Published in final edited form as:
Exp Clin Psychopharmacol. 2013 February ; 21(1): 66–73. doi:10.1037/a0030843.

Smoking and the bandit: A preliminary study of smoker and non-smoker differences in exploratory behavior measured with a multi-armed bandit task

Merideth A. Addicott, Ph.D.[1,2], **John M. Pearson, Ph.D.**[3,4], **Jessica Wilson, B.S.**[4,5], **Michael L. Platt, Ph.D.**[3,4,6], and **F. Joseph McClernon, Ph.D.**[1,2,7]

[1]Department of Psychiatry and Behavioral Research, Duke University and Duke University School of Medicine, Durham NC USA 27710

[2]Duke-UNC Brain Imaging and Analysis Center, Duke University and Duke University School of Medicine, Durham NC USA 27710

[3]Department of Neurobiology, Duke University and Duke University School of Medicine, Durham NC USA 27710

[4]Center for Cognitive Neuroscience, Duke University and Duke University School of Medicine, Durham NC USA 27710

[5]Department of Psychology and Neuroscience, Duke University and Duke University School of Medicine, Durham NC USA 27710

[6]Department of Evolutionary Anthropology, Duke University and Duke University School of Medicine, Durham NC USA 27710

[7]Durham Veterans Affairs Medical Center, Duke University and Duke University School of Medicine, Durham NC USA 27710

Abstract

Advantageous decision-making is an adaptive trade-off between exploring alternatives and exploiting the most rewarding option. This trade-off may be related to maladaptive decision-making associated with nicotine dependence; however, explore/exploit behavior has not been previously investigated in the context of addiction. The explore/exploit trade-off is captured by the multi-armed bandit task, in which different arms of a slot machine are chosen to discover the relative payoffs. The goal of this study was to preliminarily investigate whether smokers differ from non-smokers in their degree of exploratory behavior. Smokers (n = 18) and non-smokers (n = 17) completed a six-armed bandit task as well as self-report measures of behavior and personality traits. Smokers were found to exhibit less exploratory behavior (i.e. made fewer switches between slot machine arms) than non-smokers within the first 300 trials of the bandit task. The overall proportion of exploratory choices negatively correlated with self-reported measures of delay

Corresponding Author: Merideth Addicott, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Box 3527, Lakeview Pavilion E Suite 300, Durham NC 27705, merideth.addicott@duke.edu, fax: 919-681-0016.

The authors Addicott, Pearson, Wilson, Platt, and McClernon have no conflicts of interest to disclose.

All authors contributed in a significant way to the manuscript and have read and approved the final manuscript.

aversion and nonplanning impulsivity. These preliminary results suggest that smokers make fewer initial exploratory choices on the bandit task. The bandit task is a promising measure that could provide valuable insights into how nicotine use and dependence is associated with explore/exploit decision-making.

Keywords

exploration; exploitation; smoking; tobacco; multi-armed bandit task

Substance dependence is often considered the result of maladaptive decision-making. Consistent with this, nicotine dependent individuals exhibit a preference for immediate gains, more impulsivity (i.e., prefer smaller, sooner over larger, later rewards), and less sensitivity to consequences compared to healthy controls on tasks measuring outcome-contingent decision-making (Bickel, Odum, & Madden, 1999; Fishbein et al., 2005; Grant, Contoreggi, & London, 2000; Lane, Cherek, Tcheremissine, Steinberg, & Sharon, 2007; Lejuez et al., 2003). These performance deficits help inform why substance abusers continue to use drugs in the face of increasingly negative consequences (Bechara et al., 2001). One dimension of decision-making, the explore/exploit trade-off, has not previously been examined in the context of substance dependence but may offer new insights on the behavioral correlates of tobacco and other addictions.

Many daily decisions that we make, from conducting a job search to buying groceries, involve a choice between exploiting a familiar option with a known reward (e.g., job satisfaction, brand preference) versus exploring other options, hoping to discover a more rewarding alternative (for review, see Cohen, McClure, & Yu, 2007). Over time, our needs may change or the value of the rewards we obtain may diminish, thus prompting a re-evaluation (i.e. exploration) of alternative options. There are advantages and disadvantages to both exploitation and exploration: exploitation may have the advantage of familiarity and a guarantee of some reward, but at the expense of information that could be gathered by exploration. There may be better rewards out there, but gathering this information (i.e. exploring) costs time and effort that may not be recuperated unless it results in the discovery of a more rewarding alternative. The dilemma is when to stay and exploit versus when to leave and explore in order to maximize long-term rewards. The balance between exploration and exploitation is a fundamental behavioral adaptation to a changing world (Cohen et al., 2007) that is pertinent to both animals foraging for food and to people hunting for jobs or breakfast cereal. Importantly, a tendency towards exploitation may promote the development and maintenance of habitual behaviors, including extreme habits such as substance dependence (for review, see Graybiel, 2008).

The multi-armed bandit problem represents a classic mathematical formulation of the explore/exploit tradeoff (for review, see Cohen et al., 2007; John C. Gittins, 1989; J. C. Gittins & Jones, 1974; Sutton & Barto, 1998). The bandit task, adapted for behavioral research, is analogous to a gambler concurrently playing *n* slot machines (also known as “one-armed bandits”). Here, there is a direct trade-off between exploiting a single arm for its expected payoff and exploring other arms for potentially larger rewards (Kaelbling, Littman,

& Moore, 1996). In the so-called “restless” version of the bandit problem, the expected value of each slot machine changes with time, so that none of the machines are always advantageous or disadvantageous; rather, ongoing re-evaluation of outcomes and behavioral flexibility are necessary in order to advantageously switch between exploration and exploitation to maximize long-term gain.

The explore/exploit trade-off is a critical component of reinforcement learning, which is the method by which agents learn in response to environmental outcomes. In a new environment, exploration is necessary because response outcomes are uncertain, but with experience, the subject can predict future outcomes with more confidence and therefore exploit the option with the greatest expected reward (Daw & Doya, 2006; Graybiel, 2008; Sutton & Barto, 1998). In most theories of reinforcement learning, changes in reward-seeking behavior are driven by prediction errors that signal the difference between the predicted and the actual reward. Little or no prediction error implies a stable environment in which exploitation is the optimal strategy, while larger errors may signal uncertainty or instability and trigger exploration (Cohen et al., 2007). Research has shown that phasic dopaminergic signaling in the mesocorticolimbic pathway correlates with prediction errors (Schultz, 1998), and most importantly, this dopaminergic signaling has been implicated in both goal-directed learning and the transition from goal-directed to habitual behavior that occurs during the development of drug addiction (Canales, 2005; Everitt & Robbins, 2005). In particular, nicotine has been shown to enhance phasic dopamine signaling in the dorsal striatum and nucleus accumbens (Rice & Cragg, 2004; Zhang & Sulzer, 2004), which may be a mechanism involved in the development of tobacco addiction. However, very little research has investigated the effects of nicotine dependence on reinforcement learning in human subjects, and the relationship between nicotine dependence and explore/exploit decision-making is unknown.

Recently, a small but growing body of research has accumulated on the cognitive, psychological and neuroanatomical mechanisms that underlie the explore/exploit trade-off using the bandit task (Biele, Erev, & Ert, 2009; Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006; Jepma, te Beek, Wagenmakers, van Gerven, & Nieuwenhuis, 2010; Pearson, Hayden, Raghavachari, & Platt, 2009; Steyvers, Lee, & Wagenmakers, 2009), and given its relevance to reinforcement learning, explore/exploit decision-making could provide new insights into addiction. Potentially, a tendency towards exploitation of known rewards may be a cause or consequence of habitual drug use, but to our knowledge, this facet of decision-making has not been investigated in the context of substance dependence. The present study compared explore/exploit performance between smokers and non-smokers on a 6-armed bandit task to investigate whether smoking behavior could be associated with the increased tendency to exploit known (yet diminishing) rewards at the expense of exploration for other (potentially larger) rewards. Our hypothesis is based on previous research that suggests smoking is characterized by automatic tobacco use (Piper et al., 2008), nicotine enhances dopaminergic signaling implicated in reward learning (Rice & Cragg, 2004; Zhang & Sulzer, 2004), and smokers have impaired performance on decision-making tasks that require flexible cognitive control (Businelle et al., 2009; Chiu, Lohrenz, & Montague, 2008; Lejuez et al., 2003; Nesic, Rusted, Duka, & Jackson, 2011; Xiao et al., 2008). Secondly, we hypothesized that the extent of exploratory behavior would be negatively associated with smoking dependence motives such as automaticity and loss of control. Exploratory correlations between bandit performance and self-reported traits such as impulsivity and delay aversion were also conducted to investigate individual differences in explore/exploit decision-making.

Method

Participants

Thirty-seven smokers and non-smokers (21 women), aged 20–55 years (mean ± S.D.: 34 ± 10 years), participated in this study. Potential participants were recruited from the Raleigh-Durham-Chapel Hill area via internet advertisements and by word-of-mouth. Inclusion criteria were general good health and exclusion criteria included reports of significant health problems (e.g., hypertension), current use of smokeless tobacco or nicotine replacement therapy, current use of psychoactive medication, a positive breath alcohol concentration and a positive urine drug screen for illicit drugs on the study day. Smokers had to report smoking ≥ 10 cigarettes per day for ≥ two years, and produce an afternoon expired-air carbon monoxide (CO) level > 10 ppm (in order to establish smoking status). Non-smokers had to report smoking less than 50 cigarettes in their lifetime, report not smoking in the last 6 months, and produce an afternoon CO level ≤ 5 ppm. This experiment was approved by the Duke University Institutional Review Board. Informed consent was obtained from all participants.

Bandit Task

This version of the “restless bandit” task was adapted from previous studies (Berry & Fristedt, 1985; Daw et al., 2006; John C. Gittins, 1989; J. C. Gittins & Jones, 1974; Jepma et al., 2010; Pearson et al., 2009; Whittle, 1988). On each trial, six slot machines were depicted on a computer screen and participants selected one to play by pressing a corresponding number on the keypad. Following the selection, the number of points awarded was displayed on the screen for 500 msec. The number of points paid off by each slot machine (i.e., target) gradually changed from trial to trial, independently of other slot machine targets. The target values for our version of the bandit task were calculated as follows: targets began with an initial point-value of 50 on the first trial and subsequent values were drawn from a Gaussian distribution with a standard deviation (σ) = 2.8 around a moving mean and rounded to the nearest integer. Target values were randomly adjusted according to a biased random walk,

$$r_{i,t+1}=\lambda(r_{i,t}-\theta)+\theta+\eta \quad (1)$$

where $r_{i,t}$ is the reward value of the *i*th target on trial *t*, θ is the asymptotic mean reward value (equal to 50), and λ is a central tendency parameter that represents the tendency of *r* to drift back toward θ. η is a Gaussian random variable with mean zero and standard deviation σ. We used parameter values of λ = 0.015 and σ = 2.8, since this yielded payouts variable enough to encourage exploration and a low likelihood that a single target would remain most profitable for the entire task. The number of points awarded by target *i* on trial *t* was allowed to range between 0 and 100 (the resulting range was −4 to 105). A single version of the task was administered to all participants, and all participants received the same pattern of target point values. The target point values for the first 100 trials of the task are shown in Figure 1. Participants were told the goal of this task was to earn as many points as possible. They could earn up to an additional $10 based on the ratio of the number of points they earned to the total number of points possible. All 1500 trials were presented in a single block and the task lasted approximately 25 minutes. The bandit task was programmed in Matlab (MathWorks, Inc. Natick, MA) using the Psychophysics Toolbox (Brainard, 1997).

Procedure

Participants completed questionnaires and decision-making tasks as part of a larger battery of self-report and computer-based measures. All participants completed a smoking history survey which contained questions regarding lifetime and recent tobacco use in addition to the Barratt Impulsiveness Scale (BIS-11; Patton, Stanford, & Barratt, 1995), which measures trait impulsivity on three subscales: attention, motor, and nonplanning, and a Delay Questionnaire (DQ; Clare, Helps, & Sonuga-Barke, 2010). The DQ is a five-point Likert Scale (1 = strongly disagree to 5 = strongly agree) that assesses an individual’s delay discounting and aversion (e.g., “I often give up on things that I cannot have immediately” and “I hate waiting for things”, respectively). All participants were administered the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988) before and after the testing session to test for changes in mood. Smokers completed the Fagerström Test for Nicotine Dependence (FTND; Heatherton, Kozlowski, Frecker, & Fagerstrom, 1991) and the Wisconsin Inventory for Smoking Motives (WISDM; Piper et al., 2008). Four subscales of the WISDM (automaticity, craving, loss of control, and tolerance) were included in our analyses. These subscales constitute a distinct common factor that has been shown to predict dependence criteria (Piper et al., 2008). Smokers also completed the Shiffman-Jarvik nicotine withdrawal scale (Shiffman & Jarvik, 1976) before and after the session to test for changes in symptoms of nicotine withdrawal. Smokers were allowed to smoke prior to the study and were offered a smoking break half-way though the testing session.

Behavioral Modeling of the Bandit Task

Choices made in the bandit task were classified as exploratory or exploitative according to model-based account of participants’ individual choice behavior (previously described in Daw et al., 2006; Jepma & Nieuwenhuis, 2011; Pearson et al., 2009). Four reinforcement-learning models, which each calculate the estimated target pay-offs differently, were initially fit to the participants’ data and compared using the Bayesian Information Criterion in order to determine the best fitting model. The results from the best fitting model are reported here; see Supplementary Information for a description of the other three models. On each trial, selection of the target with highest estimated action value in the model was coded as exploitative, all other choices as exploratory.

As in previous studies, the best fitting model valued the slot machine options according to a softmax rule and Kalman filter (Anderson & Moore, 1979; Daw et al., 2006). The softmax rule chooses machine options probabilistically based on their expected point values:

$$P(i|\beta,Q_i)=\frac{{}_{e}\beta Q_i}{\sum_{je}\beta Q_j} \quad (2)$$

where $P(i\backslash\beta,Q_i)$ is the probability of choosing option i, and β is a so-called “greediness” parameter, with $\beta = \infty$ corresponding to perfectly greedy (uniformly exploitative) choice. In other words, the logarithmic value of the explore/exploit ratio is proportionate to the softmax parameter, β, multiplied by the difference in action value between the explore options and the exploit option. That is, β is the logarithmic value of the explore/exploit ratio per unit of action value. A larger value of β will lead to a higher percentage of exploit trials. See Figure S1 in the Supplementary Information for a graph of representative β values. The Kalman filter (Anderson & Moore, 1979) is a Bayes-optimal filtering process used to infer the values of unseen targets. Here, the posterior probability estimates for the target values took the form of normal distributions with mean and variance for all targets updated each trial according to a drift rule:

$$\mu_i \leftarrow (1-\zeta)\mu_i+\zeta\theta \quad (3)$$

$$\sigma_i^2 \leftarrow (1-\zeta)^2\sigma_i^2+D^2 \quad (4)$$

where μ_i and σ_i are the mean and standard deviation of the previous estimate of each target’s value, ζ is a central tendency of options to drift toward an asymptotic value, θ, and D reflects the growing uncertainty in an unchosen target’s value over time due to drift. Due to random changes in the target values over time, uncertainties of unseen targets grow each trial, and mean values decay slowly back toward a subject-specific asymptotic value. Note that participants did not know the true value of the central tendency, so $\zeta \neq \lambda$ in general. In addition, for the chosen target, we calculated learning parameters as follows:

$$\delta_i=r-\mu_i \quad (5)$$

$$\alpha_i=\frac{\sigma_i^2}{\sigma_i^2+\sigma_0^2} \quad (6)$$

With r the outcome on the current trial, μ_i the mean of the chosen target, and σ_0 the prior standard deviation of the target. As usual, δ is the reward prediction error and α the learning rate, used to update the chosen target according to

$$\mu_i \leftarrow \mu_i+\alpha_i\delta_i \quad (7)$$

$$\sigma_i^2 \leftarrow (1-\alpha_i)\sigma_i^2 \quad (8)$$

As a result, each trial yields a single σ and α, along with vectors μ and σ.

Data Analysis

The bandit task consisted of 1500 trials that were divided into 5 blocks of 300 trials each for analysis. The main dependent variable was the percentage of trials per block coded as “exploratory.” In addition, we explored two other trial-to-trial variables to investigate qualitative differences in bandit performance between groups. The learning rate is the rate at which values of the targets are updated (i.e., the sensitivity to the most recent pay-off of each machine arm), and is determined optimally via the Kalman filter. Learning rates are higher in more variable environments, which should positively correlate with exploration. Alternately, lower learning rates indicate behavior appropriate to a stable environment, which would favor exploitation. The range in action value is the difference between the estimated value of the chosen option and the maximum estimated pay-out during exploratory trials; large values of this measure indicate a large reward was forfeited in order to choose a suboptimal option. The reward ratio (total points/maximum possible points) was also calculated for each participant. The distribution of exploratory choices and reward ratios were normalized using a logit transformation for analysis (Gart & Zweifel, 1967).

Demographic differences between groups were analyzed with independent-sample t-tests and chi-square tests. Mood questionnaires and bandit task data were analyzed using mixed-model analyses of covariance (ANCOVAs), with time or block of trials as the within-subject factor, smoking status as the between-subject factor, and education level as the covariate of no interest. The Greenhouse-Geisser correction was reported when Mauchley’s test of sphericity was violated. Significant main effects were followed up with multivariate and univariate ANCOVAs. Effect sizes are eta-squared. All analyses were performed with SPSS (Chicago, IL) with alpha set to .05. Data from two smokers who fell asleep during the bandit task were excluded; thus, the final subject sample was 18 smokers and 17 non-smokers.

Potential differences between men and women were preliminarily tested using an ANCOVA with sex and smoking status as between-subject factors, block of trials as the within-subject factor, and education level as the covariate of no interest. There were no main effects of sex or interactions with sex; therefore, data from men and women were combined for further analyses.

Results

Participant characteristics

Participant characteristics are shown in Table 1. There were no differences between the two groups in age, sex, or racial distribution; however, smokers had fewer years of education than non-smokers, $t = 4.1$, $p < .001$. The relationships between education, self-report data, and performance outcomes were explored with Pearson correlations. Years of education correlated with scores on the BIS, $r = -.39$, $p = .020$, and the FTND, $r = -.74$, $p < .001$, in addition to the percentage of exploratory choices, $r = -.43$, $p = .010$, and the reward ratio, $r = .50$, $p = .002$, from the bandit task. Therefore, all subsequent analyses controlled for education level. There were no between-group differences in BIS or DQ scores. There were no significant changes in nicotine withdrawal symptoms among smokers pre- to post-session, nor were there changes in positive or negative affect pre- to post-session among both groups. Eight smokers chose to smoke during the cigarette break, and there were no differences between smokers that smoked and those that did not on pre- to post-session nicotine withdrawal symptoms, positive and negative affect, or task outcomes.

Bandit task

Averaged across all trials of the bandit task, the proportion of choices coded as exploratory was 23% and the rest of the choices were coded as exploitative. Participants earned approximately $8.60 for their performance and there was no difference in the amount earned between smokers and nonsmokers. The estimated marginal means of the percentage of exploratory choices (controlling for education level) made by non-smokers and smokers across the five blocks of trials are shown in Figure 2. The initial ANCOVA performed on the logit transformed data revealed a significant block x smoking status interaction, $F(4,128) = 5.6$, $p = .003$, $\eta^2 = .133$, and follow-up tests showed that nonsmokers had a greater decrease in exploratory choices from the first to the fifth block of trials than smokers, $F(1,34) = 12.6$, $p = .001$. The initial ANCOVA also revealed a trend for a between-subject effect, $F(1,32) = 2.9$, $p = .097$, $\eta^2 = .064$, and follow-up tests found that nonsmokers had a greater percentage of exploratory choices than smokers in block 1, $F(1,34) = 11.9$, $p = .002$, and had a trend for greater exploratory choices in block 2 $F(1,34) = 3.9$, $p = .056$.

A post hoc investigation into potential qualitative differences in exploratory behavior between smokers and nonsmokers was conducted on the first block of 300 trials. According to the model, smokers had a higher learning rate (estimated marginal mean ± S.E.: 0.73 ± 0.11) than non-smokers (0.36 ± 0.11), $F(1,34) = 4.7$, $p < .038$, but there was no statistical difference in the measure of the range in action value.

Analyses on the within- and between-group differences in the number of points earned on the bandit task found no effect of block or smoking status on the reward ratio. In general, participants earned an average of 88% of the total possible points and there was a negative partial correlation (controlling for education) between the reward ratio and the percentage of exploratory choices, $r = -.50$, $p = 0.003$, averaged across all 1500 trials.

Correlations were performed to explore potential relations between bandit performance and self-reported behaviors. Partial correlations (controlling for education) showed a negative relationship between the percentage of exploratory choices across 1500 trials and the BIS nonplanning subscale, $r = -.38$, $p = .029$, and the DQ, $r = -.38$, $p = .027$, for all participants, and the WISDM subscales loss of control, $r = -.61$, $p = .010$, and tolerance, $r = -.57$, $p = .018$, for smoking participants. In other words, as individuals' ratings for nonplanning-type impulsive behaviors and delay aversion/discounting increased, there was a corresponding increase in the percentage of exploitative choices made on the bandit task. Similarly, as smokers' ratings increased for the smoking dependence motives concerning a loss of volitional control over smoking behavior and nicotine tolerance, there was a corresponding increase in the percentage of their exploitative choices. There were no significant relationships between the self-report scores and the reward ratio.

Discussion

The present study is the first to preliminarily investigate differences in explore/exploit behavior among smokers and non-smokers using a multi-armed bandit task. In support of our hypothesis, we found that smokers made fewer exploratory choices within the first block consisting of 300 trials. The number of exploratory choices made by smokers remained the same while exploratory choices made by non-smokers decreased from the first to the fifth block of trials. Both groups earned a similar number of points based on their selections. This novel result suggests that behavioral strategies guiding explore/exploit decision-making in the bandit task may be related to smoking status. It is possible that either a combination of factors contributing to an increased risk for tobacco addiction (e.g., a behavioral tendency to exploit known rewards) or the effects of habitual tobacco use (e.g., the chronic effects of nicotine on dopaminergic reward signaling) contribute to explore/exploit choices. This study identifies and provides initial support for some potential correlates of explore/exploit behavior, including impulsivity, delay aversion, smoking motives and education level.

The majority of choices made in the bandit task were exploitative, and these choices positively related to the number of rewards earned on the task. This underscores the advantage of exploiting a slot machine arm with the maximum number of points available. However, when the value of the selected machine arm begins to decline, exploration is necessary to find the subsequent best option. Non-smokers made more exploratory choices in the first block of the task, and the significant decrease in this behavior from the first to the fifth block suggests an effect of learning on their performance. Perhaps they explored more than necessary to maximize their gain in the first block, and the number of exploratory choices decreased as their familiarity with response-outcomes improved. In comparison, the percentage of exploratory choices made by smokers remained constant throughout the task. This appears to be congruent with previous reports of diminished cognitive flexibility or response adaptation in other substance abusing subjects (Lane et al., 2007; Ornstein et al., 2000; Verdejo-Garcia, Lopez-Torrecillas, de Arcos, & Perez-Garcia, 2005). However, it is unlikely that explore/exploit behavior on the bandit task simply reflects cognitive flexibility. A post hoc analysis of the range in action value indicated that smokers’ exploratory choices were not qualitatively different from nonsmokers’ choices, nor did they earn fewer points, which suggests that smokers were not exploring less strategically. Interestingly, the learning rate, the rate that action value is updated based on new information, was higher among smokers in the first block. This suggests that smokers were more sensitive to the most recent pay-off of each machine arm at the start of the task. That is, smokers reacted to random noise in payouts over early trials more than non-smokers. Based on these results, smokers did not have worse performance on the bandit task compared to nonsmokers, and additional studies are needed to better understand smoker/non-smoker differences in explore/exploit strategy.

Intuitively, too much exploration in the bandit task may be considered risky or novelty-seeking behavior, while too much exploitation may be considered risk-averse (Steyvers et al., 2009). Steyvers and colleagues compared bandit task performance with self-reported measures of personality traits associated with risk-seeking behavior, but found little to no relationship (Steyvers et al., 2009). However, among the participants in this study, the proportion of exploratory choices decreased as self-reported delay aversion and nonplanning impulsivity traits increased. A potential interpretation for this relationship is that participants who report disliking and discounting delayed rewards may exploit more, thus they are analogously sacrificing larger net gains in favor of immediate rewards. A tendency against planning for future events (e.g., low scores on the BIS nonplanning scale such as “I plan for job security” and “I plan tasks carefully”) may also contribute to exploitative behavior. The absence of group differences on these measures suggests that impulsivity and delay aversion do not account for smoker/nonsmoker differences in bandit performance. It appears that exploratory choices on the bandit task are not representative of risky or impulsive decision-making, but may reflect a degree of deliberation or planning for future outcomes during goal-directed behavior. Ultimately, the optimal strategy on the bandit task is a well-timed trade-off between exploration and exploitation, and more research is needed to understand how personality traits are associated with bandit task performance.

The bandit task may be particularly appropriate for the study of tobacco addiction since nicotine affects several neuromodulators implicated in explore/exploit behavior and reward prediction, including acetylcholine (ACh), norepinephrine (NE) and dopamine (DA) (Summers & Giacobini, 1995). Acetylcholine and NE have been proposed to play an important role in the neural signaling of expected and unexpected uncertainty (Yu & Dayan, 2005). Expected uncertainty may promote exploitation when the points from a chosen arm of the bandit task vary within an expected degree across trials. Alternatively, unexpected uncertainty may promote exploration when the points from a chosen arm begin to decrease more than expected. Dopaminergic signaling has been correlated with prediction errors (Schultz, 1998), which help determine if changes in outcome are expected or unexpected. How acute and chronic nicotine exposure affects these neuromodulators in relation to explore/exploit behavior is unknown, but future studies with the bandit task and other models of reinforcement learning may offer new insights into the development and maintenance of addictive behaviors.

The development of substance dependence may also be related to the balance between explore and exploit behavior. Adaptive decision-making can be described as a carefully timed trade-off between exploring options to evaluate their outcomes and exploiting options with the most desirable outcomes. In contrast to this goal-directed behavior, habitual behaviors are fixed and repetitive actions that are insensitive to changes in outcome contingency. Habitual learning is thought to play an important role in tobacco addiction, which is primarily characterized by automatic, habitual tobacco use and a loss of volitional control over this habit, in addition to craving and tolerance (Piper et al., 2008). Within the group of smokers there were negative correlations between exploratory choices and self-reported loss of control (e.g., strongly endorsing “My smoking is out of control”) and tolerance (e.g., strongly endorsing “I usually want to smoke right after I wake up”) to their smoking habit, as measured by the WISDM. Hypothetically a behavioral preference for exploitation at the expense of exploration may promote the development of habitual drug use. Alternatively, the smoker/non-smoker differences could be a result of chronic nicotine exposure. Our results provide initial support for a relationship between exploitative choices on the bandit task and smoking motives, albeit reproducing these results with a larger subject sample would strengthen this interpretation.

In our sample of community volunteers, smokers had fewer years of education than non-smokers, which agrees with previous studies showing smoking to be more prevalent among low educated men and women (Giskes et al., 2005; Winkleby, Fortmann, & Barrett, 1990). Groups were matched by age, sex and race, yet differences in education level were significant. This limitation was addressed by statistically controlling for education in our data analyses. The inclusion of this covariate did not affect the direction or significance of our results. However, there remains a possibility that education level is an important moderator of bandit performance, either independently or in relation to other factors. Higher levels of education correlated with less impulsivity on the BIS, more exploitative choices and a higher reward ratio earned on the bandit task. Hypothetically, less impulsive individuals with more education are better able to exploit the slot machine arm with the maximum point values. To our knowledge, this is the first study to show that education level predicts performance on the bandit task. Additional studies are needed to confirm a difference in bandit performance between smokers and non-smokers, and these studies can avoid the limitation addressed here by balancing education levels between groups.

A limitation of the present study is that the bandit task consisted of 1500 trials, as we anticipated that differences might develop between groups as the task progressed. Unexpectedly, group differences were most significant in the first 300 trials. Previous studies using bandit problems have administered tasks consisting of 300 trials (Daw et al., 2006; Jepma et al., 2010; Steyvers et al., 2009). While the convergence of behavior was an interesting result, the large number of trials resulted in fatigue among two participants and this led to the removal of their data. A second limitation is that in-depth mental health and substance use histories were not obtained in this preliminary study of smoker/non-smoker differences. Smokers have a high rate of co-morbid psychiatric illnesses and other substance dependence (Kalman, Morissette, & George, 2005), which could have influenced their behavior. Future studies could improve the interpretation of explore/exploit behavior by investigating whether these preliminary results extend to individuals who are dependent on drugs other than nicotine.

In summary, the explore/exploit trade-off is an evolutionary adaptation to changes in environmental reinforcement contingencies and studies have shown that the ability to act according to dynamic changes in reward estimation is preserved across species (e.g., Keasar, Rashkovich, Cohen, & Shmida, 2002; Pearson et al., 2009; Racey, Young, Garlick, Pham, & Blaisdell, 2011). The bandit task, with its roots in reinforcement learning, is an excellent tool to investigate the balance between exploratory and exploitative behavior. To date, the bandit task has been used to study decision-making behavior among healthy control subjects and in animal models, and to our knowledge, this is the first study to preliminarily investigate explore/exploit decision-making among nicotine-dependent subjects. Our results suggest that the extent of exploratory behavior may be related to smoking status and other individual characteristics such as delay aversion, impulsivity, smoking motives, and education level.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Katherine Hosford for her help with data collection and the anonymous reviewers for providing useful feedback on earlier versions of this manuscript.

References

Anderson, BDO.; Moore, JB. Optimal Filtering. Englewood Cliffs; 1979.

Bechara A, Dolan S, Denburg N, Hindes A, Anderson SW, Nathan PE. Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. Neuropsychologia. 2001; 39:376–389. [PubMed: 11164876]

Berry, DA.; Fristedt, B. Bandit problems: sequential allocation of experiments. New York: Chapman and Hall, London; 1985.

Bickel WK, Odum AL, Madden GJ. Impulsivity and cigarette smoking: Delay discounting in current, never, and ex-smokers. Psychopharmacology. 1999; 146:447–454. 91460447.213 [pii]. [PubMed: 10550495]

Biele G, Erev I, Ert E. Learning, risk attitude and hot stoves in restless bandit problems. Journal of Mathematical Psychology. 2009; 53:155–167.10.1016/j.jmp.2008.05.006

Brainard DH. The psychophysics toolbox. Spatial Vision. 1997; 10:433–436. [PubMed: 9176952]

Businelle MS, Kendzor DE, Rash CJ, Patterson SM, Coffey SF, Copeland AL. Heavy smokers perform more poorly than nonsmokers on a simulated task of gambling. Substance Use and Misuse. 2009; 44:905–914. [PubMed: 19415570]

Canales JJ. Stimulant-induced adaptations in neostriatal matrix and striosome systems: Transiting from instrumental responding to habitual behavior in drug addiction. Neurobiology of Learning and Memory. 2005; 83:93–103.10.1016/j.nlm.2004.10.006 [PubMed: 15721792]

Chiu PH, Lohrenz TM, Montague PR. Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. Nature Neuroscience. 2008; 11:514–520. nn2067 [pii]. 10.1038/nn2067

Clare S, Helps S, Sonuga-Barke EJ. The quick delay questionnaire: a measure of delay aversion and discounting in adults. [Validation Studies]. Attention Deficit Hyperactivity Disorders. 2010; 2:43–48.10.1007/s12402-010-0020-4 [PubMed: 21432589]

Cohen JD, McClure SM, Yu AJ. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B. 2007; 362:933–942.

Daw ND, Doya K. The computational neurobiology of learning and reward. Current Opinion in Neurobiology. 2006; 16:199–204.10.1016/j.conb.2006.03.006 [PubMed: 16563737]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441:876–879. nature04766 [pii]. 10.1038/nature04766 [PubMed: 16778890]

Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience. 2005; 8:1481–1489.10.1038/Nn1579

Fishbein D, Hyde C, Eldreth D, London ED, Matochik J, Ernst M, Kimes A. Cognitive performance and autonomic reactivity in abstinent drug abusers and nonusers. Experimental and Clinical Psychopharmacology. 2005; 13:25–40. 2005-01534-004 [pii]. 10.1037/1064-1297.13.1.25 [PubMed: 15727501]

Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with applications to quantal bioassay. Biometrika. 1967; 54:181–187. [PubMed: 6049534]

Giskes K, Kunst AE, Benach J, Borrell C, Costa G, Dahl E, Mackenbach JP. Trends in smoking behaviour between 1985 and 2000 in nine European countries by education. Journal of Epidemiology and Community Health. 2005; 59:395–401.10.1136/jech.2004.025684 [PubMed: 15831689]

Gittins, JC. Multi-armed bandit allocation indices. Chichester; New York: Wiley; 1989.

Gittins, JC.; Jones, DH. A dynamic allocation index for the sequential design of experiments. In: Gani, JM., editor. Progress in Statistics. Amsterdam: North Holland; 1974.

Grant S, Contoreggi C, London ED. Drug abusers show impaired performance in a laboratory test of decision making. Neuropsychologia. 2000; 38:1180–1187. [PubMed: 10838152]

Graybiel AM. Habits, rituals, and the evaluative brain. Annual Review of Neuroscience. 2008; 31:359–387.10.1146/annurev.neuro.29.051605.112851

Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. British Journal of Addiction. 1991; 86:1119–1127. [PubMed: 1932883]

Jepma M, Nieuwenhuis S. Pupil diameter predicts changes in the exploration-exploitation tradeoff: Evidence for the adaptive gain theory. Journal of Cognitive Neuroscience. 2011; 23:1587–1596.10.1162/jocn.2010.21548 [PubMed: 20666595]

Jepma M, te Beek ET, Wagenmakers EJ, van Gerven JMA, Nieuwenhuis S. The role of the noradrenergic system in the exploration-exploitation trade-off: a psychopharmacological study. Frontiers in Human Neuroscience. 2010; 4:1–13. [PubMed: 20204154]

Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. Journal of Artificial Intelligence Research. 1996; 4:237–285.

Kalman D, Morissette SB, George TP. Co-morbidity of smoking in patients with psychiatric and substance use disorders. American Journal on Addictions. 2005; 14:106–123.10.1080/10550490590924728 [PubMed: 16019961]

Keasar T, Rashkovich E, Cohen D, Shmida A. Bees in two-armed bandit situations: foraging choices and possible decision mechanisms. Behavioral Ecology. 2002; 13:757–765.

Lane SD, Cherek DR, Tcheremissine OV, Steinberg JL, Sharon JL. Response perseveration and adaptation in heavy marijuana-smoking adolescents. Addictive behaviors. 2007; 32:977–990. S0306-4603(06)00246-2 [pii]. 10.1016/j.addbeh.2006.07.007 [PubMed: 16930850]

Lejuez CW, Aklin WM, Jones HA, Richards JB, Strong DR, Kahler CW, Read JP. The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. Experimental and Clinical Psychopharmacology. 2003; 11:26–33. [PubMed: 12622341]

Nesic J, Rusted J, Duka T, Jackson A. Degree of dependence influences the effect of smoking on cognitive flexibility. Pharmacology Biochemistry and Behavior. 2011; 98:376–384.10.1016/j.pbb.2011.01.015

Ornstein TJ, Iddon JL, Baldacchino AM, Sahakian BJ, London M, Everitt BJ, Robbins TW. Profiles of cognitive dysfunction in chronic amphetamine and heroin abusers. Neuropsychopharmacology. 2000; 23:113–126. [PubMed: 10882838]

Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. Journal of Clinical Psychology. 1995; 51:768–774. [PubMed: 8778124]

Pearson JM, Hayden BY, Raghavachari S, Platt ML. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. Current Biology. 2009; 19:1532–1537. S0960-9822(09)01474-2 [pii]. 10.1016/j.cub.2009.07.048 [PubMed: 19733074]

Piper ME, Bolt DM, Kim SY, Japuntich SJ, Smith SS, Niederdeppe J, Baker TB. Refining the tobacco dependence phenotype using the Wisconsin Inventory of Smoking Dependence Motives. Journal of Abnormal Psychology. 2008; 117:747–761.10.1037/A0013298 [PubMed: 19025223]

Racey D, Young ME, Garlick D, Pham JN, Blaisdell AP. Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. Learning & Behavior. 2011; 39:245–258.10.3758/s13420-011-0025-7 [PubMed: 21380732]

Rice ME, Cragg SJ. Nicotine amplifies reward-related dopamine signals in striatum. Nature Neuroscience. 2004; 7:583–584.10.1038/Nn1244

Schultz W. Predictive reward signal of dopamine neurons. Journal of Neurophysiology. 1998; 80:1–27. [PubMed: 9658025]

Shiffman SM, Jarvik ME. Smoking withdrawal symptoms in two weeks of abstinence. Psychopharmacology. 1976; 50:35–39. [PubMed: 827760]

Steyvers M, Lee MD, Wagenmakers EJ. A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology. 2009; 53:168–179.10.1016/j.jmp.2008.11.002

Summers KL, Giacobini E. Effects of Local and Repeated Systemic Administration of (−)Nicotine on Extracellular Levels of Acetylcholine, Norepinephrine, Dopamine, and Serotonin in Rat Cortex. Neurochemical Research. 1995; 20:753–759. [PubMed: 7566373]

Sutton, RS.; Barto, AG. Reinforcement learning: an introduction. Cambridge, Mass: MIT Press; 1998.

Verdejo-Garcia AJ, Lopez-Torrecillas F, de Arcos FA, Perez-Garcia M. Differential effects of MDMA, cocaine, and cannabis use severity on distinctive components of the executive functions in polysubstance users: A multiple regression analysis. Addictive Behaviors. 2005; 30:89–101.10.1016/j.addbeh.2004.04.015 [PubMed: 15561451]

Watson D, Clark LA, Tellegen A. Development and Validation of Brief Measures of Positive and Negative Affect - the Panas Scales. Journal of Personality and Social Psychology. 1988; 54:1063–1070. [PubMed: 3397865]

Whittle P. Restless bandits: activity allocation in a changing world celebration of applied probability. Journal of Applied Probability. 1988; 25A:287–298.

Winkleby MA, Fortmann SP, Barrett DC. Social-class disparities in risk-factors for disease - 8-Year prevalence patterns by level of education. Preventive Medicine. 1990; 19:1–12. [PubMed: 2320553]

Xiao L, Bechara A, Cen S, Grenard JL, Stacy AW, Gallaher P, Anderson Johnson C. Affective decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in 10th-grade Chinese adolescent smokers. Nicotine and Tobacco Research. 2008; 10:1085–1097. 794503865 [pii]. 10.1080/14622200802097530 [PubMed: 18584472]

Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005; 46:681–692.10.1016/j.neuron.2005.04.026 [PubMed: 15944135]

Zhang H, Sulzer D. Frequency-dependent modulation of dopamine release by nicotine. Nature Neuroscience. 2004; 7:581–582.10.1038/Nn1243

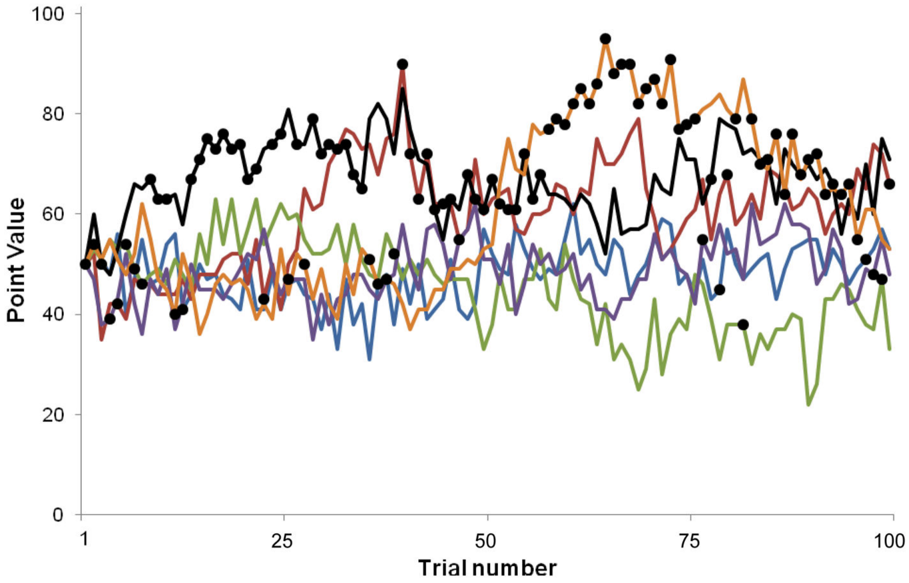

Figure 1.

Point values for the first 100 trials of the bandit task. Participants selected one of six slot machine arms to play per trial. The number of points paid off by each slot machine gradually changed from trial to trial, independently of other machines. The pattern of point values was determined according to a biased random walk (Equation 1). Each line represents the point values for a single slot machine. Dots represent the selections made by a representative subject.

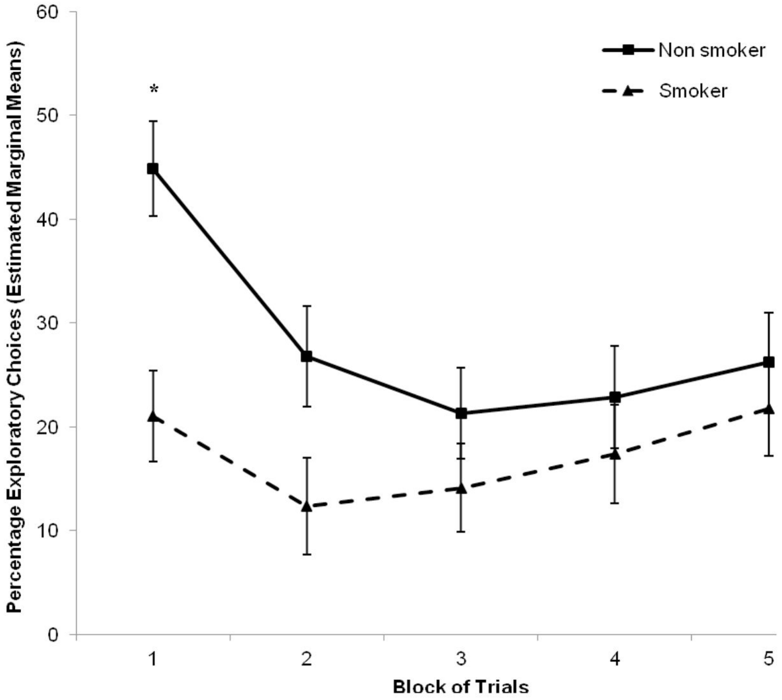

Figure 2.

The proportion of exploratory choices made on the bandit task by non-smokers and smokers. Shown are the estimated marginal means of the average percentage of exploratory choices per block of 300 trials (controlling for education). Smokers made significantly fewer exploratory choices in the first block compared to non-smokers (*$p < 0.05$). Error bars are S.E.M.

Table 1

Participant characteristics for smokers (*n* = 18) and non-smokers (*n* = 17); (Mean ± S.D.).

	Smokers	**Non-smokers**	***t* or X^2**	
Age	36 ± 8	32 ± 12		n.s.
Male/Female	8/10	7/10		n.s.
Race C/AA/H/A	6/9/1/2	6/10/1/0		n.s.
Years of Education	13 ± 2	15 ± 2	$t = 4.1$	$p < 0.001$
CO	14 ± 6	1 ± 1	$t = 7.9$	$p < 0.001$
Alcohol drinks/week	2 ± 4	2 ± 2		n.s.
Cigs/day	16 ± 8	--		
FTND	5.9 ± 2.3	--		
WISDM:				
automaticity	4.3 ± 1.7	--		
craving	4.6 ± 1.7	--		
loss of control	4.2 ± 1.3	--		
tolerance	4.6 ± 1.2	--		

C: Caucasian, AA: African American, H: Hispanic, A: Asian. CO: carbon monoxide. FTND: Fagerström Test for Nicotine Dependence. WISDM: Wisconsin Inventory of Smoking Dependence Motives. n.s.: not significant.