Title: Rapid, precise, and reliable phenotyping of delay discounting using Bayesian adaptive design optimization

Authors:
Woo-Young Ahn[1,2], Hairong Gu[2], Yitong Shen[3], Nathaniel Haines[2], Julie E. Teater[4], Jay I. Myung[2], and Mark A. Pitt[2]

[1] Department of Psychology, Seoul National University, Seoul, Korea
[2] Department of Psychology, The Ohio State University, Columbus, OH
[3] Department of Psychiatry, Indiana University-Purdue University Indianapolis, Indianapolis, IN
[4] Department of Psychiatry and Behavioral Health, The Ohio State University, Columbus, OH

Correspondence:
Woo-Young Ahn, Ph.D.
Department of Psychology
Seoul National University
Seoul, Korea 08826
Tel: +82-2-880-2538, Fax: +82-2-877-6428. Email: wahn55@snu.ac.kr

Keywords: Impulsivity, Computational Psychiatry, Bayesian Data Analysis, Adaptive Design Optimization, Delay Discounting

Number of words in the abstract: 132
Number of words in the main text (including references): 5,547
Number of Figures/Tables in the main text: 2 / 1
Number of Figures/Tables in supplemental online material: 10 / 0
Number of References: 38

Conflict of Interest: The authors declare no competing financial interest.

Abstract

Machine learning has the potential to facilitate the development of computational methods that improve the measurement of cognitive and mental functioning. In three populations (college students, patients with a substance use disorder, and Amazon Mechanical Turk workers), we evaluated one such method, Bayesian adaptive design optimization (ADO), in the area of delay discounting by comparing its test-retest reliability, precision, and efficiency with that of a conventional staircase method. The results showed that ADO led to 0.95 or higher test-retest reliability of the discounting rate within 10-20 trials (under 1-2 minutes of testing) in all three populations tested, captured approximately 10% more variance in test-retest reliability, was 3-5 times more precise, and was 3-8 times more efficient than the staircase method. The ADO methodology provides efficient and precise protocols for phenotyping individual differences.

Introduction

Precision medicine (Insel, 2014) proposes we use individually tailored treatment and prevention programs for each patient (Collins & Varmus, 2015) to maximize their efficacy. Its goal is to identify (bio)markers of individual differences and treatment outcomes on the basis of neurobiological or cognitive tests. While precision medicine is being widely used in the treatment of cancer (Friedman, Letai, Fisher, & Flaherty, 2015), there is growing interest in its application in psychiatry and general mental functioning (Insel, 2014), as reflected in the Research Domain Criteria initiative advanced by the National Institute of Mental Health.

A formidable challenge in applying precision medicine to mental functioning is improving measurement. We focus on three: reliability, precision, efficiency. It is difficult to measure reliably latent neurocognitive constructs or biological processes, such as impulsivity, reward sensitivity, or learning rate. While recent advancements in neuroscience and computational psychiatry (Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014) provide novel frameworks, cognitive tasks, and latent constructs that allow us to investigate the neurocognitive mechanisms underlying psychiatric conditions, their reliabilities have not been rigorously tested or are not yet acceptable (Hedge, Powell, & Sumner, 2017). A recent large-scale study suggests that the test-retest reliability of cognitive tasks are only modest (Enkavi, Eisenberg, Bissett, Mazza, & Poldrack, 2018). Even if a test is reliable across time, confidence in the measurement will depend on its precision at each measurement. To our knowledge, few studies have rigorously tested the precision of measures from a neurocognitive test. Lastly, cognitive tasks developed in research laboratories are not always efficient, often taking 10-20 minutes to administer. With lengthy and relatively demanding tasks, participants (especially clinical populations) can easily fatigue or be distracted (Sandry, Genova, Dobryakova, DeLuca, & Wylie, 2014), which can increase measurement error due to inconsistent responding. A by-product of low task efficiency is that the amount of data (e.g., number of participants) typically available for big data approaches to studying psychiatry is smaller than in other fields.

Bayesian adaptive testing is a promising machine-learning method that can address the aforementioned challenges and thereby improve behavioral precision medicine (Cavagnaro, Myung, Pitt, & Kujala, 2010; Myung, Cavagnaro, & Pitt, 2013). It originates from optimal experimental design in statistics (Atkinson & Donev, 1992) and from active learning in machine learning (Cohn, Atlas, & Ladner, 1994). It is an algorithm-based Bayesian methodology for designing optimal experiments that lead to rapid and accurate parameter inference about the

phenomenon under study with the fewest possible number of measurement episodes. It is a form of adaptive testing in which the values of design variables (e.g., stimulus properties and task parameters) to use in the next trial are determined online in real time based on the data collected from the preceding trials, so as to be maximally informative about the question of interest (e.g., what is the attention span of a 7-year old? how impulsive is this individual?). It differs from traditional adaptive techniques such as the staircase method (Leek, 2001) in that a parametric model of the underlying psychological process guides stimulus choice on each trial. The methodology and its variants are being applied across disciplines to improve the efficiency and informativeness of data collection (cognitive psychology (Cavagnaro, Aranovich, McClure, Pitt, & Myung, 2016; Myung & Pitt, 2009), vision (Gu et al., 2016; Lesmes, Jeon, Lu, & Dosher, 2006), psychiatry (Aranovich, Cavagnaro, Pitt, Myung, & Mathews, 2017), neuroscience (DiMattina & Zhang, 2011; Lewi, Butera, & Paninski, 2009), clinical drug trials (Wathen & Thall, 2008), and systems biology (Kreutz & Timmer, 2009)).

Here, we demonstrate the successful application of adaptive design optimization (ADO), an implementation of Bayesian adaptive testing, to improving measurement in the delay discounting task. Delay discounting is a strong candidate endophenotype for addictive disorders (Anokhin, Grant, Mulligan, & Heath, 2014; Bickel, 2015) and risky behaviors (for a review see, Green & Myerson, 2004). The construct validity of delay discounting has been demonstrated in numerous studies. For example, the delay discounting task is widely used to assess (altered) temporal impulsivity of various psychiatric disorders, including patients with substance use disorders (e.g., Green & Myerson, 2004), schizophrenia (Ahn et al., 2011; Heerey, Robinson, McMahon, & Gold, 2007), and bipolar disorder (Ahn et al., 2011). We show that in three different populations (college students, patients with substance use disorders, and the online testing community), ADO leads to rapid, precise, and reliable estimates of the delay discounting rate (k) with the hyperbolic function. Test-retest reliability of k reached up to 0.95 or higher within 10-20 trials (under 1-2 minutes of testing) with at least three times greater precision and efficiency than the staircase method (Mazur, 1987).

Methods

Experiment 1 (college students)

In Experiment 1, we recruited college students (N=58) to evaluate test-retest reliability (TRR) of the ADO and staircase (SC) methods over a period of approximately one month, a span

of time over which one might want to measure changes in impulsivity. Previous studies have typically used 1 week (e.g., Matusiewicz, Carter, Landes, & Yi, 2013), 2 weeks (Harrison & McKay, 2012), or 3-6 months (Weatherly & Derenne, 2013) between test sessions. Students visited the lab twice. In each visit they completed two ADO and two SC sessions, allowing us to measure TRR within and between sessions. In each session, students made 42 choices about hypothetical scenarios involving a larger but later reward versus a smaller but sooner reward. We examined TRR (using Pearson correlation coefficients) within each visit and between the two visits, using the discounting rate k of the hyperbolic function, as the outcome measure.

Participants. Fifty-eight students at The Ohio State University (25 males and 33 females; age range 18-37 years; mean 19.0, SD 2.9 years) were recruited. They were required to be at least 18 years of age and received course credits for their participation. For all studies reported in this work, we used the following exclusion criterion: a participant is excluded from further analysis if the participants' standard deviation (SD) of a parameter value is two SD greater or smaller than the group mean. In other words, we excluded participants who seemingly made highly inconsistent choices during the task.

Delay discounting task. Each participant completed two sessions, which were separated by approximately one month (mean=28.3 days, SD=5.3 days). In each session, a participant completed four delay discounting tasks: two ADO-based tasks and two staircase-based tasks. Each ADO-based or staircase-based task included 42 trials. The order of task completion (ADO then staircase versus the reverse) was counterbalanced across participants.

In the traditional staircase method, a participant initially made a choice between $400 now and $800 at seven different delays: one week, two weeks, one month, six months, one year, 3 years, and 10 years. Order of the delays was randomized for each participant. By adjusting the immediate amount, the choices were designed to estimate the participant's indifference point for each delay (1). See (Ahn et al., 2011; Green & Myerson, 2004) for the details of the procedure.

In the ADO method, the sooner delay and a later-larger reward were fixed as 0 day and $800. A later delay and a sooner reward were experimental parameters that were optimized on each trial. Based on the ADO framework and the participant's choices so far, the most informative design (a later delay and a sooner reward) was selected on each trial.

Computational modeling. We applied ADO to the hyperbolic function, which has two parameters (k: discounting rate and β:inverse temperature rate). The hyperbolic function has the form $V = A / (1 + kD)$, where an objective reward amount A after delay D is discounted to a

subjected reward value V for an individual whose discounting rate is k (>0). Typically in a delay discounting task, two options are presented on each trial: a sooner-smaller (SS) reward and a later-lager (LL) reward. The subjective values of the two options are modeled by the hyperbolic function. We used softmax (Luce's choice rule) to translate subjective values into the choice probability on trial t:

$$P(LL\ over\ SS) = \frac{1}{1 + e^{\beta(V_{ss(t)} - V_{LL(t)})}}$$

Where V_{ss} and V_{LL} are subjective values of the SS and LL options. To estimate the two parameters of the hyperbolic model in the staircase method, we used the hBayesDM package (Ahn, Haines, & Zhang, 2017). The hBayesDM package (https://github.com/CCS-Lab/hBayesDM) offers hierarchical and non-hierarchical Bayesian analysis of various computational models and tasks using the Stan software (Carpenter et al., 2016). The hBayesDM function of the hyperbolic model for estimating a single subject's data is *dd_hyperbolic_single*. Note that updating of our ADO framework is based on each participant's data only. Thus for fair comparisons between ADO and staircase methods, we used an individual (non-hierarchical) Bayesian approach for the analysis of data from the staircase method. In ADO sessions, the parameters, means and SDs of the parameter posterior distributions of the hyperbolic model, are automatically estimated on each trial. Note that estimation of discounting rate (k) was of primary interest in this project. Estimates of the inverse temperature rate (a measure of response consistency or a degree of exploration/exploitation), β, are provided in the Supplemental Figures, but will not be discussed further.

Experiment 2 (patients meeting criteria for a substance use disorder)

In Experiment 2, we recruited 35 patients meeting DSM-V criteria for a substance use disorder (SUD) to assess the performance of ADO in a clinical population. The experimental design was the same as in Experiment 1 except that there was only a single visit.

Participants. Twenty-eight individuals meeting Diagnostic and Statistical Manual of Mental Disorders (5th ed. DSM-V) criteria for a substance use disorder and receiving treatment for addiction problems participated in the experiment (25 males and 10 females; age range 22-57 years; mean 35.8, SD 10.3 years). All patients were recruited through in-patient units at The Ohio State University Wexner Medical Center, seeking a treatment for their addiction problems.

All patients received the Structured Clinical Interview for DSM-V Axis I disorders (SCID-I), which was conducted by trained graduate students and a study coordinator (Y.S.). Final diagnostic determinations were made by Woo-Young Ahn on the basis of patients' medical records and the SCID-I interview. Exclusion criteria for all individuals included head trauma with loss of consciousness for over 5 minutes, a history of psychotic disorders, and history of seizures or electroconvulsive therapy, and neurological disorders. Participants received gift cards for their participation (worth of $10/hr).

Delay discounting task and computational modeling. The task and methods for computational modeling in Experiment 2 were identical to those in Experiment 1. For a subset of participants in Experiment 2 (15 out of 35), the upper bound for discounting rate (*k*) during ADO was set as 0.1 for computing efficiency and we noted that some participants' *k* values reached ceiling (=0.1). For the other participants (n=20), the upper bound was set to 1. We report results that are based on all 35 patients (**Figure 2A & 2B**) as well as results without participants whose *k* values reached the ceiling of 0.1 (**Figure S9**).

Experiment 3 (large online sample)

In Experiment 3, we evaluated the durability of the ADO method, assessing it in a less controlled environment than the preceding experiments and with a larger and broader sample of the population, (808 Amazon Mechanical Turk workers). Each participant completed two ADO sessions, each of which consisted of 20 trials, which was estimated from Experiments 1 and 2 to be sufficient.

Participants. Eight hundred and eight individuals through Amazon Mechanical Turk (MTurk; 353 males and 418 females (37 individuals declined to report their sex); age mean 35.0, SD 10.8 years) were recruited. They were required to reside in the United States and be at least 18 years of age, and received $10/hr for their participation. Out of 808 participants, 71 participants (8.78%) were excluded based on the exclusion criteria (see Experiment 1)

Delay discounting task. Each participant completed two consecutive ADO-based tasks, each of which consisted of 20 trials (c.f., 42 trials per session in Experiments 1 and 2). There was no break between the two tasks, so participants experienced the experiment as a single session. The task was identical to the ADO version in Experiment 1.

All participants received detailed information about the study protocol and gave written informed consent in accordance with the Institutional Review Board at The Ohio State University, OH, USA.

Results

Past work customized the staircase method to yield very good TRR (Green & Myerson, 2004). In visits 1 and 2 of Experiment 1 (college students), we obtained mean values of 0.910 and 0.932, respectively. Nevertheless, ADO bested this performance, yielding values of 0.964 and 0.977, an improvement of approximately 10% in terms of the amount of variance accounted for (**Figures S1 and S2; Figures S3 and S4** show the results for all participants, including the outliers).

Where ADO excels more significantly over the staircase method is in efficiency and precision. We measured the efficiency of the method by calculating how many trials are required to achieve the maximum TRR, which was assessed cumulatively at each trial (**Figure 1**). With ADO, we achieved over 0.95 TRR within 10-20 trials at visit 1. At visit 2, TRR exceeded 0.95 within 10 trials. With the staircase method, TRR failed to reach 0.9 even at the end of experiment (42 trials) at visit 1, and reached 0.9 only after 39 trials at visit 2. ADO yielded approximately 3-5 times more precise estimates of discounting rate as measured by the smaller standard deviation of the posterior distribution of the parameter, *k* (ADO visit 1: 0.122, visit 2: 0.098; SC visit 1: 0.413, visit 2: 0.537; **Figure S5**).

ADO also showed superior performance when examined across visits separated by one month (**Figure S6**). TRR measures converged at around 0.8 within 10 trials and were highly consistent with each other. In contrast, with the staircase method, the trajectories of the four measures were much more variable and asymptote, if at all, below 0.8 towards the end of the experiment. The results of Experiment 1 show that ADO leads to rapid, reliable, and precise measures of discounting rate.

Figure 2A and 2B show that even in the patient population (Experiment 2, patients with a SUD), ADO still led to rapid, reliable, and precise estimates of discounting rates, again outperforming the staircase method. With ADO, maximum TRR was 0.976 and it reached this value within approximately 15 trials. Consistent with the results of Experiment 1, the staircase method led to a smaller maximum TRR (0.899) and it took on average 25 trials to reach this maximum (**Figure S7**). Precision of the parameter estimate was five times higher when using

ADO than the staircase method (0.073 vs. 0.371). **Figures S8** shows the results for all participants including the outliers in Experiment 2. While the upper bound of *k* was set as 0.1 for 15 patients and some patients' *k* values reached ceiling, **Figure S9** suggests that the results largely remain the same whether we exclude those patients or not.

In Experiment 3 (Amazon Mechanical Turk workers), ADO again led to an excellent maximum TRR (0.965), greater than 0.9 within 10 trials as shown in **Figure 2C-D**. **Figure S10** shows the results for all participants, including outliers.

Table 1 summarizes the results across the three experiments. Comparison of the two methods clearly shows that ADO is (1) more reliable (capturing 10% more variance in TRR), (2) approximately 3-5 times more precise (smaller SD of individual parameter estimates), and (3) approximately 3-8 times more efficient (fewer number of trials required to reach maximum or 0.9 TRR). As might be expected, when tested in a less controlled environment (Experiment 3), precision suffers (0.371), being more comparable to that found with the staircase method, while reliability and efficiency hardly change.

Discussion

In three different populations, we have demonstrated that ADO led to highly reliable, precise, and rapid measures of discounting rate. ADO outperformed the staircase method in college students (Experiment 1) and in patients meeting DSM-V criteria for SUDs (Experiment 2). It held up very well in a less restrictive testing environment with a broader sample of the population (Experiment 3). The results of this study are consistent with previous studies employing ADO (Cavagnaro et al., 2016; Hou et al., 2016), showing improved precision and efficiency. This is the first study demonstrating the advantages of ADO-driven delay discounting in healthy controls and psychiatric/online populations.

The staircase method is an impressive heuristic method that delivers such good TRR (close to 0.90 in our study) that there is little room for improvement. Nevertheless, ADO is able to squeeze out additional information to increase reliability further. Where ADO excels relative to the staircase method is in precision and efficiency. The model-guided Bayesian inference that underlies ADO is responsible for this improvement. Unlike the staircase method, which follows a simple rule of increasing or decreasing the value of a stimulus, ADO has no such constraint, choosing the stimulus that is expected to be most informative on the next trial. Trial after trial,

this flexibility pays significant dividends in precision and efficiency, as the results of the three experiments show.

The benefits of ADO also come with costs. For example, trials that are most *informative* can be ones that are also difficult for the participant (Ahn & Busemeyer, 2016). Repeated presentation of difficult trials can frustrate and fatigue participants. Another issue is that for participants who respond consistently, the algorithm will quickly narrow to small region of the design space and present the same trials repeatedly with the goal of improving precision even further. It is therefore important to implement measures that mitigate such behavior. We did so in the present experiment by inserting easy trials among difficult ones once the design space narrowed, keeping the total number of trials fixed. Another approach is to implement stopping criteria, such as ending the experiment once parameter estimation stabilizes for a three consecutive trials.

Both ADO and staircase methods are different version of a task, and as such led to slightly different values of discounting rates: The correlation between k from ADO and k from the staircase method is around 0.7. That the association is not higher should not be surprising. As mentioned above, ADO is more flexible than the staircase method in the design choices selected from trial to trial. While the staircase method is constrained to choosing among a few neighboring designs, there are no such limitations on ADO. This difference in flexibility will impact the final parameter estimate, especially in a short experiment. While we cannot say whether estimates using ADO are closest to individuals' true internal states, its high consistency within and especially across visits (**Figure S4**) demonstrates a degree of trustworthiness.

While we believe that ADO is an exciting, promising method that offers the potential to advance the current state of the art in precision medicine and computational psychiatry, in all fairness, we should mention a few major challenges and limitations in its practical implementation. One is the requirement of ADO that a computational/mathematical model of the experimental task is available. Also, the model should provide a good account (fit) of choice behavior; otherwise ADO might lead to even poorer TRR or other psychometric measures. We believe the success of ADO in the delay discounting task is partly thanks to the availability of a reasonably good and simple hyperbolic model with just two free parameters. The mathematical details of ADO and programming code for ADO experiments can be another serious hurdle. To reduce such barriers and allow even users with limited knowledge in ADO algorithms to utilize

ADO in their research, we are developing user-friendly tools (Python-based package, web-based and smartphone platforms) for the research and clinical community.

Lastly, while we demonstrate the promise of an ADO method only in the area of delay discounting in this work, our methodology can be easily extended to other cognitive tasks that are of interest to researchers in psychiatry, psychology, decision neuroscience, and related fields where experimentation is at the core of scientific advances. For example, we can apply ADO to tasks involving value-based or social decision making, including choice under risk and ambiguity (Levy, Snell, Nelson, Rustichini, & Glimcher, 2010) and social interactions (e.g., Xiang, Lohrenz, & Montague, 2013). In addition, ADO can be used to optimize the sequence of stimuli and improve functional magnetic resonance imaging (fMRI) measurement (Bahg et al., 2018), which will reduce the cost of data acquisition and improve the quality of neuroimaging data.

In conclusion, the results of the current study suggest that machine-learning based tools such as ADO can improve the measurement of latent neurocognitive processes and thereby assist in the development of assays for precision medicine in mental health and more generally advance measurement in the behavioral sciences.

Table 1. Comparison of ADO and Staircase in their reliability, precision, and efficiency of estimating temporal discounting rates.

Measures		ADO	Staircase
Maximum test-retest reliability (TRR)	Experiment 1 (College students), Visit 1	0.964	0.910
	Experiment 1 (College students), Visit 2	0.977	0.932
	Experiment 2 (Patients w/ SUDs)	0.976	0.899
	Experiment 3 (Amazon Mturk) **	0.965	N/A
Within-subject variability (SD of individual parameter estimates, a measure of precision)	Experiment 1 (College students), Visit 1	0.122 (0.105)	0.413 (0.252)
	Experiment 1 (College students), Visit 2	0.098 (0.070)	0.537 (0.409)
	Experiment 2 (Patients w/ SUDs)	0.073 (0.063)	0.371 (0.180)
	Experiment 3 (Amazon Mturk) **	0.339 (0.262)	N/A
Number of trials required to reach 0.9 test-retest reliability, a measure of efficiency	Experiment 1 (College students), Visit 1	7	Failed to reach 0.9 even after 42 trials
	Experiment 1 (College students), Visit 2	5	39
	Experiment 2 (SUD Patients)	11	27
	Experiment 3 (AMT participants) **	11	N/A

** **Note**: Experiments 1 and 2 had 42 trials. Experiment 3 had 20 trials.

Figure 1. Comparison of ADO and Staircase (SC) test-retest reliability of temporal discounting rates when assessed cumulatively in each trial (ADO) or every third trial (SC) (Experiment 1, college students) over two visits separated by approximately one month. In each visit, a participant completed two ADO sessions and two SC sessions. Test-retest reliability was assessed cumulatively in each trial.

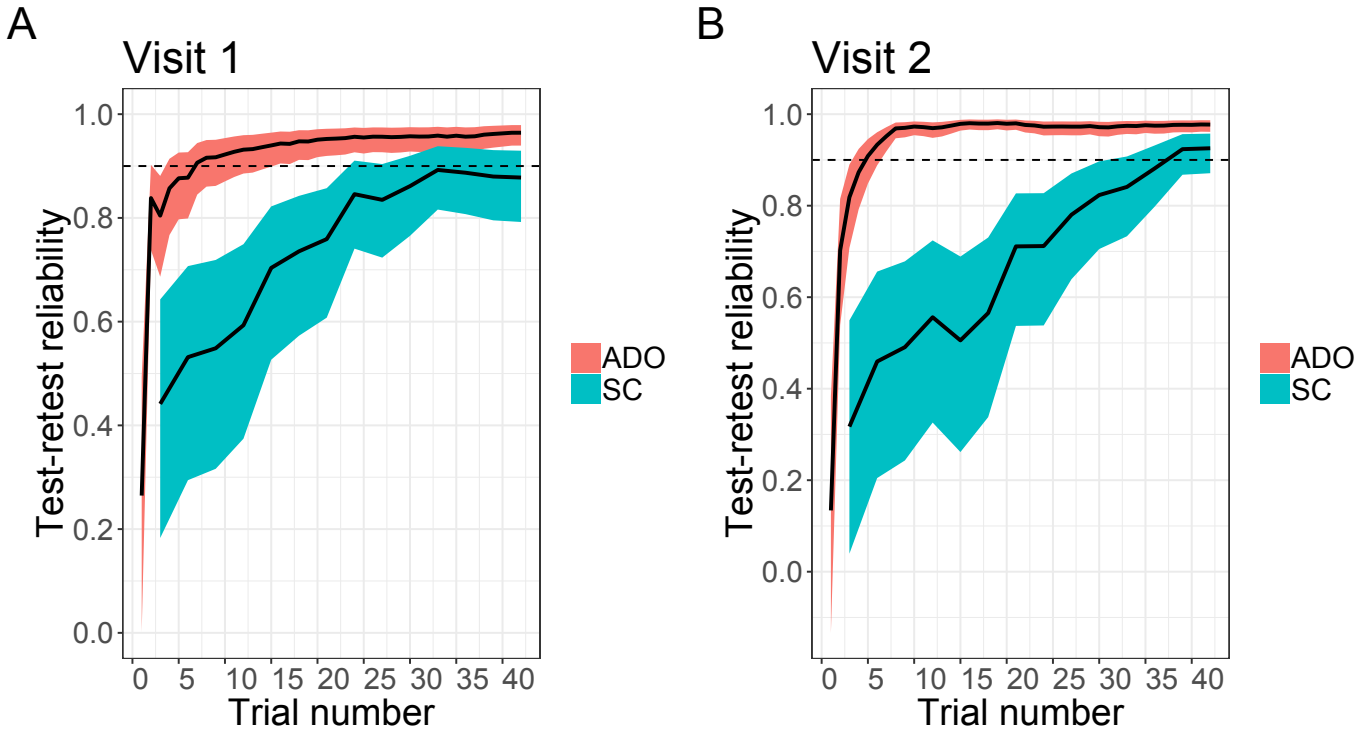


Figure 2. Reliability and efficiency of the ADO method in Experiments 2 and 3 (A) Test-retest reliability of temporal discounting (log(k); the natural logarithms of temporal discounting rates) among patients with substance use disorders (SUDs) across two ADO sessions (Experiment 2) (B) Test efficiency as measured by the cumulative test-retest reliability across trials (Experiment 2). Dashed line = 0.9 test-retest reliability. (C-D) Same as A-B but for the online (Amazon MTurk) workers (Experiment 3). Unlike Experiments 1 and 2, only 20 trials were administered per session.

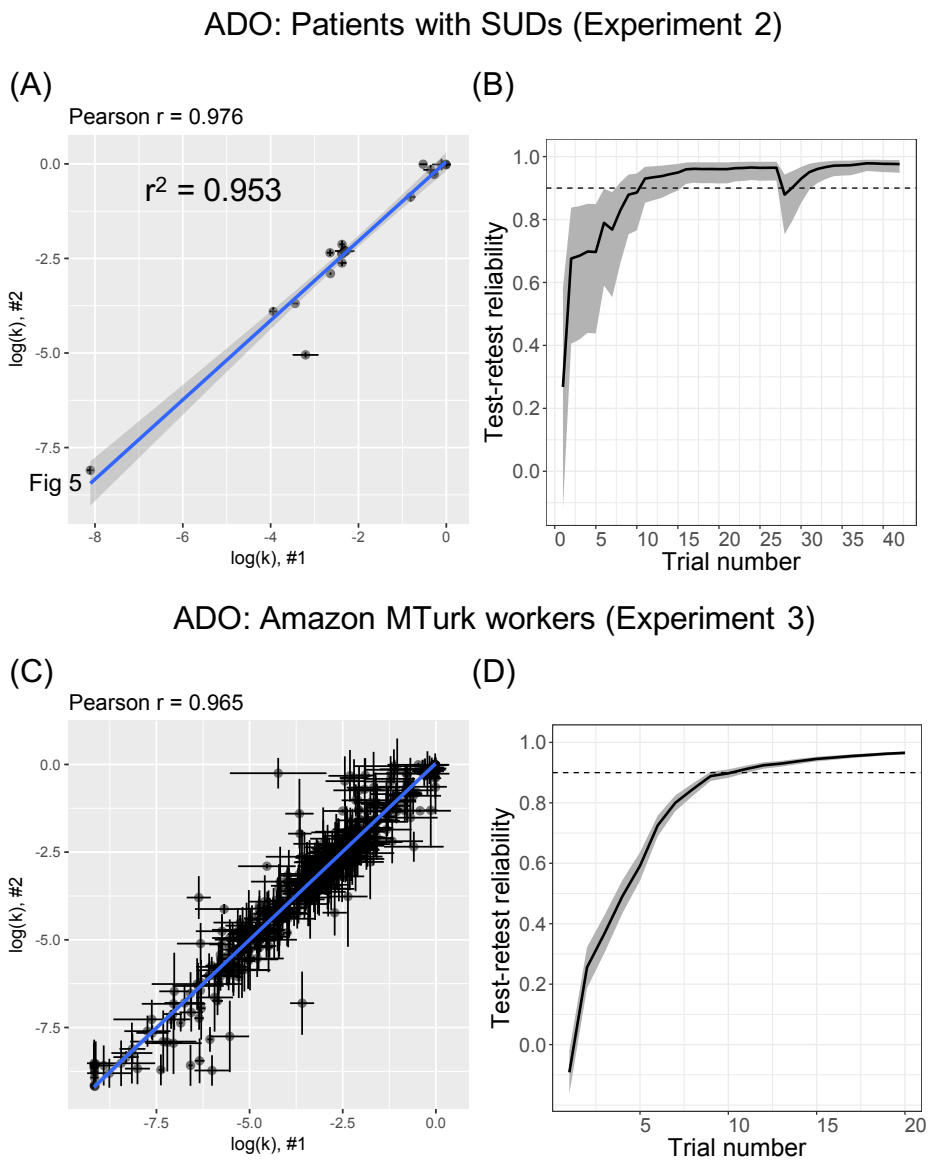


Figure S1. Test-retest reliability of temporal discounting rates across two ADO sessions (Experiment 1, college students, with all 42 trials per session)

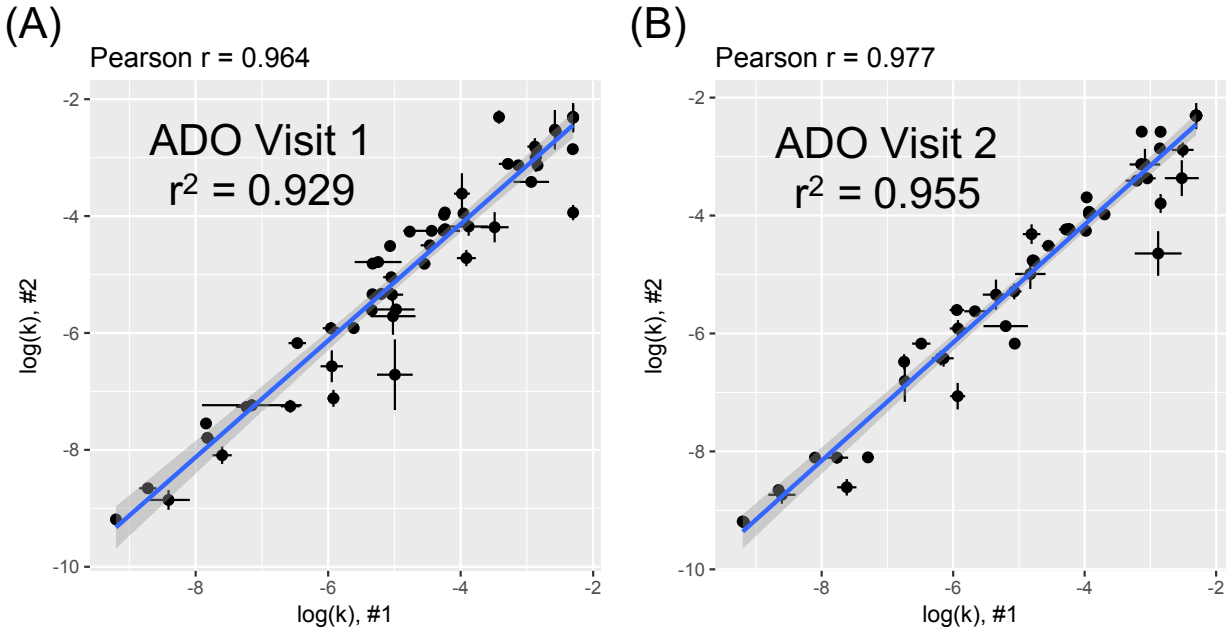


Figure S2. Test-retest reliability of discounting rates across two staircase sessions (Experiment 1, college students, with all 42 trials per session).

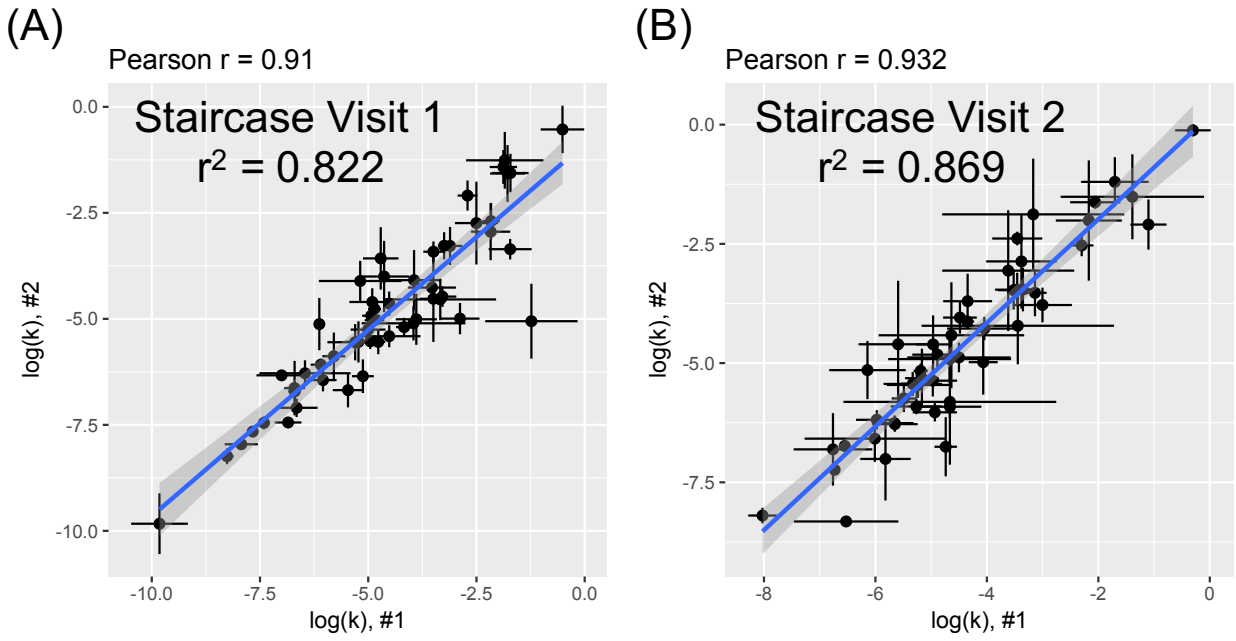


Figure S3. Test-retest reliability of temporal discounting rates (log(k), A & C) and inverse temperature parameters (β, B & D) among college students (Experiment 1) with ADO including outliers, which are indicated as red circles. See **Methods and Materials** for the description of outliers. (A & B) At visit 1, (C & D) At visit 2, which was separated by approximately one month from visit 1.

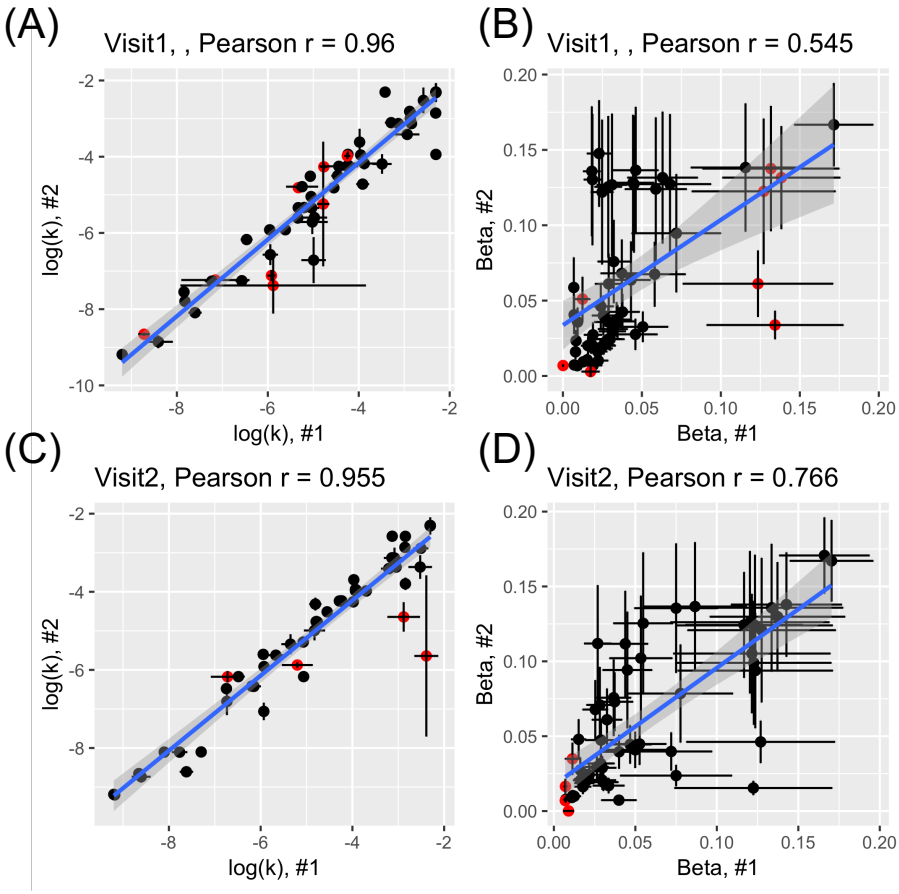


Figure S4. Test-retest reliability of temporal discounting rates (log(k), A & C) and inverse temperature parameters (β, B & D) among college students (Experiment 1) with the staircase method including outliers, which are indicated as red circles. See **Methods and Materials** for the description of outliers. (A & B) At visit 1, (C & D) At visit 2, which was separated by approximately one month from visit 1.

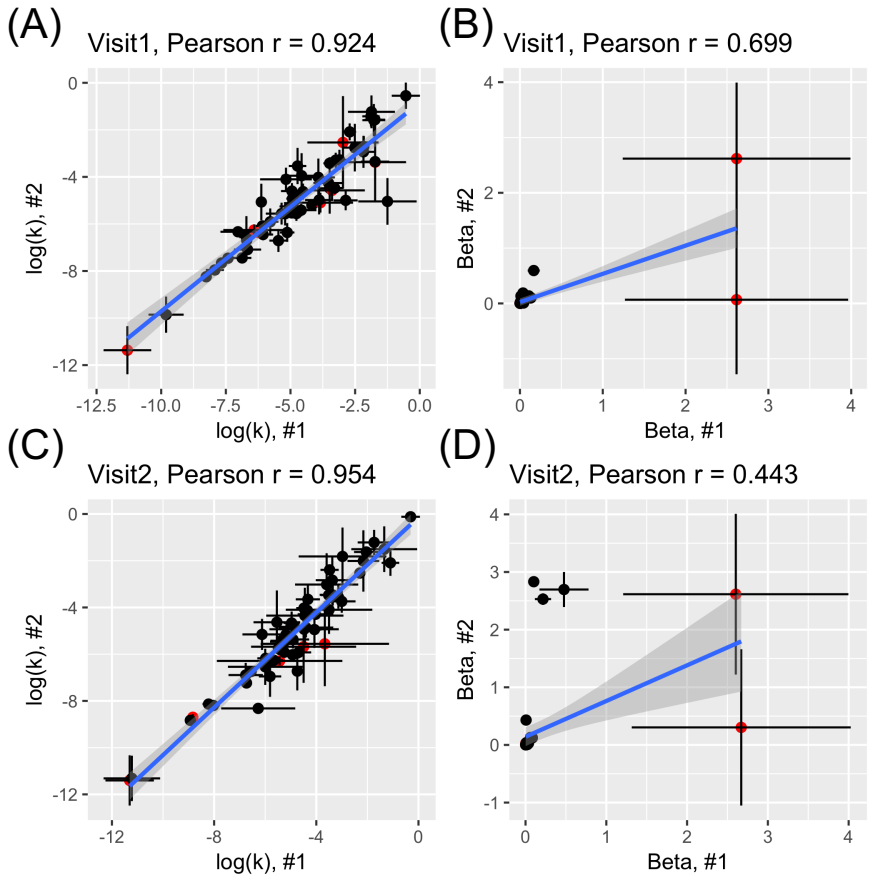


Figure S5. Comparison of ADO and staircase methods with respect to their precision of parameter estimates (Experiment 1, college students). As a measure of precision, we used the standard deviation (SD) of an individual parameter (natural log of the posterior distribution of the temporal discounting rate (k)). Thus, the smaller SD is, the greater its precision is. (A) At visit 1 (B) At visit 2, which was separated by approximately one month from visit 1.

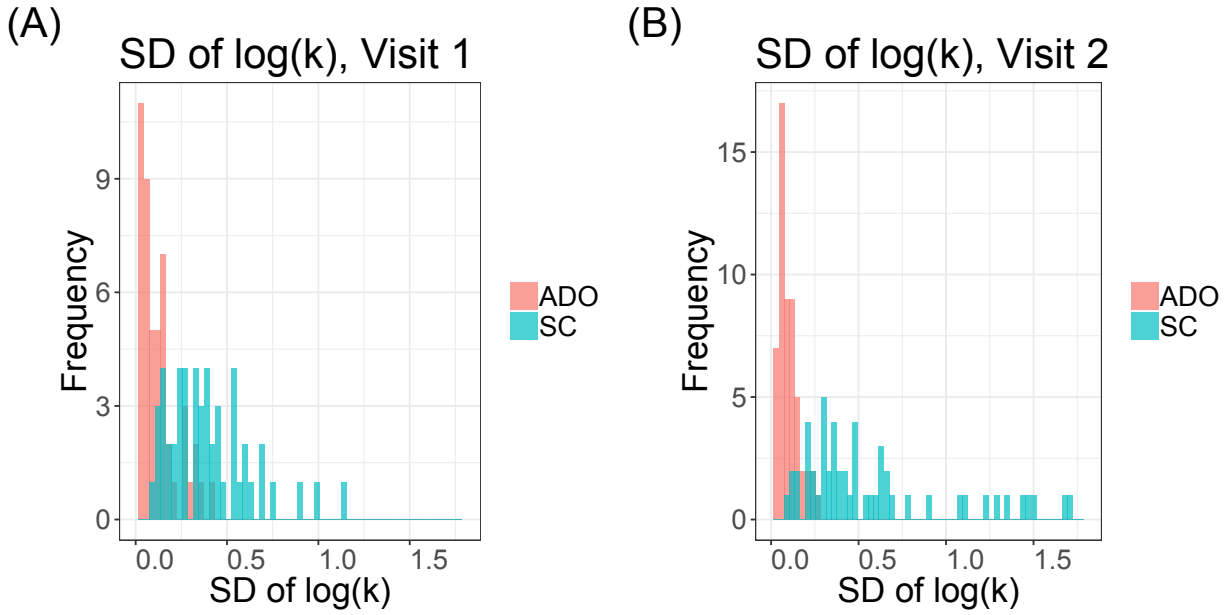


Figure S6. Comparison of efficiency of the ADO (A) and the staircase (B) methods (Experiment 1, college students). Efficiency is measured as the cumulative test-retest reliability in each trial (ADO) or every third trial (staircase). The two visits were separate by approximately one month. Each line represents a different test-retest reliability comparison across the two visits (1. visit1-session1 vs visit2-session1 (red); 2. visit1-session1 vs visit2-session2 (green); 3. visit2-session2 vs visit2-session1 (cyan); 4. visit1-session2 vs visit2-session2 (purple)).

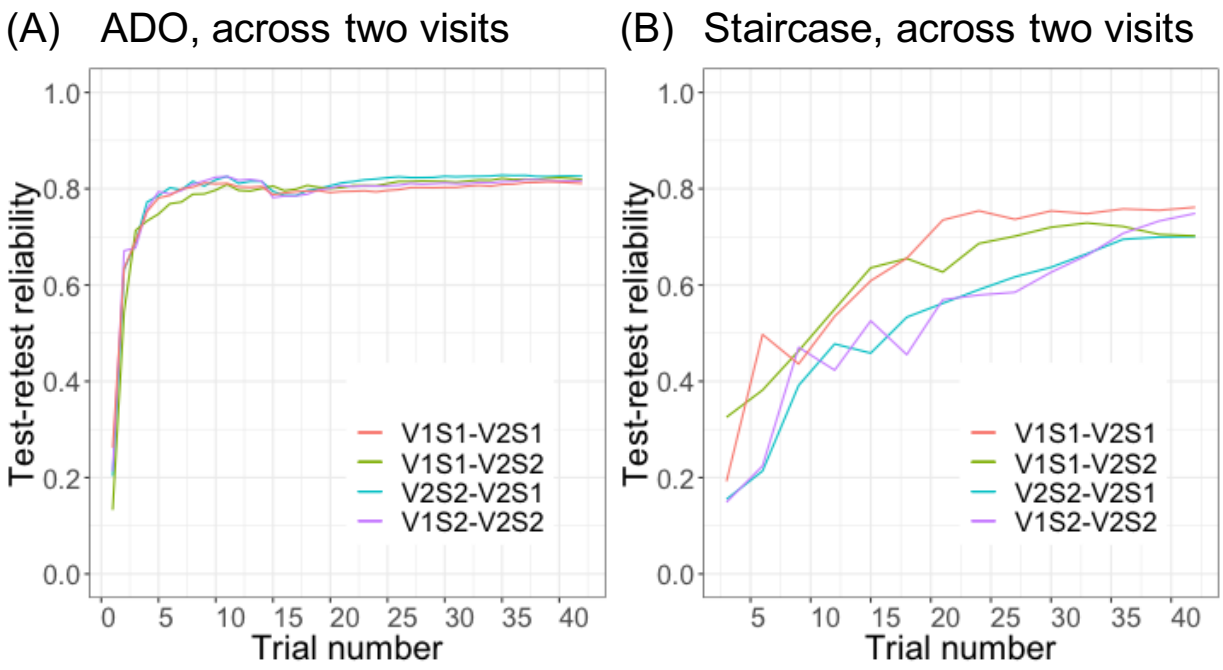


Figure S7. Reliability and efficiency of the staircase method in Experiment 2 (patients with substance use disorders (SUDs)). (A) Test-retest reliability with all 42 trials per session. (B) Test efficiency as measured by the cumulative test-retest reliability in every third trial. See **Figure 2A** to compare the performance of ADO and staircase methods in this population.

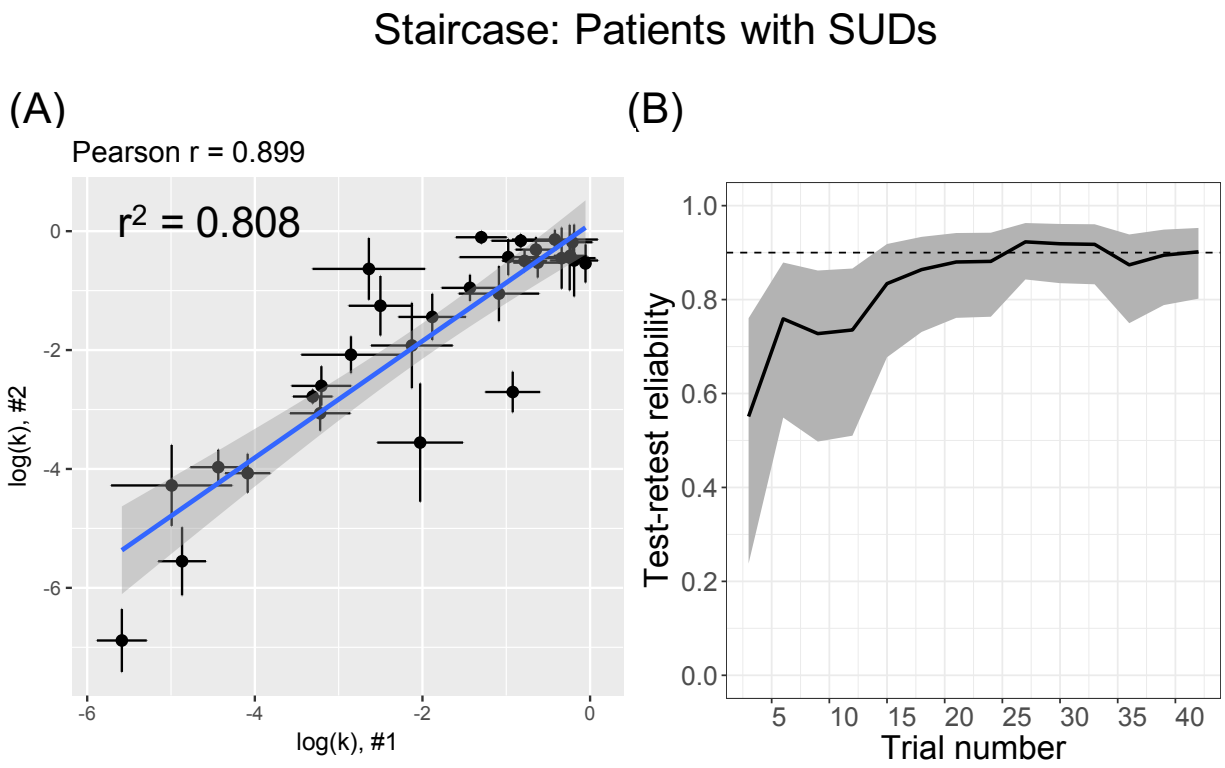


Figure S8. Test-retest reliability of discounting rates (log(k), A & C) and inverse temperature parameters (β, B & D) among patients with substance use disorders (Experiment 2) with ADO and staircase methods including outliers, which are indicated as red circles. See **Methods and Materials** for the description of outliers. (A & B) With ADO, (C & D) With the staircase method.

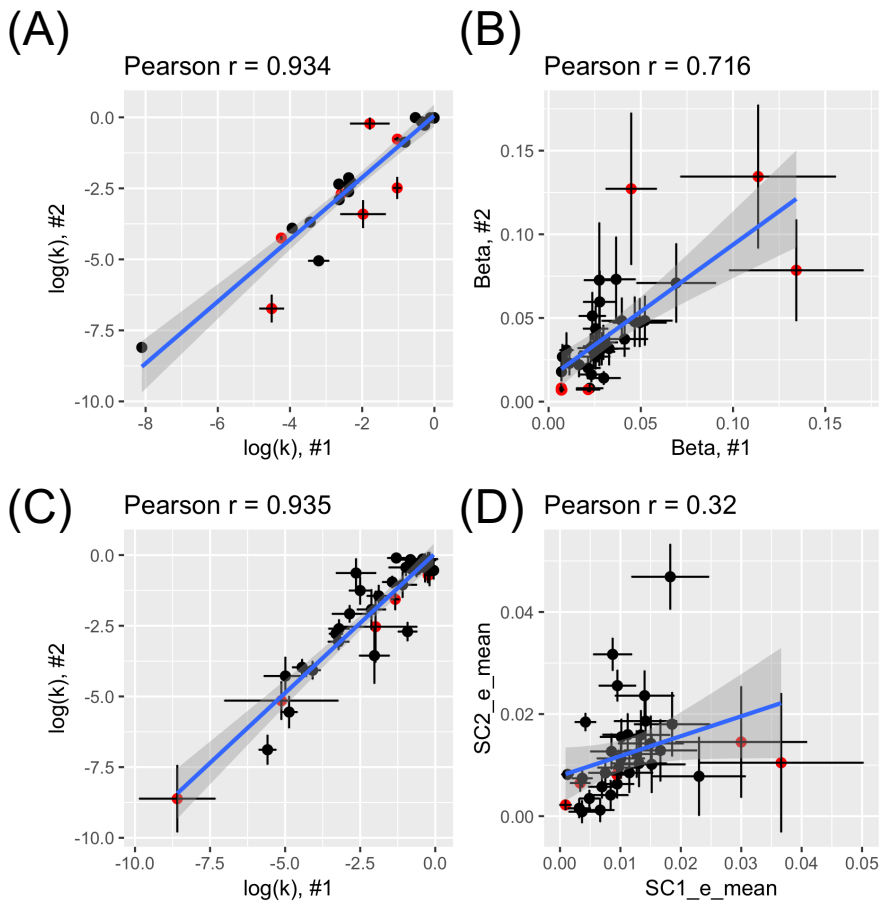


Figure S9. Reliability and efficiency of the staircase method in Experiment 2 (N=15 patients with substance use disorders (SUDs)) with ADO. Out of 35, 14 patients whose *k* values reached the upper bound (=0.1) were first excluded and 6 additional patients were excluded with the 2SD rule (c.f., **Figure 2A &2 B**). (A) Test-retest reliability with all 42 trials per session. (B) Efficiency metric using cumulative test-retest reliability across trials.

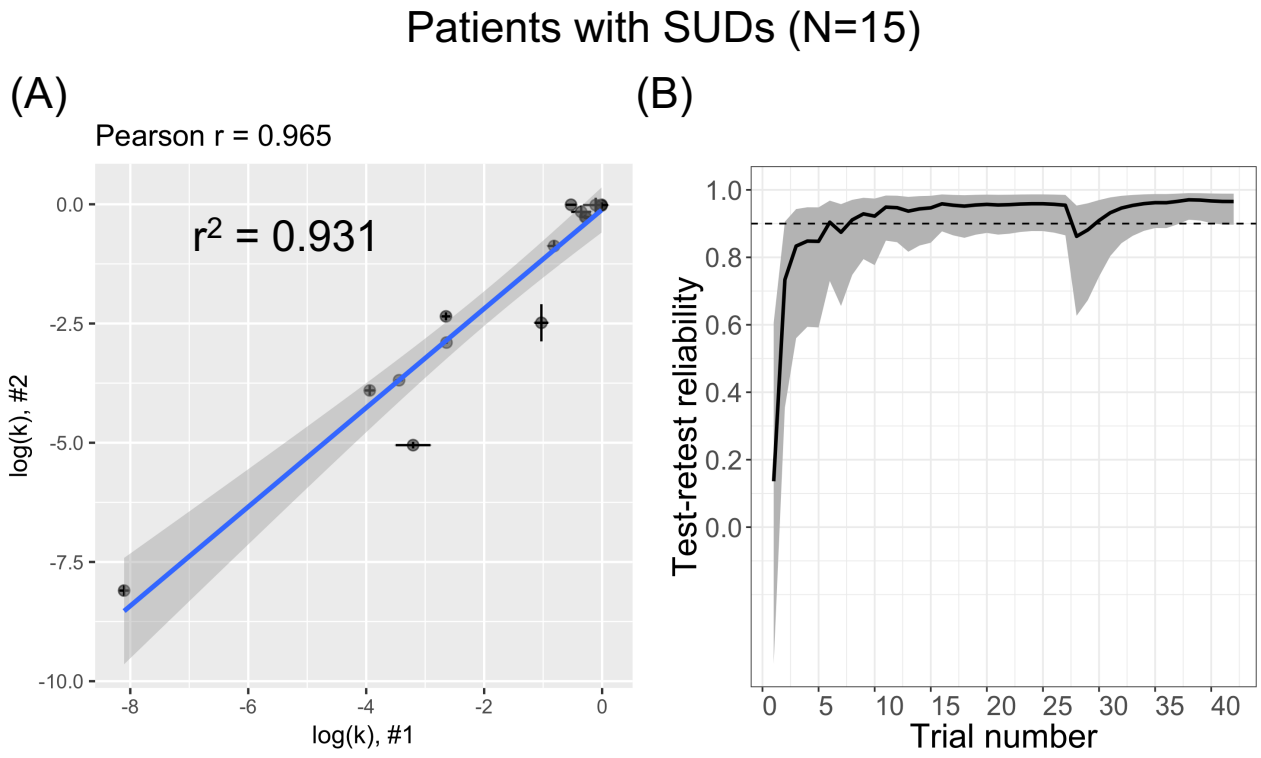
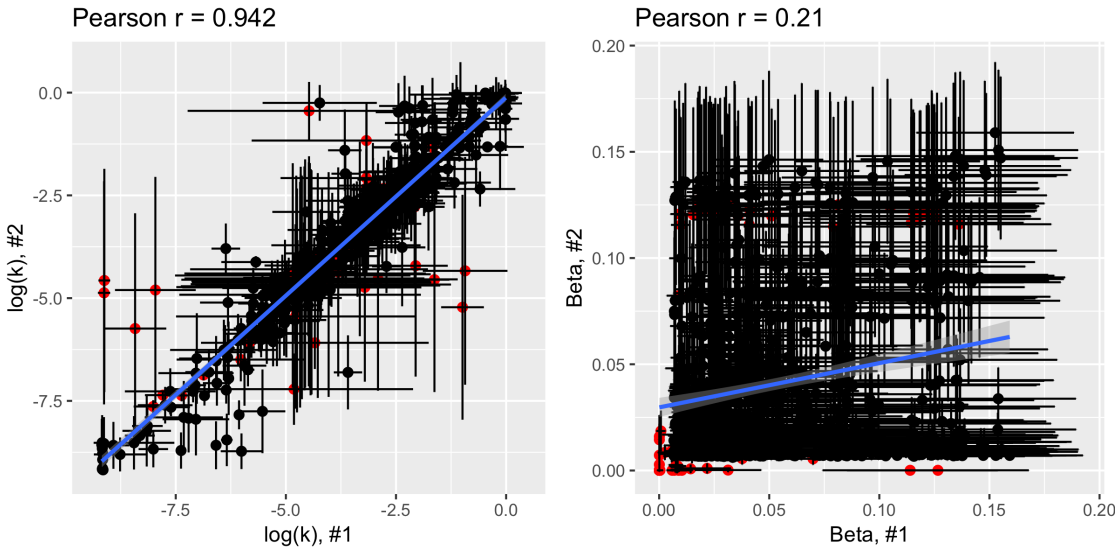


Figure S10. Test-retest reliability of discounting rates (log(k)) and inverse temperature parameters (β) among large online Amazon MTurk participants (Experiment 3) with ADO including outliers, which are indicated as red circles. See **Methods and Materials** for the description of outliers.



Acknowledgement

The research was supported by National Institute of Health Grant R01-MH093838 to M.A.P. and J.I.M, and the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT, & Future Planning (2018R1C1B3007313) to W.-Y.A. We thank Andrew Rogers, Hunter Hahn, and Zoey Butka for their assistance in data collection.

Author contributions

W.-Y.A., J.M., and M.P. conceived and designed the experiments. Y.S. and N.H. collected data. W.-Y.A. performed the data analysis and drafted the paper. J.M. and M.P. provided critical revisions. All authors wrote the paper and approved the final version of the paper for submission.

References

Ahn, W.-Y., & Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences*, *11*, 1–7.

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*, 24–57. http://doi.org/10.1162/CPSY_a_00002

Ahn, W.-Y., Rass, O., Fridberg, D. J., Bishara, A. J., Forsyth, J. K., Breier, A., et al. (2011). Temporal discounting of rewards in patients with bipolar disorder and schizophrenia. *Journal of Abnormal Psychology*, *120*(4), 911–921. http://doi.org/10.1037/a0023333

Anokhin, A. P., Grant, J. D., Mulligan, R. C., & Heath, A. C. (2014). The Genetics of Impulsivity: Evidence for the Heritability of Delay Discounting. *Biological Psychiatry*, *77*(10), 887–894. http://doi.org/10.1016/j.biopsych.2014.10.022

Aranovich, G. J., Cavagnaro, D. R., Pitt, M. A., Myung, J. I., & Mathews, C. A. (2017). A model-based analysis of decision making under risk in obsessive-compulsive and hoarding disorders. *Journal of Psychiatric Research*, *90*, 126–132. http://doi.org/10.1016/j.jpsychires.2017.02.017

Atkinson, A. C., & Donev, A. N. (1992). Optimum Experimental Designs. *El Observador De Estrellas Dobles*, 344.

Bahg, G., Sederberg, P. B., Myung, J. I., Li, X., Pitt, M. A., Lu, Z.-L., & Turner, B. M. (2018). Real-time Adaptive Design Optimization in Functional MRI Experiments. Manuscript under review.

Bickel, W. K. (2015). Discounting of delayed rewards as an endophenotype. *Biological Psychiatry*, *77*(10), 846–847. http://doi.org/10.1016/j.biopsych.2015.03.003

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, *52*(3), 233–254.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Comput*, *22*(4), 887–905.

Cohn, D., Atlas, Les, & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, *15*(2), 201–221. http://doi.org/10.1007/BF00993277

Collins, F. S., & Varmus, H. (2015). A New Initiative on Precision Medicine. *The New England Journal of Medicine*, *372*(9), 793–795.

DiMattina, C., & Zhang, K. (2011). Active Data Collection for Efficient Estimation and Comparison of Nonlinear Neural Models. *Neural Computation*, *23*(9), 2242–2288.

Enkavi, A. Z., Eisenberg, I., Bissett, P., Mazza, G. L., & Poldrack, R. A. (2018). A large-scale analysis of test-retest reliabilities of self-regulation measures. *PsyArxiv*.

Friedman, A. A., Letai, A., Fisher, D. E., & Flaherty, K. T. (2015). Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, *15*(12), 747–756. http://doi.org/10.1038/nrc4015

Green, L., & Myerson, J. (2004). A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychological Bulletin*, *130*(5), 769–792. http://doi.org/10.1037/0033-2909.130.5.769

Gu, H., Kim, W., Hou, F., Lesmes, L. A., Pitt, M. A., Lu, Z.-L., & Myung, J. I. (2016). A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function. *Journal of Vision*, *16*(6), 15–15. http://doi.org/10.1167/16.6.15

Harrison, J., & McKay, R. (2012). Delay Discounting Rates are Temporally Stable in an Equivalent Present Value Procedure Using Theoretical and Area under the Curve Analyses. *The Psychological Record*, *62*(2), 307–320. http://doi.org/10.1007/BF03395804

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *103*(3), 1–21. http://doi.org/10.3758/s13428-017-0935-1

Heerey, E. A., Robinson, B. M., McMahon, R. P., & Gold, J. M. (2007). Delay discounting in schizophrenia. *Cognitive Neuropsychiatry*, *12*(3), 213–221. http://doi.org/10.1080/13546800601005900

Hou, F., Lesmes, L. A., Kim, W., Gu, H., Pitt, M. A., Myung, J. I., & Lu, Z.-L. (2016). Evaluating the performance of the quick CSF method in detecting contrast sensitivity function changes. *Journal of Vision*, *16*(6), 18–18. http://doi.org/10.1167/16.6.18

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry*, *171*(4), 395–397. http://doi.org/10.1176/appi.ajp.2014.14020138

Kreutz, C., & Timmer, J. (2009). Systems biology: experimental design. *The FEBS Journal*, *276*(4), 923–942. http://doi.org/10.1111/j.1742-4658.2008.06843.x

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept Psychophys*, *63*(8), 1279–1292. http://doi.org/10.3758/BF03194543

Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Res*, *46*(19), 3160–3176. http://doi.org/10.1016/j.visres.2006.04.022

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural Representation of Subjective Value Under Risk and Ambiguity. *Journal of Neurophysiology*, *103*(2), 1036–1047. http://doi.org/10.1152/jn.00853.2009

Lewi, J., Butera, R., & Paninski, L. (2009). Sequential Optimal Design of Neurophysiology Experiments. *Neural Computation*, *21*(3), 619–687.

Matusiewicz, A. K., Carter, A. E., Landes, R. D., & Yi, R. (2013). Statistical equivalence and test–retest reliability of delay and probability discounting using real and hypothetical rewards. *Behavioural Processes*, *100*, 116–122.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. *Commons, ML.; Mazur, JE.; Nevin, JA*, 55–73.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. http://doi.org/10.1016/j.tics.2011.11.018

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518. http://doi.org/10.1037/a0016104

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3-4), 53–67. http://doi.org/10.1016/j.jmp.2013.05.005

Sandry, J., Genova, H. M., Dobryakova, E., DeLuca, J., & Wylie, G. (2014). Subjective Cognitive Fatigue in Multiple Sclerosis Depends on Task Length. *Frontiers in Neurology*, *5*(1), 24. http://doi.org/10.3389/fneur.2014.00214

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92. http://doi.org/10.1016/j.conb.2013.12.007

Wathen, J. K., & Thall, P. F. (2008). Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine*, *27*(27), 5586–5604. http://doi.org/10.1002/sim.3381

Weatherly, J. N., & Derenne, A. (2013). Testing the Reliability of Paper-Pencil Versions of the fill-in-the-blank and Multiple-Choice Methods of Measuring Probability Discounting for Seven Different Outcomes. *The Psychological Record*, *63*(4), 835–862. http://doi.org/10.11133/j.tpr.2013.63.4.009

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108.