University

Nathaniel Daw
ndaw@princeton.edu
Princeton University

Andrew Caplin
ac1@nyu.edu
New York University

Sebastian Musslick
sebastian_musslick@brown.edu
Brown University

Abstract

Recent years have witnessed significant advances in our understanding of bounds on rationality in both cognitive psychology and economics. These two fields have been making separate progress, but time is ripe for unifying these efforts. In this article, we introduce recently developed economic tools, themselves rooted in the psychometric tradition, to quantify individual differences in the capacity for cognitive control. These tools suggest that a reliable assessment of the capacity for cognitive control may be accomplished by examining task performance as a function of reward. We demonstrate through simulation studies that an incentive-informed measure of task performance does a better job of recovering individual differences in one's capacity for cognitive control, compared to the commonly used congruency effect. Furthermore, we show that the economic approach can be used to predict control-dependent behavior across different task settings. We conclude by discussing future directions for the fruitful integration of behavioral economics and cognitive psychology with the aim of improved measurement of individual differences in the capacity for cognitive control.

Keywords: mental effort; executive function; individual differences; rational inattention; Bayesian revealed preference

Introduction

Cognitive control is a collection of mechanisms that enables us to adapt information processing in service of current task goals. Our capacity to exert control is crucial for accomplishing complex tasks and is related to various real-life outcomes such as work success and good health (Diamond, 2013). Yet, numerous factors can constrain our capacity to exert control (Musslick & Cohen, 2021), including its efficacy and cost. Individual differences in these factors, and the resulting capacity for control, have been theorized to predict performance outside the laboratory (Musslick, Cohen, & Shenhav, 2019). Unfortunately, traditional performance-based measures of cognitive control capacity (i.e. accuracy, response time) remain poor predictors of performance in the real world (Saunders et al., 2018). In this article, we leverage a psychometric framework from behavioral economics—grounded in rational inattention—to derive a novel, choice-based measure of control-dependent behavior and assess its internal and external validity. Our results suggest that the estimation of factors determining the capacity for cognitive control can be improved by assessing performance as a function of incentives.

Researchers have developed a variety of paradigms to index a given person's capacity to exert cognitive control. Most commonly, these paradigms require the participant to categorize a target stimulus while ignoring one or multiple distractors (Dreisbach & Fröber, 2019). For instance, in the Stroop task, participants are required to name the ink color of a color word (e.g., say "red" in response to "GREEN") while ignoring the word. Such interference tasks allow for the computation of a congruency cost: the difference in performance (e.g., accuracy) on trials in which target and distractor are associated with the same response (congruent trials, e.g. "RED" in the Stroop task) and trials in which the distractor interferes with the target to produce a competing response (incongruent trails, e.g., "GREEN" in the Stroop task). Congruency costs have commonly been applied to assess real-world self-control outcomes (e.g., addiction treatment compliance, healthy diets; Streeter et al., 2008; Allan, Johnston, & Campbell, 2010). However, the test-retest ability of interference-based measures in cognitive control tasks is surprisingly low (Hedge, Powell, & Sumner, 2018; Draheim, Tsukahara, Martin, Mashburn, & Engle, 2021). Moreover, task performance measures that are supposed to target the same construct (e.g., the inhibition of distraction information) are poorly correlated with one another (Rouder & Haaf, 2019; Whitehead, Brewer, & Blais, 2019) and are poor predictors of other measures, such as questionnaires that tap into real-life outcomes (Dang, King, & Inzlicht, 2020). This issue is complicated in choice-based paradigms, aiming to assess participants' preference for executing control-demanding tasks (e.g., Kool, McGuire, Rosen, & Botvinick, 2010; Westbrook, Kester, & Braver, 2013) as such preferences can be determined by a multitude of factors other than one's capacity for cognitive control (Musslick et al., 2019).

Mounting theoretical work suggests that the low predictive validity of performance- and choice-based measures derives from a confound between a person's *capacity* for cognitive control and their *willingness* to exert it. For instance, building on a cost-benefit model of control allocation (Shenhav, Botvinick, & Cohen, 2013), Musslick et al. (2019) demonstrate that traditional metrics, such as the congruency effect, can be driven by a variety of factors unrelated to a person's capacity for cognitive control, such as the amount of practice on a task. Related simulation work shows that the performance-enhancing influence of positive affect can be explained in terms of both a reduction in the cost of cognitive control and an enhanced sensitivity to rewarding outcomes (Grahek, Musslick, & Shenhav, 2020). Thus, assessing one's capac-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ity for cognitive control requires dissociating it from factors influencing motivation to exert it. Prior theoretical work suggests that proper estimation of the capacity for control may require an assessment of performance as a function of incentives (Musslick, Cohen, & Shenhav, 2018). This comports with recent efforts in economics on utilizing the psychometric information contained in task accuracy across different incentives (Caplin & Martin, 2015). Borrowing from production theory, this line of work suggests that the amount of attention a person is willing to supply for a given incentive reflects the underlying cost of attention—just like the quantity of product a company is willing to supply for a given price reflects the cost of production (Caplin, Csaba, Leahy, & Nov, 2020; Caplin, 2021). Caplin et al. (2020) leverage this operationalization to quantify the costs of attention from task accuracy without confounding the effect of incentives. This suggests that a similar measure of control-demanding behavior—as a function of incentives—can prove useful for indexing an individual's capacity for cognitive control.

In this article, we leverage the psychometric theory of rational inattention, as introduced by Caplin et al. (2020), to derive an incentive-informed measure of cognitive control capacity. After introducing the measurement procedure, we derive a formal relationship between the introduced metric and the traditional congruency effect. We then build on a well-established model of control allocation, to simulate individual differences in task performance. We examine the internal and predictive validity of the introduced measures, by extracting them from the simulated performance and by correlating them with internal variables that determine an agent's capacity for cognitive control. We demonstrate that the assessment of performance across incentives provides a better index of cognitive control capacity and that it may provide a reliable predictor of cognitive performance across individuals and tasks. We conclude by outlining directions for the joint advancement of economic theory and psychometrics.

The Psychometrics of Rational Inattention

Our psychometric framework builds upon the intellectual tradition of bounded rationality, assuming that human cognition is rational within certain boundaries (Chater & Oaksford, 1999; Anderson, 1990; Griffiths, Lieder, & Goodman, 2015). While an unboundedly rational agent always chooses the 'best' action to take, the actions chosen by a boundedly rational agent may not always be the absolute 'best' but are on average as good as possible without requiring an unreasonable amount of effort.

To formalize this framework in an experimental setting, we use A to denote the set of *actions* selected by a participant (e.g., responding "red" or "green" in a Stroop task). The value of a chosen action typically depends on the experimental condition (e.g., whether the color of the Stroop stimulus is red or green). We call the set of experimental conditions states, denoted by S. As we will demonstrate below, a Stroop task with congruent and incongruent trials may be characterized by two states (congruent and incongruent) with two respective actions (correct or incorrect response). Finally, how good an action $a \in A$ is in a state $s \in S$ is quantified by an expected utility function $U : S \times A \to \mathbb{R}_{\geq 0}$. A participant that is boundedly rational must always choose according to an action strategy that maximizes the expected value of $U(s,a) - K(\text{experiment})$. Here $U(s,a)$ represents the *utility* the agent receives and K represents the (mental) cost the agent has to pay to resolve this choice problem. What we mean by an action strategy is simply the probability of choosing an action given a state, denoted as $p(a|s)$. Hence the expected value of $U(s,a) - K(\text{experiment})$ amounts to

$$\sum_{s \in S} \sum_{a \in A} U(s,a) p(a|s) p(s) - K(\text{experiment})$$

where $p(s)$ denotes the probability of the state s occurring in the experiment. The quantities $p(s)$ and $p(a|s)$ can be easily determined based on the choice data. The probability of each state is set by the experimenter and the marginal probability of an action in a given state can be estimated from the participant's choice data using their choice frequency. Unfortunately, one cannot directly measure a participant's utility function U nor their cost of control K. However, U can be estimated based on the reward structure of the experiment. While there is plenty of work proposing ways to estimate U from rewards under different research contexts, (Caplin et al., 2020) provide a first, general procedure for recovering the mental cost K—henceforth referred to as Caplin et al.'s K—from only the choice data.

To assess K, we require that the same participant performs the task (i.e, identical set of experimental conditions and actions) under different utility functions. This can be accomplished by asking the participant to perform multiple instances of the same experiment but under different incentives. To illustrate this, consider a "baseline" experiment with a given utility function U that we set as a reference point. We require that in all other experiments, the same participant must have a utility function wU for some $w \in \mathbb{R}_{\geq 0}$. Next for each of the experiments, we can compute the marginal probability $p_w(s,a) = p_w(a|s) p_w(s)$ of the state s and the action a from the frequencies of each unique pair of state and action in the choice data [1]. Caplin et al. denote the expected utility of the experiment as a function of w:

$$\hat{u}(w) = \sum_{s \in S} \sum_{a \in A} wU(s,a) p_w(s,a)$$

The main quantity of interest in our framework is the expected utility normalized by the weight w:

$$\bar{u}(w) = \frac{\hat{u}(w)}{w} = \sum_{s \in S} \sum_{a \in A} U(s,a) p_w(s,a) \qquad (1)$$

where $\bar{u}(w)$ is the revealed expected utility calculated using the choice data of the w weighted experiment and the utility function of the base experiment. Caplin et al. (2020) prove

[1] Note that all marginal probabilities must add up to 1.

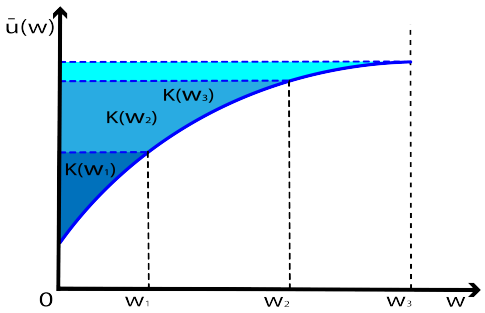

Figure 1: Caplin et al.'s K. $K(w)$ is the area of the region to the left of the revealed expected utility $\bar{u}$ curve between $\bar{u}(0)$ (the revealed expected utility for the base experiment) and $\bar{u}(w)$ (the revealed expected utility for the experiment that provides w times more of an incentive than the base experiment). w_1, w_2, and w_3 are three example values of w.

that if the participant is indeed a boundedly rational agent, $\bar{u}$ should be a monotonically non-decreasing function. If the data supports such monotonicity, one can recover K:

$$K(w) = w\bar{u}(w) - \int_0^w \bar{u}(t)dt. \tag{2}$$

Note that K is a function of w, that is, there is a separate K for each experiment (each with a different reward structure). The geometric intuition is illustrated in Figure 1. To summarize, we can recover K from the empirical choice data using the following steps:

Step 1 Recruit participants for a base experiment at which the participants each has some utility function U. Then, obtain the marginal probabilities p(a, s) for each state and action pair from each participant's choice data.

Step 2 Repeat step 1 for a set of n experiments that are identical to the base experiment except for the reward structure. Importantly, the reward structures differ in such a way that participant's utility function in each experiment is some positive multiple w of U. n should be large enough to estimate how the function $\bar{u}$ depends on w according to Eqn. 1.

Step 3 Compute the revealed expected utility $\bar{u}$ for each experiment, and then compute K according to Eqn. 2.

We illustrate the recovery method based on an application to the Stroop task.

Application to the Stroop Task

In a simple version of the Stroop task, a participant may choose from two relevant *actions*, to indicate the ink color of a color word (accurate) or not (inaccurate):

$$A = \{\text{correct}, \text{incorrect}\}.$$

Two possible experimental states (or conditions) may occur with *equal frequency*: the ink color may match the word (congruent condition) and the ink color mismatches the word (incongruent condition).

$$S = \{\text{congruent}, \text{incongruent}\}.$$

If the action is the correct response (i.e., 'accurate'), the participant gets the reward of 1 point, otherwise she gets 0 points.

Step 1 With the aformentioned Stroop task as our base experiment, the first step is to collect data. In this article, we obtain all the data from a simulated cognitive agent which we introduce in the next section. Because data are simulated, can implement the assumption that reward points equal the participant's utility. Thus, for the base experiment:

$$U(s,a) = \begin{cases} 1 & \text{if } a = \text{correct} \\ 0 & \text{if } a = \text{incorrect} \end{cases}$$

From the data, one can calculate $p(s,a)$ for each pair of state and action. Because actions can be only correct or incorrect, the marginal probability of being correct $\sum_{s\in S} p(s, a = \text{correct})$ can be interpreted as the participant's *accuracy*. Because under this particular utility function, the utility value does not depend on the state amounts to 1 whenever the action is correct, and 0 if it is incorrect, the expected utility is equal to the marginal probability of being correct and can thus be also interpreted as participant's accuracy.

Step 2 Now we need to simulate data from versions of the Stroop experiment with the same S and A, and $\exists w \in \mathbb{R}_{\geq 0}$ such that its reward/utility function

$$U_w(s,a) = \begin{cases} w & \text{if } a = \text{correct} \\ 0 & \text{if } a = \text{incorrect} \end{cases}$$

For each experiment, if the action amounts to the correct response (i.e., 'accurate'), the participant gets the reward of $w * 1 = w$ point, otherwise she gets $w * 0 = 0$ points. Akin to Step 1, the expected utility of each of these data divided by w, i.e. $\bar{u}(w)$, can be understood as the participant's response *accuracy* in the task where the reward is w points, instead of 1 point. Similarly, a participant's accuracy in only the congruent condition is equivalent to $\bar{u}(w|s = \text{congruent})$ and participant's accuracy in only the congruent condition is equivalent to $\bar{u}(w|s = \text{incongruent})$. To emphasize these operational interpretations, from now on we denote

- w as R (reward for accurate responses),
- $\bar{u}(w)$ as $A(R)$ (accuracy as a function of reward),
- $\bar{u}(w|s = \text{congruent})$ as $A_C(R)$ (congruent trial accuracy), and
- $\bar{u}(w|s = \text{congruent})$ as $A_I(R)$ (incongruent trial accuracy).

Step 3 Using our new notations, we can compute the Caplin et al.'s K as $K(R) = RA(R) - \int_0^R A(t)dt$. We next compare and contrast Caplin et al.'s K with the congruency effect, which has been the state-of-the-art indicator of the capacity for cognitive control in the literature. The congruency effect in the Stroop experiment with reward R, denoted by $\Delta A(R)$, is defined as the difference in accuracies on congruent and incongruent trials in the experiment with reward R, i.e., $\Delta A(R) = A_C(R) - A_I(R)$. These two measures have two crucial distinctions. First, Caplin et al.'s K traces its roots to

a rigorous and very general mathematically theory entailed directly by the bounded rationality paradigm, whereas the usage congruency effect is restricted to tasks like the Stroop task and lacks such theoretic connections. Second, while Caplin et al.'s K takes advantage of the accuracy function at all reward levels, the congruency effect only depends on one experiment. Since both measures were computed using the accuracy function, they also formally related:

Proposition 0.1. *If the experiment entails an equal number of congruent and incongruent trials, we can show that,*

$$K(R) - K_I(R) = \frac{R}{2}(\Delta A(R) - \Delta\bar{A}|_0^R)$$

where $K_I(R)$ refers to the Caplin et al.'s K calculated using only the data in the incongruent condition, and $\Delta\bar{A}|_0^R$ denotes the mean of congruency effects from reward 0 to R.

Proof. By definition,

$$K(R) = RA(R) - \int_0^R A(t)dt$$

Because we assume there are the same number of trials in congruent and incongruent condition,

$$= R\frac{A_C(R)+A_I(R)}{2} - \int_0^R \frac{A_C(t)+A_I(t)}{2}dt$$

$$= R\frac{\Delta A(R)+2A_I(R)}{2} - \int_0^R \frac{\Delta A(t)+2A_I(t)}{2}dt$$

$$= \frac{R}{2}\Delta A(R) + RA_I(R) - \int_0^R \frac{\Delta A(t)}{2}dt - \int_0^R A_I(t)dt$$

$$= \frac{R}{2}\Delta A(R) - \frac{R}{2}\Delta\bar{A}|_0^R + RA_I(R) - \int_0^R A_I(t)dt$$

$$= \frac{R}{2}(\Delta A(R) - \Delta\bar{A}|_0^R) + K_I(R)$$

$$K(R) - K_I(R) = \frac{R}{2}(\Delta A(R) - \Delta\bar{A}|_0^R)$$

□

The relationship between Caplin et al.'s K and the congruency effect is illustrated in Figure 2. However, while the existence of such a relationship serves as an interesting theoretic result, it does not imply that they are equally useful as estimators of one's capacity for cognitive control. We will proceed to validating these metrics using simulated data from a computational model of control allocation below. The corresponding simulation code, as well as the domain-general recovery method is available on GitHub (`https://github.com/musslick/Psychometrics-of-Rational-Inattention`).

Assessing the Internal and Predictive Validity of Caplin et al.'s K

The validity of any performance-based metric of control-demanding behavior can be assessed based on its ability to (a) track internal variables of the control system, and (b) predict behavior across tasks. So far, we considered three different metrics: the traditional congruency effect (assessed at

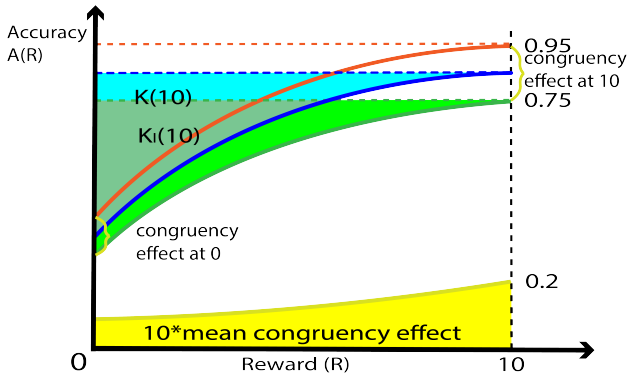

Figure 2: Caplin et al.'s K and mean congruency effect $\Delta\bar{A}$. Red curve: average accuracy in the congruent condition (A_C); green curve: average accuracy in the incongruent condition (A_I); blue curve: the average accuracy for the entire task (A); yellow curve: congruency effect (ΔA). Blue region: Caplin et al.'s K of the entire task; green region: Caplin et al.'s K of only the incongruent condition; Yellow region: 10 times the mean congruency effect ($10\Delta\bar{A}|_0^{10}$) where 10 corresponds to the highest reward.

reward $= 0$), Caplin et al.'s K, and the mean congruency effect, averaged across reward conditions. In this section, we extract these metrics from the behavior of a simulated agent in an interference task. The agent implements a model of control allocation based on the Expected Value of Control (EVC; Shenhav et al., 2013) which has been used to simulate a wide array of empirical findings concerning human performance (e.g., Musslick, Shenhav, Botvinick, & Cohen, 2015; Grahek, Shenhav, Musslick, Krebs, & Koster, 2019; Grahek et al., 2020; Shenhav, Straccia, Musslick, Cohen, & Botvinick, 2018; Bustamante, Lieder, Musslick, Shenhav, & Cohen, 2021). We then examine how well the introduced metrics relate to latent variables that determine the agent's capacity of control, and how well they can predict performance in a second, unrelated task.

The Expected Value of Control Model

The EVC theory by Shenhav et al. (2013) proposes that an agent determines the optimal amount of control by maximizing the expected value of control, that is, the expected utility of implementing a control signal with a given intensity u minus an intrinsic cost that scales with the intensity of the signal

$$\text{EVC}(u,S) = P(\text{correct}|u,S)V(R) - \text{Cost}(u)$$

where $P(\text{correct}|u,S)$ is the probability of responding correctly on a task, $V(R)$ represents the value of committing a correct response for a given incentive R (e.g., monetary reward), and $\text{Cost}(u)$ amounts to the cost of implementing the control signal with intensity u. It is hypothesized that the control system chooses to implement the control signal with the maximal expected value of control:

$$u^* = \underset{u}{\text{argmax}} EVC(u,S).$$

Parameterization of Simulated EVC Agents

For the simulations reported in this article, we assume that the probability of responding correctly in an interference task increases monotonically with the amount of control intensity allocated, following the sigmoid function:

$$P(\text{correct}|u,S) = \frac{1}{1+e^{\varepsilon \cdot u - a}}$$

where ε characterizes the control efficacy, i.e. how an increase in control signal intensity translates into changes in task accuracy. The parameter a determines the degree of task automaticity: The higher a, the easier the task, that is, the less cognitive control is needed to reach the correct outcome[2]. Note that $a = a_0 - i$ depends on the amount of interference i in the current trial. For congruent trials, we assume that the agent receives no interference $i = 0$ whereas in incongruent trials, the agent receives some amount of interference $i > 0$ (effectively lowering the probability of responding correctly).

The subjective value can be described as a function of the reward R provided for a correct response. Here we assume that the subjective value of the correct outcome amounts to

$$V(R) = v \cdot R + b$$

where R is a monetary reward that is provided in the event of a correct response, v is the reward sensitivity of the agent and b is the baseline value that the agent assigns to being accurate (accuracy bias). Finally, the EVC agent is assumed to allocate control by taking into account an intrinsic cost that scales with control signal intensity. For the simulations reported below, we chose an exponential cost function,

$$Cost(u) = e^{c \cdot u} - 1$$

where c scales the increase in the cost of control with one unit of control signal intensity u. To simulate individual differences in EVC agents, we sampled parameters of the model from the uniform intervals listed in Table 1. The intervals were chosen such that the agents would allocate at least some amount control ($u^* > 0$) for each task condition. Note that the control efficacy ε and the control cost c both characterize an agent's capacity to exert control across different tasks.

Table 1: Uniform intervals of sampled EVC parameters.

Parameter	Description	Min	Max
e	control efficacy	4	5
c	control cost	0.8	0.9
v	reward sensitivity	0.95	1.05
a_A	automaticity of Task A	-4	-2
a_B	automaticity of Task B	-4	-2
i_A	distractor interference of Task A	-2	-1
i_B	distractor interference of Task B	-2	-1
b	accuracy bias	5	5

Simulation Procedure

Following the estimation procedure outlined above, we simulated behavior for different reward conditions, ranging from $R = 0$ to $R = 10$ in steps of 0.01. For each reward condition, we assessed the agent's performance in congruent trials ($i_A = 0$) and incongruent trials ($i_A > 0$) of an interference task, henceforth referred to as Task A. We also simulate the behavior of an EVC agent in an additional task that we refer to as Task B. We ensured that Task A and Task B were unrelated to one another, by sampling task automaticities and interference effects for both tasks independently for each agent. Thus, an agent's automaticity a_B and interference i_B for Task B can be considered independent from their automaticity a_A and interference i_A for Task A. We simulated behavior for a total of 1000 EVC agents whose parameters were drawn from uniform distributions listed in Table 1.

Analysis

For each reward and trial condition, we computed the agent's congruency effect by subtracting $P(\text{correct}|u,S)$ in incongruent trials from $P(\text{correct}|u,S)$ in incongruent trials, i.e. $\Delta A = P(\text{correct}|u,S) - P(\text{correct}|u,S)$. We also derived Caplin et al.'s $K(R = 10)$, by integrating the accuracy of each simulated agent across all reward conditions (Figure 2, blue area). Finally, we computed the related, mean congruency effect across reward conditions (cf. Proposition 0.1).

To examine the internal validity of each metric, we assessed the Pearson correlation of that metric with (1) the control efficacy ε as well as (2) the control cost c, across the 1000 simulated agents. We assessed the predictive validity by correlating the metrics derived from Task A with the congruency effect at r=0 in Task B. The integral term involved in calculating K was approximated using the trapezoid rule. To remove the effect of simulation sample size, non-parametric bootstrapping with 6000 samples was used to obtain the standard error of the absolute value of each correlation coefficient and the p-value of the difference in correlation coefficients.

Results: Internal Validity

To compare the internal validity in task A, we correlated our measures each with the simulating parameter values. We found that all of our three measures correlate significantly with the control cost and control efficacy parameters ($|t| > 10$, $\text{df} = 998$, $p < 10^{-15}$). However, the correlation between the control cost and the congruency effect for $R = 0$ ($r = 0.302$) is significantly weaker than between the mean congruency effect from $R = 0$ to $R = 10$ ($r = 0.313$, $p = 0.025$) and between Caplin et al.'s K at $R = 10$ ($r = 0.355$, $p = 0.002$). The correlation between the control cost and the mean congruency effect from $R = 0$ to $R = 10$ is also significantly weaker than that between Caplin et al.'s K at $R = 10$ ($p = 0.008$). However, the correlation between the control efficacy and the congruency effect for $R = 0$ ($r = -0.555$) is not significantly weaker than between the mean congruency effect from $R = 0$ to $R = 10$ ($r = -0.559$, $p = 0.155$). However, note that both measures yield a significantly weaker correlation than Caplin et al.'s K at $R = 10$ ($r = -0.669$, $p < 10^{-3}$). This result indicates that Caplin et al.'s K can better predict the control cost and efficacy than the congruency effect.

[2]The automaticity of a task may be influenced by a number of factors, such as prior experience with the task.

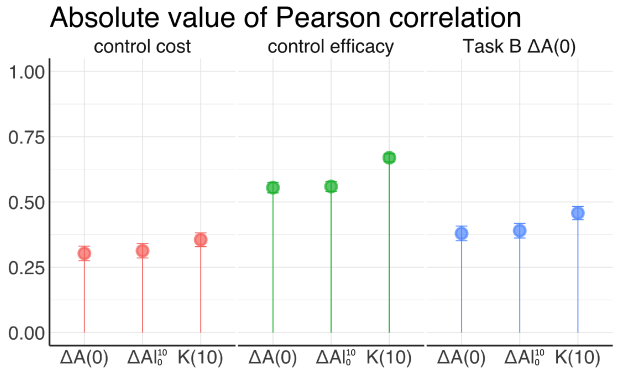

Figure 3: Correlation strength. The x-axis represents different measurements calculated from the simulated Stroop task. The y-axis represents the absolute value of Pearson correlation between the measurement and the quantity shown in the panel label. $\Delta A(0)$ is the congruency effect at reward 0, $\Delta A|_0^{10}$ is the mean congruency effect from reward 0 to 10, K(10) is Caplin et al.'s K at reward 10. Error bars are bootstrapped confidence intervals using 6000 samples.

Results: Predictive Validity

We found that all three metrics derived from the agents' performance in Task A correlate significantly with the congruency effect at $R = 0$ in Task B ($t > 12$, df $= 998$, $p < 10^{-15}$). Similarly, the correlation between the the congruency effect at reward 0 in task B and the congruency effect for $R = 0$ in task A ($r = 0.379$) is significantly weaker than between the mean congruency effect from $R = 0$ to $R = 10$ in task A ($r = 0.390$, $p = 0.017$). Caplin et al.'s K at $R = 10$ in task A correlates more strongly with the congruency effect in Task B than the other two measures ($r = 0.458$, $p < 10^{-3}$). This indicates that Caplin et al.'s K can predict better the congruency effect of another cognitive control task.

General Discussion

People rely on cognitive control to adjust how they process information in order to achieve their goals. Although reliable measurement of the capacity to exert cognitive control is crucial for predicting important outcomes (e.g., mental and physical health, job success; Diamond, 2013), current measures of control (e.g., the congruency effect) show poor internal and external validity (Hedge et al., 2018; Saunders et al., 2018). Here we propose that one of the causes of this issue is that the current measurement approaches are focused on measuring control capacity at a fixed point. While there is extensive research showing that motivation increases how much cognitive control people exert (Botvinick & Braver, 2015; Parro, Dixon, & Christoff, 2018; Shenhav, Fahey, & Grahek, 2021), current approaches measure cognitive control capacity at a fixed motivational point, rather than considering how the amount of control changes as a function of incentives. In this work, we introduced Caplin et al.'s K as a novel measurement technique that relies on the integration of task performance across reward conditions. This measurement device is implied by economic theories of rational inattention, resulting in a formal relationship with existing measures (Caplin et al., 2020; Caplin, 2021).

Building on the bounded rationality framework, we exposed a relationship between Caplin et al.'s K and the congruency effect measured as a function of linearly increasing incentives. We examined the validity of these metrics based on the behavior of simulated agents implementing a well-established theory of control allocation (Shenhav et al., 2013; Musslick et al., 2015). We find that Caplin et al.'s K does a better job of predicting latent variables that determine an agent's capacity for cognitive control (control efficacy and control cost) compared to the traditional congruency effect. Moreover, Caplin et al.'s K can be leveraged to predict an agent's behavior across unrelated interference tasks. These results provide initial support for the idea that considering how performance varies as a function of incentive levels is crucial for assessing the capacity for cognitive control.

While Caplin et al.'s K shows promise in advancing the quantification of control-demanding behavior, there are important next steps to take. First, our simulations call for replication with data collected from human participants. However, humans may not satisfy our simplifying assumption that expected utility scales linearly with the reward offered for correct performance. Addressing this may require a variation of prize probabilities as opposed to rewards (Caplin et al., 2020) or the consideration of prospect theory (Tversky & Kahneman, 1992). Moreover, calculating Caplin et al.'s K for noisy empirical data might require fitting the accuracy function $A(w)$. This could be done either via assuming the shape of the function or relying on symbolic regression (Udrescu et al., 2020). Second, here we have focused exclusively on task accuracy, but accuracy is known to tradeoff with reaction times in cognitive tasks (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). Thus, future work is needed to incorporate existing models of the reaction time-accuracy tradeoff into the economic theory of rational inattention.

Despite current limitations, the introduced measurement device opens up new vistas for the systematic characterization of cognitive control across paradigms. Given that Caplin et al.'s K can be computed for arbitrary experiment conditions, it may help quantify and tie together the demands for cognitive control across different paradigms (e.g., task switching or dual-tasking). This includes varying demands for cognitive control as a function of expected trial frequencies, e.g., the relative proportion of congruent versus incongruent trials (Logan & Zbrodoff, 1979). Finally, Caplin et al.'s K promises to expose a formal relationship between the avoidance of cognitive control (assessed in terms of task preferences; Kool et al., 2010) and task performance. A combination of psychometrics and rational inattention may offer novel paths to integrating the various ways in which researchers assess cognitive control. Such an integrative effort may not only further the cognitive control capacity but advance concomitant theorizing of rational inattention in behavioral economics.

References

Allan, J. L., Johnston, M., & Campbell, N. (2010). Unintentional eating. what determines goal-incongruent chocolate consumption? *Appetite*, *54*(2), 422–425.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. New York: Psychology Press. doi: 10.4324/9780203771730

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Botvinick, M. M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology*, *66*(1), 83-113.

Bustamante, L., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. (2021). Learning to overexert cognitive control in a stroop task. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(3), 453–471.

Caplin, A. (2021, October). *Economic data engineering* (Working Paper No. 29378). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w29378` doi: 10.3386/w29378

Caplin, A., Csaba, D., Leahy, J., & Nov, O. (2020). Rational inattention, competitive supply, and psychometrics. *The Quarterly Journal of Economics*, *135*(3), 1681–1724.

Caplin, A., & Martin, D. (2015). A Testable Theory of Imperfect Perception. *The Economic Journal*, *125*(582), 184–202. doi: 10.1111/ecoj.12130

Chater, N., & Oaksford, M. (1999, February). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. doi: 10.1016/S1364-6613(98)01273-X

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in cognitive sciences*, *24*(4), 267–269.

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135–168.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242.

Dreisbach, G., & Fröber, K. (2019). On how to be flexible (or not): Modulation of the stability-flexibility balance. *Current Directions in Psychological Science*, *28*(1), 3–9.

Grahek, I., Musslick, S., & Shenhav, A. (2020). A computational perspective on the roles of affect in cognitive control. *International Journal of Psychophysiology*, *151*, 25–34.

Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. (2019). Motivation and cognitive control in depression. *Neuroscience & Biobehavioral Reviews*, *102*, 371–381.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, *50*(3), 1166–1186.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665–682.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a stroop-like task. *Memory & cognition*, *7*(3), 166–174.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775.

Musslick, S., Cohen, J. D., & Shenhav, A. (2018). Estimating the costs of cognitive control from task performance: theoretical validation and potential pitfalls. In *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 800–805). Madison, WI.

Musslick, S., Cohen, J. D., & Shenhav, A. (2019). Decomposing individual differences in cognitive control: a model-based approach. *In Proceedings of the 41st Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Montreal, CA*, 2427-2433.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *Reinforcement Learning and Decision Making Conference.*

Parro, C., Dixon, M. L., & Christoff, K. (2018). The neural basis of motivational influences on cognitive control. *Human brain mapping*, *39*(12), 5097–5111.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review*, *26*(2), 452–467.

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., Inzlicht, M., & Vazire, S. (2018). Reported self-control is not meaningfully associated with inhibition-related executive function: A bayesian analysis. *Collabra: Psychology*, *4*(1).

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Fahey, M. P., & Grahek, I. (2021). Decomposing the motivation to exert mental effort. *Current Directions in Psychological Science*, *30*(4), 307-314. doi: 10.1177/09637214211009510

Shenhav, A., Straccia, M. A., Musslick, S., Cohen, J. D., & Botvinick, M. M. (2018). Dissociable neural mechanisms track evidence accumulation for selection of attention versus action. *Nature communications*, *9*(1), 1–10.

Streeter, C. C., Terhune, D. B., Whitfield, T. H., Gruber, S., Sarid-Segal, O., Silveri, M. M., ... others (2008). Performance on the stroop predicts treatment compliance in

cocaine-dependent individuals. *Neuropsychopharmacology*, *33*(4), 827–836.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., & Tegmark, M. (2020, December). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *arXiv:2006.10782 [physics, stat]*.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? load, trait, and aging effects revealed by economic preference. *PLoS One*, *8*(7), e68210.

Whitehead, P. S., Brewer, G. A., & Blais, C. (2019). Are cognitive control processes reliable? *Journal of experimental psychology: learning, memory, and cognition*, *45*(5), 765.