Experimental and Clinical Psychopharmacology
2008, Vol. 16, No. 6, 555–564

1064-1297/08/$12.00 DOI: 10.1037/a0014245

Development of an Automatic Response Mode to Improve the Clinical Utility of Sequential Risk-Taking Tasks

Timothy J. Pleskac
Michigan State University

Thomas S. Wallsten, Paula Wang, and C. W. Lejuez
University of Maryland–College Park

Sequential risk-taking tasks, especially the Balloon Analogue Risk Task (BART), have proven powerful and useful methods in studying and identifying real-world risk takers. A natural index in these tasks is the average number of risks the participant takes in a trial (e.g., pumps on the balloons), but this is difficult to estimate because some trials terminate early because of the consequences of those risks (e.g., when the desired number of balloon pumps exceeds the explosion point). The standard corrective strategy is to use an adjusted score that ignores such event-terminated trials. Although previous data supports the utility of this adjusted score, the authors show formally that it is biased. Therefore, the authors developed an automatic response procedure, in which respondents state at the beginning of each trial how many risks they wish to take and then observe the sequence of events unfold. A study comparing this new *automatic* and the original *manual* BART shows that the automatic procedure yields unbiased statistics whereas maintaining the BART's predictive validity of substance use. The authors also found that providing respondents with the expected-value-maximizing strategy and complete trial-by-trial feedback increased the number of risks they were willing to take during the BART. The authors interpret these results in terms of the potential utility of the automatic version including shorter administration time, unbiased behavioral measures, and minimizing motor involvement, which is important in neuroscientific investigations or with clinical populations with motor limitations.

Keywords: sequential choice, BART, risk taking, missing data

Laboratory-based gambling tasks are increasingly being used to study risk-taking behavior among both normal and clinical populations (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; Lane, Cherek, Pietras, & Tcheremissine, 2004; Lejuez et al., 2002; Paulus, Rogalsky, Simmons, Feinstein, & Stein, 2003; Rogers et al., 1999). A common attribute among these tasks is that they require participants to make repeated risky decisions often with real money at stake. This attribute gives rise to several advantages. One is that, unlike risk propensity scales (e.g., Eysenck, Pearson, Easting, & Allsopp, 1985; Zuckerman, 1994), these tasks yield measures derived from actual behavior and therefore do not require assumptions about accurate self-report and recall bias, which empirical evidence from cognitive psychology suggests are problematic for risk propensity scales (see, e.g., Bernstein, Whitttlsea, & Loftus, 2002; Hyman & Loftus, 1998; Tversky & Kohler, 1994). This advantage along with promising data from laboratory-based gambling tasks showing that they can identify real-world risk takers (e.g., Lejuez et al., 2003; Stout et al., 2005) has led some researchers to suggest that laboratory-based gambling tasks are among the best risk assessment measures available in the clinical literature (see Harrison, Young, Butow, Salkeld, & Solomon, 2005). Beyond an assessment of risk-taking propensity, laboratory-based gambling tasks also make it possible to examine risky decision making at both the cognitive (e.g., Busemeyer & Stout, 2002; Pleskac, 2008; Wallsten, Pleskac, & Lejuez, 2005; Yechiam, Busemeyer, Stout, & Bechara, 2005) and neural level (e.g., Bechara, Damasio, Damasio, & Lee, 1999; Fecteau et al., 2007; Leland & Paulus, 2005).

Despite these advantages, there are challenges with laboratory-based gambling tasks. A primary challenge is that the observed risk-taking behavior during the tasks is partly a function of the participant and partly [that of] the statistical environment of the tasks. To understand this statement, consider the sequential risk-taking paradigm (the focus of this article). This set includes Slovic's (1966) devil task (see also Hoffrage Weber, Hertwig, & Chase, 2003); the Balloon Analogue Risk Task (BART; Lejuez et al., 2002); and the Angling Risk Tasks (Pleskac, 2008). On any given trial of any of these tasks, participants must sequentially choose between a risky play option and a safe stop option. The

Timothy J. Pleskac, Department of Psychology, Michigan State University; Thomas S. Wallsten, Paula Wang, and Carl W. Lejuez, Department of Psychology, University of Maryland-College Park.

A National Institute of Health grant R01 DA 18647 and R21 DA14699 awarded to Carl Lejuez and a National Institute of Mental Health Research Service Award (MH019879) awarded to Indiana University supported this work. We thank Stephen G. West, Anthony J. Bishara, and Ryan Jessup for insightful comments on earlier versions of the paper.

Correspondence concerning this article should be addressed to Timothy J. Pleskac, Department of Psychology, Michigan State University, East Lansing, MI 48824. E-mail: tim.pleskac@gmail.com

statistical environment is such that one of two events can occur if respondents choose the risky play option. Either they (a) win a reward and get another choice or (b) they lose the gamble, forfeit their winnings for the trial and the trial ends. If, however, participants choose to stop playing then they collect their earnings for that trial and the trial ends. Clearly, how many risks a respondent would have taken on trials that end in a loss is unobserved. As a result, the risk propensity measure in these sequential tasks typically is calculated only on trials in which participants chose to stop themselves. The resulting *adjusted score* is the average number of play options taken on trials for which the participant chose to stop (Lejuez et al., 2002). This score, however, is biased as it rests on an incorrect implicit assumption that the trial was terminated independent from the behavior of the participant.

In this article, we formally and empirically establish the bias of the adjusted score and identify a methodological solution to eliminate the bias in the form of a new response mode for sequential risk-taking tasks. Next, we introduce the basic structure of the sequential risk-taking paradigm, focusing on the BART, in particular. Then we show formally how the interaction between the statistical environment of the task and respondents' risk taking behavior during the task interact to produce a biased measure of performance. Finally, using our theoretical understanding of behavior in this task (see Wallsten et al., 2005), we modify the standard BART and create a new version (automatic BART) that produces an unbiased measure of risk taking. Finally, we present an empirical study that reveals the magnitude of the bias as well as show that the automatic version has similar external validity as the standard BART.

The Bias of the Adjusted Score

To be concrete about the problem of the biased behavioral indices in sequential risk-taking tasks, we consider the specific case of the BART (Lejuez et al., 2002). The problem and the solution we develop in this paper, however, applies to all related sequential risk-taking tasks. We chose to focus on the BART in particular for several reasons. First, the BART is currently the most widely used and tested sequential risk-taking task and has been shown to relate to real-world risk taking (e.g., smoking, illegal drug use, having unprotected sex) in a broad range of populations (Aklin, Lejuez, Zvolensky, Kahler, & Gwadz, 2005; Bornovalova, Gwadz, Kahler, Aklin, & Lejuez, 2008; Crowley, Raymond, Mikulich-Gilbertson, Thompson, & Lejuez, 2006; Lejuez et al., 2002, 2003, 2007; Lejuez, Aklin, Bornovalova, & Moolchan, 2005; Lejuez, Simmons, Aklin, Daughters, & Dvir, 2004), with more recent evidence indicating its link to neurobehavioral correlates of risk behavior (Fecteau et al., 2007; Fein & Chang, 2008). Second, the BART, in our opinion, is a more engaging task for adults (our population of focus) than other sequential risk-taking tasks like the devil task (Slovic, 1966), which was originally designed and used with young children (see, e.g., Hoffrage et al., 2003; Jamieson, 1969; Montgomery, 1974). Finally, the cognitive processes, which we draw on in our study, are the arguably the best understood in the BART of all the tasks (see Pleskac, Yechiam, & Lejuez, 2008; Wallsten et al., 2005).

On each trial of the BART, respondents inflate a simulated balloon presented on a computer screen. The risky play option is to pump up the balloon (by pressing a "Pump" button), which if chosen can inflate the balloon and reward participants with a constant amount of money (typically 5¢) placed in a temporary bank off the screen. Sometimes, however, pumping the balloon causes it to explode, which means respondents lose their accumulated money and the balloon trial ends. The safe option is to choose to stop pumping the balloon by pressing a button (labeled as "Collect $$$"), which allows participants to collect their earned money from the temporary bank and begin another balloon trial (for more details see Lejuez et al., 2002).

The behavioral statistic most commonly analyzed in the BART and other sequential risk-taking tasks is the adjusted score (e.g., Lejuez et al., 2002). It is the mean number of play (i.e., pump) responses on trials that do not end in an explosion. The adjusted score can be contrasted with the *unadjusted score*, which is the mean number of play responses on all trials (trials that end with an explosion and when the participant chooses to stop). The unadjusted score, however, is a poor index of risk-taking behavior because it includes trials that ended before the participant chose to terminate the trial and thus biases the unadjusted score downward.

To a lesser degree, however, the adjusted score is biased low as well. This fact is illustrated by considering a third score: the *target score*. This score is the average number of play choices (pumps) respondents would have executed on all trials, regardless of whether individual trials were terminated by the stochastic gambling environment or by the participant. The target score is actually the statistic of interest in sequential risk taking tasks because it is an unbiased estimator of how far participants would have pumped on each trial. The target score is unobservable with the current response mode in sequential risk taking tasks, which leaves the adjusted score as the next-best statistic to index risk-taking behavior. A side-by-side comparison of the target and adjusted scores reveals that the adjusted score is based on an assumption that trials ending in failures (explosions) do so independently of respondents' behavior. The independence assumption is incorrect. This is because in the BART and more generally in all the sequential risk-taking paradigms the more times respondents choose the risky option the more likely the trial will terminate with an explosion (failure). Consequently, the adjusted score tends to filter out the longer response sequences, which causes the average adjusted score to be lower than the average targeted score.

The appendix provides a formal proof that the adjusted score is always less than the target score. Several observations should be noted about our proof. First, it does not rely on any specific assumption about the psychological process generating choice behavior. Instead, the proof shows that for any choice process the adjusted score is biased relative to the target score. Second, the proof holds for a large class of sequential risk-taking tasks. These include, for example,

tasks like the BART in which the probability of a failure increases with each successful play option (see Equation A8) as well as for tasks in which the probability of a failure remains constant (see Pleskac, 2008). In fact, the only property necessary for the adjusted score in the sequential risk-taking paradigm to be biased is that the probability of a failure (explosion) is between 0 and 1, exclusive, for all trials and for each sequential choice. Finally, the proof says nothing about the magnitude of the bias. The obvious question is whether it is large enough to matter for either clinical testing or research purposes.[1]

Automatic BART

To assess the magnitude of the bias, the targeted number of pumps per trial must be made observable. We developed the automatic BART for this purpose. Instead of sequentially pumping the balloon, as in the standard or manual paradigm, the automatic BART asks participants to simply type the target number of pumps they wish to take at the beginning of a trial. Once they accept this value, participants watch the balloon as it automatically expands until either the stated number of pumps is reached or it explodes. We expected that risk-taking behavior in the automatic version would be similar to the manual version based on our knowledge of the cognitive strategy that participants use during the manual BART. Using formal cognitive models (Pleskac, 2008; Wallsten et al., 2005) as well as posttask surveys (Pleskac, 2004), we have found that when completing the manual BART respondents set a target number of pumps before each balloon and then pump to that target. Our working hypothesis was that during the automatic BART participants would simply enter their already chosen targeted pump number. A within-subject comparison between the adjusted and unadjusted scores in the automatic BART and the manual BART can serve as an initial test of this prediction. If the cognitive strategies are roughly equivalent between the automatic and manual BART, then the scores should be statistically equal to those scores in the standard paradigm. Furthermore, if in fact there is statistical equivalence between the adjusted and unadjusted scores, then the difference between the target score in the automatic paradigm and the adjusted score in the manual paradigm can serve as an estimate of the magnitude of the bias of the adjusted score.

To test this prediction and assess the magnitude of the bias of the adjusted score we conducted a study where participants completed both the manual and the automatic versions. As a further check on the similarity of the two paradigms, we compared correlations with a subset of self-report measures that are related to the adjusted BART score. In the study, we included indices of (a) overall self-reported substance use, (b) sensation-seeking, and (c) impulsivity. In addition, we obtained (d) an anxiety measure hypothesized to be negatively related to the adjusted BART score (Loewenstein, Weber, Hsee, & Welch, 2001; Manes et al., 2002; Reiss, Peterson, Gursky, & Mcnally, 1986), and (e) a measure of depressive symptoms as an indicator of divergent validity (Lejuez et al., 2002).

Increasing the Number of Risky Responses

Finally, in the study we also addressed a second limitation of the original work with the BART. Specifically, almost all respondents—real world risk takers and risk avoiders—rarely choose to play the risky option enough during a given trial so as to maximize their expected earnings. For example, in the BART participants on average choose to play the risky option approximately 30 times when they should do so on average 64 times to maximize their expected earnings.[2] Choosing the risky option more or less than 64 times per trial will decrease expected earnings and the earnings become less and less as the distance from 64 pumps increases (see Lejuez et al., 2002; Pleskac, 2008; Wallsten et al., 2005). As a result, in past studies those participants who were more risk-seeking in the task (pumping closer to 64 times per balloon but not more than 64) tend to earn more. In other words, risk-taking in the task has been confounded with earning more money. This pattern of behavior may be problematic if interventions and/or experimental manipulations, such as stress induction or drug administration, are used to alter behavior in the BART so that increasing risk-taking also increases earnings. Past results show that extensive experience with the task does not appreciably change risk taking in the task (Lejuez et al., 2003). Therefore, in addition to removing the bias in the risk taking measure, we attempted to increase responding by (a) altering the instructions for both versions to give participants explicit instructions about the best decision strategy to use during the BART and (b) providing explicit event feedback on all trials (details in both cases are provided in the method section).

Method

Participants

Participants were 75 undergraduate students at the University of Maryland, College Park (34 female and 41 male; M age = 20.9; SD = 2.05). Fifty-three percent described themselves as White, 32% as Black or African American, 8% as Asian or Southeast Asian, 5% as Hispanic or Latino, and the remaining 2% marked "Other" or chose not to respond to the question. Participants completed a set of questionnaires, reported past real-world risky behavior, and played both the manual and the automatic versions of the BART. Participants were paid between $10 and $25 depending on their final scores. The specific amount was scaled from the overall earnings in both versions.

[1] Note one way to answer this question would be to model the distribution $P(y)$ as we did in Wallsten et al. (2005) or in Pleskac (2008). However in that case, the validity of the estimated magnitude of the bias depends on the validity of the model. Thus, we leave the theoretical result as it stands and turn to an empirical assessment of the magnitude of the bias in the BART.

[2] This maximizing strategy assumes the goal of participants is to maximize expected value (as opposed to utility) and they have full knowledge of the structure of the task (see Pleskac, 2008).

Our exclusive use of college students is worth noting before proceeding. As outlined by Tull, Patterson, Bornovalova, Hopko, and Lejuez (2008), and in line with original development work with the BART, we chose to study college students instead of a clinical sample because we are interested in risk taking propensity as a vulnerability to risky substance use and wanted to limit any possible confounds chronic substance use might involve. Also in line with our original development work with the BART, this more narrow focus will then set the stage for an extension to clinical samples in future work.

Materials and Apparatus

Self-Report Measures

The self-report battery included measures to assess convergent validity with risk-related measures (i.e., positive relationship with sensation seeking and impulsivity) and anxiety (i.e., negative relationship with anxiety sensitivity), a measure to establish divergent validity (depression), and a criterion measure of substance use.

Sensation Seeking Scale. The Sensation Seeking Scale (Zuckerman, Eysenck, & Eysenck, 1978) examines optimal levels of stimulation. This 40-item questionnaire presents participants with forced choices between two opposite statements to determine the degree to which people seek varied, novel, complex, intense situations and experiences (Zuckerman, 1994). This instrument has excellent psychometric properties with internal consistencies ranging from .83 to .86 (Zuckerman, 1994; Zuckerman et al., 1978).

Eysenck Impulsivity Subscale. The Impulsivity subscale of the Eysenck Impulsiveness Scale includes 19 forced choice (yes or no) items, with higher scores indicating greater levels of impulsivity. Internal consistency for the subscale is good ($\alpha = .84$; Eysenck et al., 1985), and is consistent across males and females.

Anxiety Sensitivity Index (ASI). The ASI is a 16-item self-report questionnaire using a 5 point Likert scale ranging from 0 (very little) to 4 (very much) designed to assess the degree to which an individual is concerned about the possible and negative consequences of arousal-based anxiety symptoms. The ASI has high levels of internal consistency ($\alpha = .79$-$.90$; Peterson & Reiss, 1992).

Center for Epidemiological Studies – Depression Scale (CES-D). The CES-D is a short self-report scale designed to measure individual's depressive symptomatology over the immediately prior past 2 weeks. Alpha reliabilities for the CES-D are .80 or higher depending on the sample (Radloff, 1977). Scores of 16 and above indicate high depressive symptoms.

Risky Substance Use Behaviors

Alcohol was assessed as lifetime heaviest frequency of binge drinking. In line with its assessment on the Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993), binge drinking was defined as five drinks in one time period for men and four drinks for women. Never was scored as 0, once as 1, less than monthly as 2, 2–4 times per month as 3, 2–3 times per week as 4, and 4 or more days per week as 5. Binge drinking was used instead of a simple frequency measure to tap more closely the risky aspect of alcohol use. Both *nicotine* and *marijuana* were assessed as lifetime heaviest frequency of use along the same 0–5 scale with any use per day being the critical event. Other drug use was assessed as number of drug classes ever used across stimulants (e.g., cocaine, amphetamine, meth), CNS depressants (e.g., sedatives, heroin), and hallucinogens (e.g., ecstacy, PCP). Any use was assessed instead of frequency because of low frequency across each, and the resulting high negative skew. Given high internal consistency in the current sample across these 4 items [lifetime heaviest frequency of binge drinking (0–5), lifetime heaviest frequency of nicotine use (0–5), lifetime heaviest frequency of marijuana use (0–5), and lifetime use of other drug classes across stimulants, CNS depressants, and hallucinogens (0–3); $\alpha = .81$], and consistent with the assessment of risky substance use in previous BART studies (e.g., Lejuez et al., 2002), we formed a composite measure of lifetime risky substance use by converting each of the four items to z-scores and summing these z scores.

Manual and automatic BART. In both versions the maximum number of pumps possible was set to 128 for each balloon with an explosion a priori equally likely to occur on any given pump subject to the constraint that within each sequence of 10 balloons the average explosion point was on pump 64. All participants saw the same balloons in the same order to limit extraneous variability.

In comparing the manual and automatic versions of the BART, we made two changes from the standard BART paradigm. One was to (truthfully) inform participants that they could earn the most money on average if they pumped 64 times on each trial (see Lejuez et al., 2002). The primary purpose of this instruction was to attempt to increase the number of times participants choose the risky option during the BART. In addition, it served to minimize learning across the paradigms. Their instructions read in part:

> The explosion point varies across balloons, ranging from the first pump to the 128th pump. The ideal number of pumps is 64. What that means is that if you were to make the same number of pumps on every balloon, your best strategy would be to make 64 pumps for every balloon. This would give you the most money over a long period of time. However, the actual number of pumps for any particular balloon will vary, so the best overall strategy may not be the best strategy for any one balloon.

The second change was to inform participants of the pump on which the balloon would have exploded on trials that they successfully terminated. The effect of this change was to provide event feedback on all trials, not just on those ending in explosions.

The manual BART proceeded as described earlier. A trial began with a balloon and the pump and collect response buttons on the screen; and the participant pumped until deciding to stop or the balloon exploded. In the automatic

condition, the balloon appeared on the screen along with a single box in which the participant entered the number of pumps he or she wanted to wanted to use. The balloon then automatically inflated to that number unless it exploded sooner. In both tasks, a pump inflated the balloon and when an explosion occurred a "pop" sound was emitted from the computer.

Procedure

Participants were recruited via advertisements across the University of Maryland campus. Although the advertisement invited individuals at all levels of risk, we also used the phrase "Are you a risk taker?" to increase the likelihood of recruiting a representative number of individuals at the upper end of the risk-taking continuum. Aside from undergraduate status, no other inclusion/exclusion criteria were used.

Upon arrival at the laboratory and after giving informed consent, participants completed two packets of self-report assessment measures and the two versions of the BART. The order of the manual BART and automatic BART, as well as the questionnaires was counterbalanced. The entire session lasted approximately 1 hour.

Results

Figure 1 displays the average unadjusted and adjusted scores for both the manual and automatic conditions as well as the target score for the automatic condition. Recall that in the manual paradigm the unadjusted score is the average number of pumps taken on all balloons. The adjusted score is the average number of pumps taken on only the balloons that did not end in an explosion. These same measures can be inferred in the automatic paradigm from the participants' responses and the actual point of explosion of the balloons. Finally, the target score is the average stated number of pumps for all balloons, which is observable only in the automatic condition. Note that within each response mode, the unadjusted scores are about 9 to 12 pumps less than the adjusted, which is consistent with the premise that the unadjusted score is a poor estimator for the target score.

For both the adjusted and unadjusted scores, we ran a mixed model analysis of variance where the between groups factor was gender and the within factor was the response mode (adjusted and manual). As Figure 1 shows whether we use adjusted ($F(1, 73) = 7.9$, $p < .01$ $MSE = 246.9$, $\eta_p^2 = .10$), unadjusted ($F(1, 73) = 6.5$, $p < .05$, $MSE = 81.2$, $\eta_p^2 = .08$), or the target score, $t(73) = 2.5$, $p < .05$, $d = .41$, females were more risk averse than males. Figure 1 also shows, consistent with our a priori prediction, that the adjusted ($F(1, 73) = 2.44$, *ns*, $MSE = 48.5$) and the unadjusted ($F(1, 73) = .72$, *ns*, $MSE = 29.4$) scores are statistically equivalent across the two conditions. Importantly, the interaction between gender and response mode was not significant for both the adjusted ($F(1, 73) = 1.22$, *ns*, $MSE = 48.5$) and the unadjusted ($F(1, 73) = .78$, *ns*, $MSE = 29.4$) scores. These results imply that (a) the change in response procedure did not change participants' risk-taking behavior and that (b) regardless of response mode females were more risk averse than males. As further evidence of this statistical equivalence of the two different response modes, the linear correlation between the adjusted scores for the manual and automatic condition across genders was $r = .71$, $p < .01$.

The adjusted scores across genders ($M = 53.6$ in the manual and $M = 55.5$ in the auto) are also higher than the average of $M = 30.4$ obtained across prior studies ($N = 448$; $SE = 2.2$; 95% CI = 26.0 < score <34.6) (see Aklin et al., 2005; Crowley et al., 2006; Jones & Lejuez, 2005; Lejuez et al., 2005; Lejuez et al., 2002, 2003). The change is undoubtedly because of the combined effects of providing optimal instructions and full feedback.

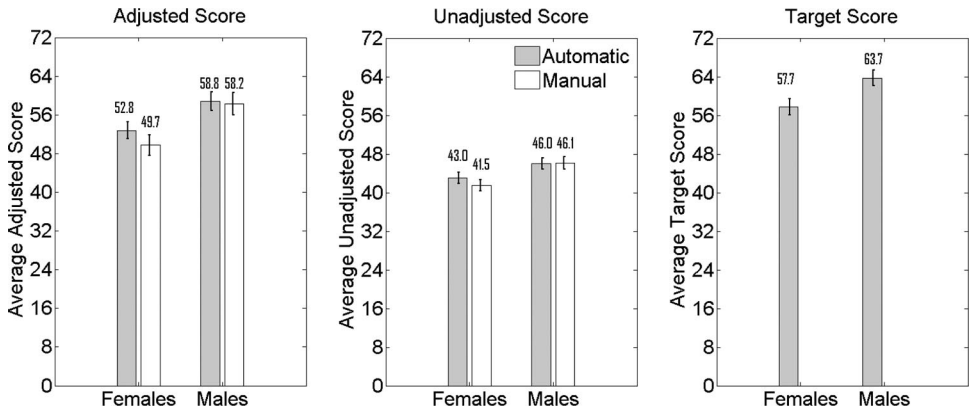

Figure 1. The adjusted, unadjusted, and target scores in the manual and automatic conditions. The unadjusted score is the average number of pumps taken on all balloons, and this score is observed in the manual paradigm and inferred in the automatic. The adjusted score is the average number of pumps taken on only the balloons that did not end in an explosion. The target score is the average stated number of pumps for all balloons in the automatic condition. The values in the figures are the means. Error bars indicate ±1 *SE*.

Returning to comparisons of the automatic and manual conditions, as a further check on the similarity of the two paradigms, we compared the linear correlations of the self-report measures with the target scores in the automatic condition and with the adjusted scores in the manual condition each. The results are shown in Table 1. The pattern is virtually identical for the two different response modes implying the external validity of the BART is not compromised by the new automatic response procedure.[3]

The results summarized in Figure 1 and Table 1 provide no evidence that participants adopted different response strategies in the automatic versus the manual condition. That conclusion and the fact that there were a substantial number of trials that ended in an explosion for the manual ($M = 13.7$; $Mdn = 13.0$; $SD = 3.4$) and automatic version ($M = 14.1$; $Mdn = 14.0$; $SD = 2.8$) permits us to estimate the size of the bias in the adjusted score in either condition. Because gender did not interact with performance between the automatic and manual BART we collapsed across gender for all subsequent analyses. Comparing the adjusted scores to the target score in the automatic reveals a significant 4.9 pump bias, $t(74) = 8.9$, $p < .01$, $d = .31$ in the automatic condition and a 7.0 pump bias in the manual, $t(74) = 5.3$, $p < .01$, $d = .26$, respectively.

Empirically, the automatic response procedure has two further potential advantages over the manual response procedure. First, as we have already discussed, the automatic procedure makes the intended pumping behavior observable on the balloon trials that ended in an explosion. These trials are potentially quite interesting because these were the trials when respondents were on average the most risk seeking. On exploding balloons during automatic BART respondents wanted to pump on average 65.9 ($SD = 10.8$) times. This was significantly more than the nonexploding balloons in the automatic condition, $t(73) = 8.8$, $p < .01$, $d = .67$ and manual condition, $t(73) = 7.7$, $p < .01$, $d = .50$.[4] A final advantage of the automatic BART is that the task takes less time to administer. On average the automatic BART took 3.5 minutes ($Mdn = 3.2$, $SD = 1.5$) while the manual BART took twice as long ($M = 7.1$, $Mdn = 6.8$, $SD = 1.9$), $t(74) = 14.2$, $p < .01$, $d = 1.5$.

Discussion

Sequential risk-taking tasks are a powerful tool for identifying and studying real world risk-taking. There are, however, two challenges in using and interpreting the results from these tasks. The first challenge is that behavior in such tasks often terminates before the so-called optimal stopping point is reached. For example, pumping behavior in the BART is typically well beneath the maximizing value of 64 pumps. This behavioral pattern can lead to problems in interpreting levels of and changes to risk taking during the task might mean. A second limitation for sequential risk taking tasks has been that some of the trials end in a failure and consequently how many more risks participants intended to take on those trials is left unobserved. As a solution to this problem, researchers have begun using an adjusted score as a measure of risk taking behavior. This score is the average number of risks participants took on trials that did not end in a failure. However, we have shown that this solution introduces its own problems. In particular, the adjusted score is a biased estimate of how many risks participants intended take on all the trials. In this paper, we directly addressed these limitations of sequential risk-taking tasks and identified possible solutions for them. We first consider the issue of less than maximizing behavior in sequential risk taking tasks and then turn to the issue of the biased adjusted score.

A Solution for Less Than Maximizing Behavior in the BART

Past studies with the BART have shown that participants typically exhibit adjusted scores between 26 and 35 pumps. A concern about this level of pumping in the BART is that manipulations hypothesized to increase risk taking in the BART (e.g., stress induction or drug administration) would cause respondents to make more money and thus confound an increase in risk taking with performing more optimally in the task. Based on the hypothesis that the low pumping behavior might be the result of insufficient experience with the BART past research investigated whether increasing the number of trials from 30 to 90 trials would eliminate the problem (see Lejuez et al., 2003). The increase in trials had little to no effect. In this paper, we took two alternative steps to address this problem. One was to modify the instructions to tell participants that the expected-value maximizing strategy is to pump 64 times on each trial. The other was to provide event feedback on trials ending in explosions as well as on trials terminated by the respondent. These changes addressed the problem directly and were successful. Indeed, the average target score of 61 approached the expected-value maximizing strategy of 64 pumps. In fact, 29 out of 75 (39%) of the participants had mean target scores greater than 64, and thereby earned less than their counterparts who were closer to 64 pumps on average. Future work should address which of these changes (instructions, feedback, or the automatic response mode) have a larger effect in moving behavior closer to the maximizing goal.

A Solution for the Bias of the Adjusted Score

In terms of the bias of the adjusted score we have shown both formally and empirically that this adjustment still yields a measure of risk taking that is a biased estimator of the parameter of interest: the targeted pumps on all balloons. Although the bias in the adjusted score is less than the bias of the unadjusted score, it is nevertheless still present. The precise amount of bias, of course, will vary with experi-

[3] Similar correlational patterns were observed with the adjusted scores from the automatic condition.

[4] A computer error prevented us from calculating the standard deviations and the average targeted pumps entered on nonexploding balloons for one participant.

Table 1
Linear Correlations Between Substance Use, the Target Score for the Automatic Condition, the Adjusted Score for the Manual Condition and Self-Report Measures

	Ave. (*SD*)	Substance use	Gender	Automatic	Manual	Sensation seeking	Impulsivity	Anxiety
Gender	.45 (—)	−.10						
Automatic	60.4 (10.8)	.28*	.28*					
Manual	53.6 (13.9)	.32**	.31*	.62**				
Sensation seeking	20.6 (6.9)	.53**	.25*	.33**	.28*			
Impulsivity	7.6 (4.1)	.21	.21	.17	.06	.37**		
Anxiety	34.2 (8.2)	−.09	−.20	−.30**	−.31*	−.26*	.06	
Depression	14.0 (9.5)	.06	−.16	.13	−.02	−.05	.21	.36**

Note. * $p < .05$. ** $p < .01$.

mental conditions, and might not be constant over the full range of possible scores, but we know from our formal proof (see Appendix) that it will never be zero. Moreover, because the proof is so general, we know that the bias exists to some degree for all sequential risk-taking tasks, such as Slovic's devil task or the Angling Risk Tasks (Pleskac, 2008), so long as the task is in fact risky (i.e., the probability of a success or a failure is between 0 and 1 exclusive).

Empirically, the Cohen's *d* estimates imply that the bias of the adjusted score in this study is a small to medium sized effect. The size of the effect may be larger or smaller in other tasks, depending on their exact parameters. Even a small to medium bias, however, can have consequences. For example, conclusions and/or diagnoses that depend on the numerical magnitude of the adjusted BART score may be especially vulnerable to this bias (e.g., classifying people as risk averse or risk seeking according to whether their adjusted score exceeds some threshold value).

We have also put forth a solution to the problem of the bias of the adjusted score in the form of an automatic response mode for sequential risk-taking tasks. This response mode makes participant's intended number of responses on each trial observable and thus circumvents the necessity of calculating the adjusted score in the first place. This methodological solution is of course applicable to all sequential risk-taking tasks. We see this a priori methodological solution as a better alternative than the enlistment of an after-the-fact statistical correction as is commonly recommended for situations plagued by missing data (e.g., Little & Rubin, 2002; Schafer & Graham, 2002). Statistical corrections carry their own set of assumptions, the validity of which determines the validity of the correction (Schafer & Graham, 2002).

Another alternative to avoid the bias is to implement a deceptive strategy originally advocated with sequential risk taking tasks (see Jamieson, 1969; Montgomery, 1974; Slovic, 1966). The strategy was to tell respondents a failure could occur after any one of their choices to play the gamble (e.g., pump the balloon in the BART), but the failure event is set to occur after the very last possible sequential choice (e.g., 128 in our study). We, however, believe the deceptive component of this strategy makes it less attractive than the automatic response mode, especially in cases of multiple trials where the respondent is likely to learn the true explosion pattern.

An advantage of the automatic response mode is that it corresponds with the strategy respondents use to complete the manual BART (see Pleskac, 2008; Wallsten et al., 2005). Specifically, using formal models of the respondents' underlying cognitive processes, these authors showed that during the manual BART respondents set a target number of pumps before each balloon and then stochastically pumped to that target. Indeed as further evidence of this correspondence, we found that the change from a manual to an automatic response procedure does not appear to change behavior in the BART.

Despite this at least surface level correspondence between the two different response modes, there may still be cognitive or emotional differences between the two response modes, which may affect the ability of the BART (or any other sequential risk taking task) implemented with an automatic response mode to predict real world-risk taking. For instance, one might hypothesize that the automatic response mode removes an aspect of impulsivity from the response procedure. This may be a crucial element as impulsivity and/or inhibitory control are typically thought to be a crucial aspect of drug-use (Fillmore & Rush, 2002; Fillmore, Rush, & Hays, 2002). Similarly, others may contend that the automatic procedure taps a completely different cognitive process—a more analytical mode of processing (Kahneman, 2003; Sloman, 2002; Stanovich & West, 2000)—that also may affect its clinical diagnosticity (but see Reyna & Farley, 2006). These are all possible hypotheses that deserve future attention. The current data, however, suggest that these alternatives are not at play or at least have a minimal impact on the observed risk taking behavior.

Limitations and Future Directions

A few limitations of this work should be noted. First, we did not test the automatic and manual versions with the typical instructions that do not explain the maximizing strategy and with conditions involving feedback only on task-terminated trials. Instead we relied on the fact that the manual BART already is well characterized with the noninformative instruction set and asymmetrical feedback condition (see, e.g., Aklin et al., 2005; Crowley et al., 2006; Jones & Lejuez, 2005; Lejuez et al., 2005; Lejuez et al., 2002, 2003). Nevertheless, further tests of the automatic BART should utilize a between groups design to understand more fully the role of these

changes. A second limitation is that our use of a college sample may restrict the generalizability of our results. Certainly, future work should seek to investigate the robustness and applicability of our results to risk taking in other populations. Despite this limitation, the fact that both the automatic and manual exhibited a similar level of association with substance use suggests that either task is equally capable of identifying individuals with a propensity or vulnerability to engage in risky substance use.

Despite these limitations, this new version of the BART has at least three advantages over the manual version. First, it makes use of the responses from all trials rather than from a subset, thereby increasing the reliability of the data. Second, the distribution of target scores from the automatic BART contains more information than the adjusted score distributions, revealing intended pumping behavior for exploding balloons. Third and finally, the sessions are relatively quick and carry less motor requirements (entering a single number vs. clicking a button). This last fact makes the automatic BART an excellent paradigm to use in studies that benefit from such characteristics, such as those involving neuroimaging or with clinical populations with motor limitations.

References

Aklin, W. M., Lejuez, C. W., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behaviour Research and Therapy, 43,* 215–228.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7–15.

Bernstein, D. M., Whittlesea, B. W., & Loftus, E. F. (2002). Increasing confidence in remote autobiographical memory and general knowledge: Extensions of the revelation effect. *Memory & Cognition, 30, 432*–438.

Bornovalova, M. A., Gwadz, M., Kahler, C. W., Aklin, W. M., & Lejuez, C. W. (2008). Sensation seeking and risk-taking propensity as mediators in the relationship between childhood abuse and HIV-related risk behavior. *Child Abuse and Neglect, 32,* 99–109.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14,* 253–262.

Crowley, T. J., Raymond, K. M., Mikulich-Gilbertson, S. K., Thompson, L. T., & Lejuez, C. W. (2006). A risk-taking "set" in a novel task among adolescents with serious conduct and substance problems. *Journal of the American Academy of Child and Adolescent Psychiatry, 45,* 175–183.

Eysenck, S. B. G., Pearson, P. R., Easting, G., & Allsopp, J. F. (1985). Age norms for impulsiveness, venturesomeness and empathy in adults. *Personality and Individual Differences, 6,* 613–619.

Fecteau, S., Pascual-Leone, A., Zald, D. H., Liguori, P., Theoret, H., Boggio, P. S., et al. (2007). Activation of prefrontal cortex by transcranial direct current stimulation reduces appetite for risk during ambiguous decision making. *The Journal of Neuroscience, 27,* 6212–6218.

Fein, G., & Chang, M. (2008). Smaller feedback ERN amplitudes during the BART are associated with a greater family history density of alcohol problems in treatment-naive alcoholics. *Drug and Alcohol Dependence, 92,* 141–148.

Fillmore, M. T., & Rush, C. R. (2002). Impaired inhibitory control of behavior in chronic cocaine users. *Drug and Alcohol Dependence, 66,* 265–273.

Fillmore, M. T., Rush, C. R., & Hays, L. (2002). Acute effects of oral cocaine on inhibitory control of behavior in humans. *Drug and Alcohol Dependence, 67,* 157–167.

Harrison, J. D., Young, J. M., Butow, P., Salkeld, G., & Solomon, M. J. (2005). Is it worth the risk? A systematic review of instruments that measure risk propensity for use in the health setting. *Social Science & Medicine, 60,* 1385–1396.

Hoffrage, U., Weber, A., Hertwig, R., & Chase, V. M. (2003). How to keep children safe in traffic: Find the daredevils early. *Journal of Experimental Psychology: Applied, 9,* 249–260.

Hyman, I. E., & Loftus, E. F. (1998). Errors in autobiographical memory. *Clinical Psychology Review, 18, 933*–947.

Jamieson, B. D. (1969). The influences of birth order, family size, and sex differences on risk-taking behavior. *British Journal of Social and Clinical Psychology, 8,* 1–8.

Jones, H. A., Lejuez, C. W. (2005). Personality correlates of caffeine dependence: The role of sensation seeking, impulsivity, and risk taking. *Experimental and Clinical Psychopharmacology, 13,* 259–266.

Kahneman, D. (2003). A perspective on judgment and choice-Mapping bounded rationality. *American Psychologist, 58,* 697–720.

Lane, S. D., Cherek, D. R., Pietras, C. J., & Tcheremissine, O. V. (2004). Alcohol effects on human risk taking. *Psychopharmacology, 172,* 68–77.

Lejuez, C. W., Aklin, W. M., Bornovalova, M. A., & Moolchan, E. T. (2005). Differences in risk-taking propensity across inner-city adolescent ever- and never-smokers. *Nicotine & Tobacco Research, 7,* 71–79.

Lejuez, C. W., Aklin, W. M., Daughters, S. B., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2007). Reliability and validity of the youth version of the Balloon Analogue Risk Task (BART-Y) in the assessment of risk-taking behavior among inner-city adolescents. *Journal of Clinical Child and Adolescent Psychology, 36, 106*–111.

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., et al. (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology, 11,* 26–33.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8,* 75–84.

Lejuez, C. W., Simmons, B. L., Aklin, W. M., Daughters, S. B., & Dvir, S. (2004). Risk-taking propensity and risky sexual behavior of individuals in residential substance use treatment. *Addictive Behaviors, 29,* 1643–1647.

Leland, D. S., & Paulus, M. P. (2005). Increased risk-taking decision-making but not altered response to punishment in stimulant-using young adults. *Drug and Alcohol Dependence, 78,* 83–90.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127,* 267–286.

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., et al. (2002). Decision-making processes following damage to the prefrontal cortex. *Brain, 125,* 624–639.

Montgomery, R. (1974). Transmissioon of risk-taking through modeling at two age levels. *Psychological Reports, 34,* 1187–1196.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., & Stein, M. B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *Neuroimage, 19,* 1439–1448.

Peterson, R. A., & Reiss, S. (1992). *Anxiety sensitivity manual* (2nd ed.). Worthington, OH: International Diagnostic Systems.

Pleskac, T. J. (2004). *Evaluating cognitive sequential risk-taking models: Manipulations of the stochastic process.* Unpublished Dissertation, University of Maryland, College Park.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology-Learning Memory & Cognition, 34,* 167–185.

Pleskac, T. J., Yechiam, E., & Lejuez, C. W. (2008). A cognitive comparison of complex risky decision-making tasks. Manuscript submitted for publication.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Reiss, S., Peterson, R. A., Gursky, D. M., & Mcnally, R. J. (1986). Anxiety Sensitivity, Anxiety Frequency and the Prediction of Fearfulness. *Behaviour Research and Therapy, 24,* 1–8.

Reyna, V. F., & Farley, F. (2006). Risk and rationality in adolescent decision making: Implications for theory, practice, and public policy. *Psychological Science,* 1–44.

Rogers, R. D., Owen, A. M., Middleton, H. C., Williams, E. J., Pickard, J. D., Sahakian, B. J., et al. (1999). Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *Journal of Neuroscience, 19,* 9029–9038.

Saunders, J. B., Aasland, O. G., Babor, T. F., Delafuente, J. R., & Grant, M. (1993). Development of the Alcohol-Use Disorders Identification Test (Audit): Who collaborative project on early detection of persons with harmful alcohol-consumption. 2. *Addiction, 88,* 791–804.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Slovic, P. (1966). Risk-taking in children: Age and sex differences. *Child Development, 37,* 169–176.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral & Brain Sciences, 23,* 645–726.

Stout, J. C., Rock, S. L., Campbell, M. C., Busemeyer, J. R., & Finn, P. R. (2005). Psychological processes underlying risky decisions in drug abusers. *Psychology of Addictive Behaviors, 19,* 148–157.

Tull, M. T., Patterson, R., Bornovalova, M. B., Hopko, D. R., & Lejuez, C. W. (2008). Analogue Research Methods. In D. McKay (Ed.), *Handbook of research methods in abnormal and clinical psychology.* Thousands Oaks, CA: Sage.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review, 112,* 862–880.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16,* 973–978.

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking.* New York: Cambridge University Press.

Zuckerman, M., Eysenck, S., & Eysenck, H. J. (1978). Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology, 46,* 139–149.

(*Appendix follows*)

Appendix

A Formal Proof of the Bias of Adjusted Scores in Sequential Risk-Taking Tasks

To formally prove that the adjusted score used in the sequential risk-taking paradigm introduces a bias, let S denote a trial that ends in a success (S) (e.g., no explosion), and $P(S)$ the proportion of successful trials in a session. Let the target score Y be the random variable that would be observed if decision makers' response sequences were not terminated (by an explosion) on each trial. The score Y is the behavioral measure of interest because it characterizes the respondents risk taking for all trials during the task. However, in the standard paradigm this variable is latent because Y is not observed on trials that end in a failure (explosion). Theoretically, the score Y can range between 0 and infinity, but practically for the sequential risk-taking tasks studied here it will be limited to n, the maximum number of possible play options that can be chosen on a given trial. The distribution over Y, $P(y)$, is defined over the non-negative integers with the usual restrictions that $0 \leq P(y) \leq 1$ and $\sum_{y=0}^{n} P(y) = 1$. We denote the expected target score as $E(y)$. Only trials that end in success S contribute to the adjusted score. The expected adjusted score is $E(y|S)$ and its distribution is $P(y|S)$. By definition, the adjusted score is biased if $E(y|S) \neq E(y)$. Our proof shows that in fact $E(y|S) < E(y)$.

Bayes' Rule specifies the relationship between $P(y|S)$ and $P(y)$. Specifically,

$$P(y|S) = \frac{P(S|y)P(y)}{P(S)} \tag{A1}$$

where

$$P(S) = \sum_{y'=0}^{n} P(S|y')P(y'). \tag{A2}$$

From Equation A1 it is apparent that

$$P(y|S) \begin{cases} > P(y) & if \quad P(S|y)/P(S) > 1 \\ = P(y) & if \quad P(S|y)/P(S) = 1 \\ < P(Y) & if \quad P(S|y)/P(S) < 1 \end{cases}. \tag{A3}$$

Because Y and therefore the distribution $P(y)$ are unobservable in the standard paradigm, it is not possible to derive the exact relationship between it and $P(y|S)$. We can prove, nevertheless, that $P(y|S)$ shifts down over y relative to $P(y)$ making lower values of y more likely for $P(y|S)$, and therefore that expected adjusted scores always are less than expected target scores. That is,

$$E(y|S) = \sum_{y=0}^{n} P(y|S)y < E(y) = \sum_{y=0}^{n} P(y)y. \tag{A4}$$

To do so we have to find for what values of y the ratio $P(S|y)/P(S)$ is less than, equal to, or greater than one.

Let p_i denote the probability that response i is successful. We assume only that $p_i < 1$ (otherwise the situation is not risky). Therefore, the probability that y responses will end in success is

$$P(S|y) = \prod_{i=1}^{y} p_i.$$

As a consequence,

$$P(S|y) \to 0 \quad as \quad y \to \infty. \tag{A5}$$

Next we define y^* as the unique value of y at which $P(S|y) = P(y)$. Because $P(S)$ is constant for any n (Equation A2), and because $P(S|y)$ decreases monotonically (Equation A5), it follows that

$$\frac{P(S|y)}{P(S)} \begin{cases} > 1 & if \quad y < y^* \\ = 1 & if \quad y = y^* \\ < 1 & if \quad y > y^* \end{cases}. \tag{A6}$$

Rewriting Equation A3 using Equation A6,

$$P(y|S) \begin{cases} > P(y) & if \quad y < y^* \\ = P(y) & if \quad y = y^* \\ < P(y) & if \quad y > y^* \end{cases}. \tag{A7}$$

In words, low values of y are more likely when conditioning on successful trials than unconditionally and high values are less likely; or, as we wrote above, the distribution over y is shifted downward when calculating the adjusted score relative to the target score. From Equation A4, then, it follows that $E(y|S) < E(y)$ and the adjusted score is biased low.

Note that this bias applies to sequential risk taking tasks like the BART where

$$p_i = \frac{n - i}{n - i + 1}. \tag{A8}$$

It also applies to a broader range of sequent risk-taking paradigms with any stochastic environment that makes a failure (explosion) statistically possible ($0 < p < 1$) on each trial.

Received December 20, 2007
Revision received July 2, 2008
Accepted July 2, 2008 ■