Behavior Research Methods (2019) 51:2509–2521
https://doi.org/10.3758/s13428-018-1094-8

Variations on the balloon analogue risk task: A censored regression analysis

Michael E. Young [1] · Anthony W. McCoy [1]

Published online: 27 July 2018

Abstract

In the present project, we reexamined the balloon analogue risk task (BART) by evaluating three variations on the task: one that does not require pumping, one that controls for trial duration, and another that withholds feedback on popping until the end of each trial. To accurately assess the censored data produced by the BART, performance was compared across these variations using Bayesian analysis with censored regression. The first experiment compared a task that required pumping to one that did not, and revealed that the tendency to respond earlier than is optimal does not reflect an avoidance of effort. The second experiment included a condition in which the duration of each trial was held constant by continuing to automatically inflate a balloon to its maximum size after a cash-in response; feedback on the pop time was withheld until the end of each trial. This condition revealed that the tendency to respond earlier is not driven by a desire to finish the task quickly by cashing in early, but the results also strongly suggested that the immediate experience of popping created a greater aversion to risk (although this condition difference was inconsequential by the end of the experiment). The article concludes by considering the implications of these results for cognitive neuroscience approaches to understanding performance on the BART.

Keywords Balloon analogue risk task (BART) · Risk taking · Bayesian · Censored regression

In the study of decision making, researchers often ask people to choose between options that differ in multiple ways, with one option being better on one or more dimensions, and the other option better on the other dimensions (e.g., Chatterjee & Heath, 1996; Kahneman & Tversky, 1979; Rachlin, Raineri, & Cross, 1991; Young, Webb, Rung, & McCoy, 2014). These types of designs help determine the relative importance of the dimensions for each participant.

In one such task, the balloon analogue risk task, or BART (Lejuez et al., 2002), participants are asked to pump up a series of visual representations of a balloon. On each trial, the larger the balloon is before the participant stops pumping, the greater the reward magnitude. But each pump increases the likelihood that a balloon will pop; popping results in no reward on that trial. The BART involves the trade-off between reward magnitude and reward probability: Cashing in quickly results in a high probability of a low-magnitude reward, whereas cashing in later results in a lower probability of a higher-magnitude reward. The BART, however, involves more trade-offs than that. First, pumping the balloon requires effort. Second, pumping the balloon requires time, and cashing in earlier on each trial produces a shorter experiment as well as requiring less physical effort. Thus, a participant's decision to produce fewer pumps could reflect a desire for some combination of higher probability, less effort, or a shorter experiment, rather than a greater average reward magnitude.

In the present study we had two primary goals. First, we wanted to examine variations on the BART, to determine whether factors other than risk aversion might play a role in people's decisions to cash in earlier than expected. In Experiment 1 we examined the role of physical effort, and in Experiment 2 we examined the impact of a desire for a shorter experiment and, secondarily, the timing of the feedback. Second, we addressed these questions using a censored regression analysis, and thus the article also considers whether the so-called risk aversion on the BART is due to a statistical artifact (Pleskac, Wallsten, Wang, & Lejuez, 2008). Experiment 1 compared the traditional analytical approach, based on only analyzing trials that do not result in popping the balloon (i.e., adjusted pumps), to a censored regression approach; due to the

✉ Michael E. Young
michaelyoung@ksu.edu

[1] Department of Psychological Sciences, Kansas State University, 492 Bluemont Hall, Manhattan, KS 66506-5302, USA

advantages of censored regression, this method was then used for Experiment 2.

The multiple advantages of responding earlier rather than later on the BART may partly explain the common observation that participants cash in earlier than is "optimal," where optimal is defined as the point at which the expected value of the reward is maximized. Lejuez et al. (2002) reported that participants only pumped their blue balloon 28 times, when 64 pumps was optimal (44% of optimal), whereas they pumped the yellow and orange balloons 11 and 3.5 times, relative to the optimal 16 and 4 pumps, respectively (about 69% and 82% of optimal). The researchers described this result as evidence of risk aversion among the participants. However, the strength of the tendency to cash in too early was greater for conditions in which the additional effort and time required to maximize the expected value was greater. Paradoxically, the tendency to cash in early was greatest when the risk of popping the balloon was lowest. Importantly, the premature cashing in was also evident in studies that have used only the "thickest" balloon (i.e., the balloon least likely to pop, or the blue balloon in Lejuez et al., 2002). For example, the participants pumping thick balloons in Hunt, Hopko, Bare, Lejuez, and Robinson (2005) cashed in after an average of 35 pumps (55% of optimal), and the participants in Crowley et al. (2009) averaged 26 pumps (41% of optimal).

An alternative explanation of the reported risk aversion was explored by Pleskac et al. (2008). They noted that when the balloon pops, the researcher does not know how many times the participant would have pumped up the balloon before cashing in. Given that participants are only scored on trials in which they cash in (Lejuez et al., 2002), this creates a downward bias in the scoring of their performance, due to truncation of the distribution of the number of pumps. Pleskac et al. proposed a variation on the BART in which participants indicated the number of pumps they wished to produce before each trial started; the computer then produced this number of pumps for the participant. This automated BART produced a slightly larger number of pumps than the manual version, both when using adjusted pumps that artificially truncated the distribution of the automatic condition (to match the truncation of the standard BART) and when using the unadjusted number of pumps by censoring both distributions, by treating trials ending in a pop as involving the number of pumps produced at popping. There was no attempt to model the entire pump distribution, so as to include those values not observed due to premature balloon popping in the original BART.

Censoring is a statistical outcome common in survival analysis—a person who is still alive at the end of a medical study is treated as having a longevity score equal to his or her value at study termination. For example, in a study examining patient survival up to 10 years after surgery, a patient still alive at the end of the study would be assigned a longevity score of 10 years. This patient's score is said to be censored, to distinguish it from a patient's score who died exactly 10 years after surgery. Statisticians can analyze censored data using censored regression, which models the unobserved tail of the distribution by considering the observed data and the number of censored observations (Tobin, 1958). A standard regression must either treat censored data as missing at random (which they are not) or replace these missing values with the maximum possible value that could be observed. In analyses of the BART, the former approach is adopted when an analysis is based on the adjusted score, and the latter approach when analysis is based on the unadjusted score; both approaches result in downward bias of the estimated number of pumps (for further discussion, see Pleskac et al., 2008). Furthermore, analyzing only the adjusted pumps results in a smaller sample size, by omitting popped trials. These trials contain information—namely, that the participant had planned to wait at least that long—that is lost in the traditional analytic approach.

Given that censored regression can derive the most likely distribution of a dependent variable involving censored data, we used a Bayesian analysis to perform a multilevel censored regression for each of the designs that we assessed. This approach allowed us to determine whether our manipulations were affecting cash-in behavior, regardless of whether the methodology can produce censoring (like the original BART) or not (like the automatic BART).

In the present study, we more fully examined the risk aversion claims for the BART by systematically testing variations of the task. In Experiment 1, two versions of the BART were examined: one in which the balloon had to be pumped up, and one in which the balloon automatically increased in size and the participant only needed to respond once in order to terminate the trial, and thus cash in. If participants waited longer when the pump requirement was removed, then effort was playing a role in the decisions to cash in early when pumping was required.

In Experiment 2 we compared a version in which pumping was not required and the balloon popped at a scheduled time (the *inflate-to-pop* condition) to a version in which pumping was still not required but the balloon inflated to its maximum on each trial, regardless of when the participant cashed in (the *inflate-to-full* condition). In the inflate-to-full condition, participants did not know that the balloon popped until the trial was complete, and zero points were awarded when their cash-in time was longer than the scheduled pop time (cf. Jentsch, Woods, Groman, & Seu, 2010). In both conditions, the trial duration was held constant, in order to control for the amount of time to complete the task (there was thus no time benefit to cashing in early). The inflate-to-full condition also eliminated the data truncation caused by popping (cf. Pleskac et al., 2008). However, this method of eliminating truncation by removing the immediate experience of popping could result in less risk aversion, a situation that might also exist in Pleskac et al.'s automatic BART, in which participants are required to choose the number of pumps at the beginning of each trial.

Experiment 1

In our first experiment, participants were assigned to one of two conditions: Either the balloon inflated automatically and only a cash-in response was required, or the participants had to manually inflate the balloon (a conceptual replication of the traditional BART task). We designed this first experiment to assess what changes, if any, the removal of the pumping response would make to the measured behavior on the task. If participants in the manual condition cashed in earlier, then effort was playing at least some role in these decisions.

Participants assigned to the automatic condition simply watched the video of the balloon inflating and made a single response to cash in on the balloon (the size of the balloon at cash-in was directly related to the number of points earned on the trial). If they waited too long, the balloon popped (i.e., the video ended), and they received zero points. Participants in the manual condition were required to press the space bar to inflate the balloon. The participants in this condition watched the same video as those in the automatic condition, but the video would pause after every 100 ms. Once the video paused, participants had to choose between the risky choice (press the space bar and let the balloon continue inflating) and the safe choice (cash in and receive the available points). As with the participants in the automatic condition, if participants waited too long to cash in, the video would end and they would receive zero points. The video played for a maximum of 3,400 ms (100- to 3,400-ms range) in the automatic condition, and up to 34 × 100-ms bins in the manual condition.

In addition to the manipulation above, we also used five balloon thicknesses to create five levels of riskiness associated with waiting. In the most common implementation of the BART (Aklin, Lejuez, Zvolensky, Kahler, & Gwadz, 2005; Fein & Chang, 2008; Hopko et al., 2006; Lejuez et al., 2003), participants experience balloons with a single thickness (typically, 128 pumps max, with 64 pumps being optimal). We were interested in observing how behavior changed on the basis of the riskiness contingency, and whether this sensitivity would differ between the automatic and manual conditions; using multiple thicknesses also avoided ceiling and floor effects that might be observed with some participants. We hypothesized that participants would learn to wait longer for thicker balloons, but we had no expectations regarding differential sensitivity to this manipulation across the conditions.

All analyses were conducted using Bayesian multilevel gamma regression analysis, as implemented in the brms package for R. For the censored analysis, popped trials were classified as censored when the balloon popped—the balloon size when it popped became the censored value. For further discussion of censored regression in multilevel models, see Rodriguez (2007). For the standard analysis, only the trials ending in a cash-in were analyzed (i.e., censored trials were omitted), but we used the same Bayesian method of estimation to allow for direct comparison of the two approaches to analyzing the data.

Method

Participants A total of 114 introductory psychology students (36 men and 78 women) at Kansas State University received course credit for their voluntary participation.

Stimuli In our task, each stimulus comprised a single balloon animation with a predetermined random size at which the balloon would pop. Each participant was programmed to receive the same pop times in the same order, but their experiences differed depending on their cash-in behavior. The balloon animation was a colored circle with a diameter that increased linearly over time (see Fig. 1 for screenshots of the displays) from its minimal diameter of 0.6° visual angle to a maximum of 22.4° of visual angle.

Traditionally, the distribution of pop times had been determined by Eq. 1:

$$\text{Probability of Popping} = \frac{1}{n-i+1} \tag{1}$$

where the *probability of popping* is the probability of the balloon popping on the *i*th pump, and *n* is the maximum number of pumps (Lejuez et al., 2002; Wallsten, Pleskac, & Lejuez, 2005). The pop times were then selected from this distribution in a pseudorandom manner, such that the average pop time was maintained across blocks. For the present study, the distribution of pop times was determined by a superellipsoid function first adopted by Young, Webb, and Jacobs (2011), using Eq. 2 (see Fig. 2):

$$\text{Probability of Popping} = \left(1-\left(\frac{3400-t}{3400}\right)^{\text{power}}\right)^{\frac{1}{\text{power}}} \tag{2}$$

where the *probability of popping* is the probability of the balloon popping at time *t*, and *power* is a parameter in the superellipsoid function that defines the rate of change in probability across the duration. The probability of the balloon popping at higher powers accelerated much earlier, resulting in a balloon that popped after waiting a short time (i.e., thus representing thinner balloons). At low powers the probability of the balloon popping increased much later, resulting in a balloon that popped only after waiting a long time (i.e., thus representing thicker balloons). This equation has the benefit of having the same maximum balloon size across variations in thickness, although the larger sizes are rarely experienced for the thinner balloons. To ensure equality across the two conditions, the likelihood of popping for the automatic condition was determined at 100-ms intervals, to match the size of the intervals in the manual condition.

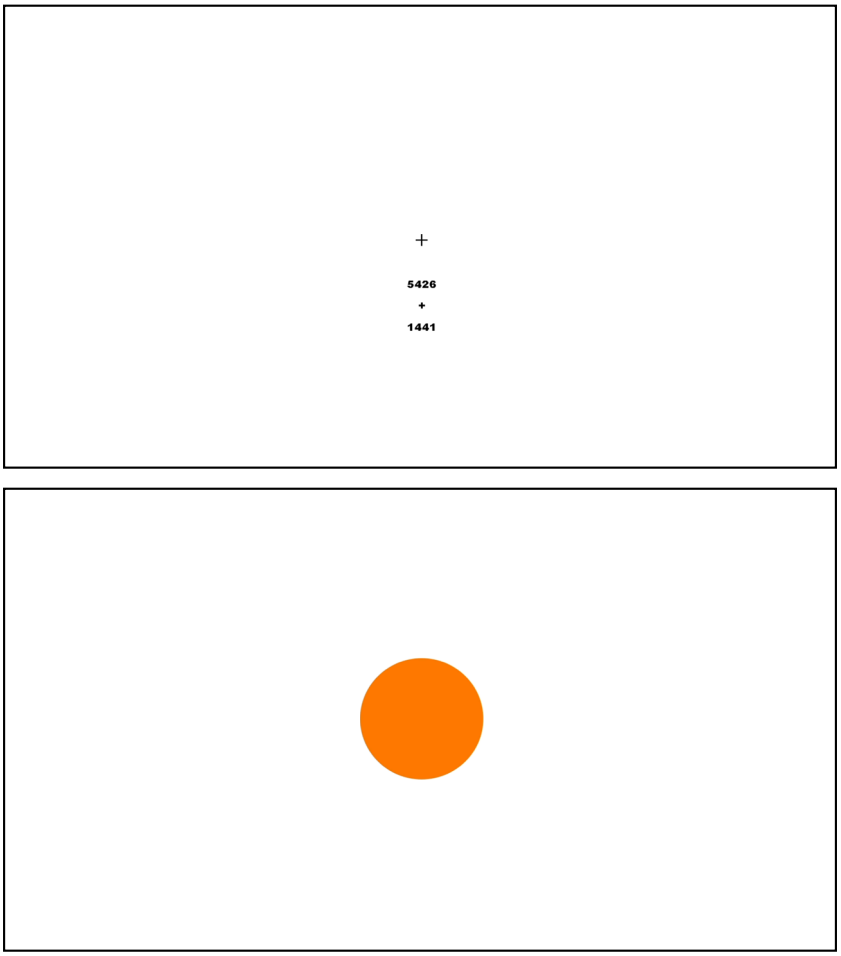

Fig. 1 Screenshots of the stimuli used in both Experiments 1 and 2. (Top) The feedback screen participants were shown at the end of each trial, which included a fixation cross, the number of points earned so far in the experiment, and the number of points earned on the most recent balloon. (Bottom) A depiction of the balloon stimulus seen by participants (after 1,000 ms, or ten pumps). Participants saw five different colored balloons

Participants also completed a questionnaire, which consisted of the UPPS-P (Whiteside & Lynam, 2001) and questions about participants' video game experience and sex. The UPPS-P is a 59-item impulsivity questionnaire that consists of five subscales. Order of completion was counterbalanced between subjects, such that half of the participants completed the UPPS-P and demographic questionnaire first, and the other half completed the BART first. These data were gathered for exploratory purposes, but no clearly significant effects emerged, and thus these variables will receive no further attention.

Procedure Participants completed all of the assigned tasks at one of four identical workstations. Participants were read the instructions and began the task: They were told they would be inflating 250 balloons and that the color of the balloon would change, which would signify that the balloon might now pop earlier or later than it had been. Five colors were experienced within each 50-trial block, but the assignment of colors to power values was counterbalanced such that each color was not predictive of power when it was encountered again later in the experiment. Data collection was programmed using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993) on four Mac Mini computers.

At the beginning of each trial, a fixation cross was presented centrally, and participants pressed any key to begin the balloon animation. In the automatic condition, the balloon animation played until either the balloon popped or the participant had cashed in using the "c" key (whichever happened first); inflation was smooth and continuous, as it would be if accomplished by a machine rather than by an individual blowing up a balloon. Immediately after the balloon animation terminated, a feedback screen was presented. The feedback screen included the central fixation cross, as well as the number of points previously earned, a plus sign, and the number of points earned on the current trial (see Fig. 1). Participants could press any key to terminate the feedback and begin the inflation of the next balloon. In the manual condition a trial progressed similarly to the trials in the automatic condition, except that the balloon animation paused every 100 ms, and the participant had to press

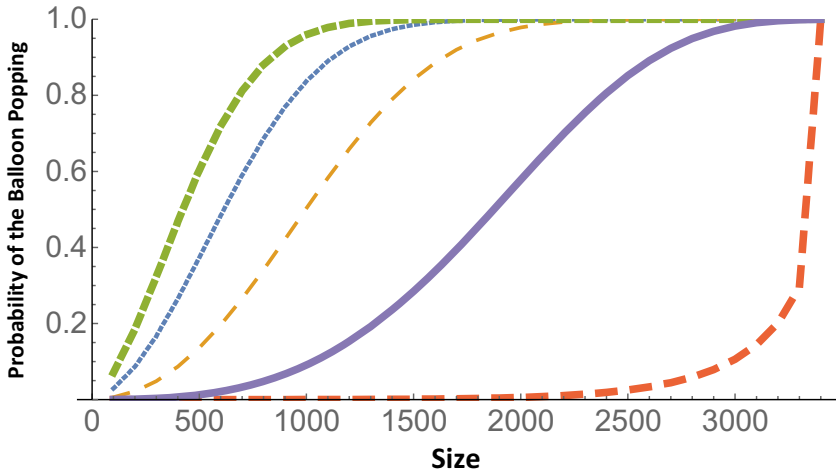

Fig. 2 Cumulative likelihood of the balloon popping before it reached a particular size, when controlled using a superellipsoid function. Each line represents a different balloon thickness, from the thinnest on the left (power = 1.25) to the thickest on the right (power = 0.25)

either the space bar, to have the balloon continue expanding, or the "c" key, to cash in.

We divided the 250 trials into five blocks of five sets of ten balloons each (see Fig. 3). At the end of each set of ten balloons, the color of the balloon changed, as a cue to the participant that the balloon thickness was different. The balloon thickness (power value) changed at the end of each set of trials, but participants had to learn this new environmental contingency through trial and error. Each power value was experienced five times, and the order of experience was Latin-square counterbalanced within participants.

Results

The data and R scripts for both Experiments 1 and 2 are available on the Open Science Framework, at https://osf.io/zq23u/. All participants completed the tasks within the allotted time of 1 h. One participant was dropped for only cashing in on one trial. Although the rest of the participants were included in order to avoid differential attrition, three participants in the manual condition eventually produced a stereotypic behavior in which they cashed in all balloons with the same number of pumps (usually one), which suggests that at least some of the participants in the manual condition were choosing a least-effort strategy. Cash-ins below 100 units in size were excluded for being too fast, which resulted in 827 trials being excluded (roughly 3% of the data).

To equate behavior across the two tasks, the time to cash in was assessed using the size of the balloon at cash in. This metric of balloon size corresponded to the number of milliseconds waited in the automatic condition, and to 100 times the number of pumps in the manual condition. Note that if the balloon popped before cash-in, for the censored regression the size of the balloon when it popped was used as the data value and also tagged as a censored data point.

Plots of the full distribution, geometric means, and optimal sizes (to maximize the expected value) are shown in the top part of Fig. 4. It is apparent that participants were cashing in much earlier than was optimal, especially for the thickest balloons (powers less than 0.75). Note that these distributions and geometric means are based on collapsing across all participants, and thus do not reflect the different numbers of trials that each participant contributed to these distributions, either due to differences in pop rates or failures to complete the experiment. A multilevel model will be used to appropriately weight each participant in the statistical analysis.

We ran a generalized linear multilevel censored regression (Pinheiro & Bates, 2004) on balloon size at cash-in and specified a gamma error distribution (with a log link function), due to the strong skew evident in the latencies. For all of the Bayesian analyses the warm-up or burn-in period was 500, with an additional 1,000 iterations to estimate the posterior distribution of each parameter; three chains were run with these values. The weak priors were $N(0.0, 0.5)$ for the condition regression weights, $N(6, 1)$ for the intercept, $N(0, 1)$ for the power slope, $Cauchy(0, 2)$ for the standard deviations, and $LKJ(2)$ for the correlations. The slope estimate of the power parameter that dictates balloon thickness and the intercept were allowed to vary across participants as random effects. Power and condition (automatic vs. manual) as well as their

Block 1					Block 2					Block 3					Block 4					Block 5				
0.75	1.25	0.50	0.25	1.00	0.25	0.50	0.75	1.00	1.25	1.25	0.25	1.00	0.50	0.75	0.50	1.00	1.25	0.75	0.25	1.00	0.75	0.25	1.25	0.50

Fig. 3 Order in which participants experienced each balloon thickness (power value) in Experiments 1 and 2

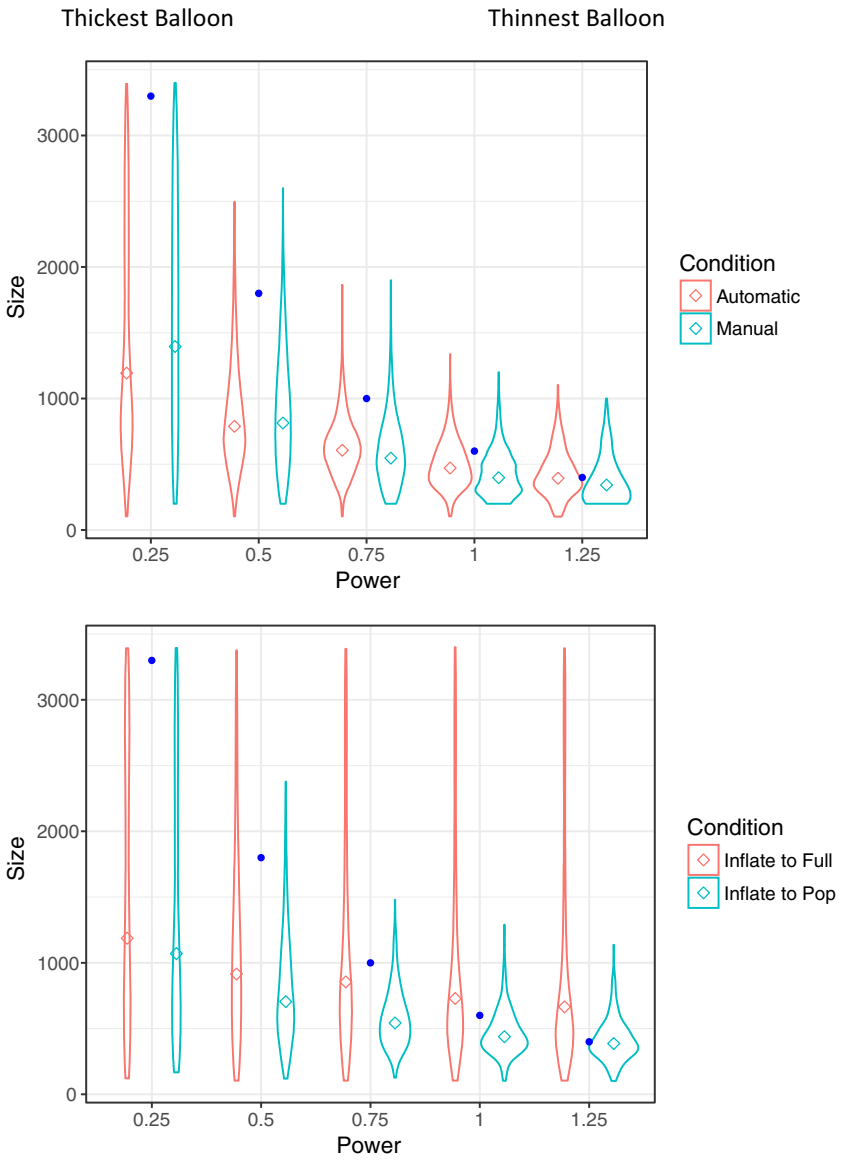

Fig. 4 Balloon size at cash-in for each condition, with the geometric mean (diamond) and optimal value (filled dot), for Experiment 1 (top) and Experiment 2 (bottom). The full data set is represented, and the results do not reflect that participants contributed different amounts of data due to popping or premature termination of the experiment

interaction were evaluated as fixed effects. To avoid issues with multicollinearity, the power value was centered and condition was effect-coded. Visual examination of the performance of the chains and the R-hat values (all 1.05 or below) indicated that convergence of the chains was good (i.e., the posterior distribution was very similar, regardless of the different starting values used in the three chains). For a Bayesian multilevel analysis of BART data for testing nonlinear models, see van Ravenzwaaij, Dutilh, and Wagenmakers (2011).

To demonstrate the consequences of only analyzing the trials on which the balloon did not pop (the standard approach to the assessment of performance on the BART), we also ran the same regression but omitted the trials on which the balloon popped, thus eliminating censored trials. The results of this analysis illustrate the degree to which the standard approach underestimates the extent to which people would pump up the balloon.

The Bayesian estimated posterior credibility intervals are shown in Table 1 for both the censored regression and the regression based only on the trials that did not end with a pop. The averages are back-transformed in the text below, for simplicity of presentation. For the more valid censored regression, the participants in the automatic condition were estimated to cash in at a balloon size of 845 units (845 ms in the automatic condition, or about eight or nine pumps in the manual condition), and they were sensitive to balloon thickness ($M = -1.15$; see the left side of Fig. 5), meaning that

Table 1 Parameter estimates (as 95% credibility intervals) for the balloon size analyses

	Experiment 1 Censored	Experiment 1 Truncated	Experiment 2 Censored
Intercept	[6.68, 6.78]	[6.51, 6.59]	[6.75, 6.97]
Power	[– 1.22, – 1.07]	[– 1.49, – 1.35]	[– 0.97, – 0.71]
Condition	[– 0.09, 0.02]	[– 0.05, 0.03]	[0.02, 0.21]
Condition × Power	[0.07, 0.23]	[0.08, 0.23]	[0.08, 0.35]

Power was centered by subtracting 0.75. Condition was effect coded as [– 1 = press, + 1 = no press] for Experiment 1 and [– 1 = inflate to pop, + 1 = inflate to full] for Experiment 2

participants tended to wait longest to cash in balloons at the 0.25 power value (thickest balloons) and to wait least to cash in balloons at the 1.25 power value (thinnest balloons). Participants in the automatic condition cashed in earlier (mean size of 821 units) than participants in the manual condition (mean size of 871 units), indicating that increased effort did not result in cashing in earlier; the difference was small (equivalent to the size difference that one pump produces), and zero was a credible value. Participants in the manual condition were more sensitive to changes in the balloon thickness (mean slope = – 1.30) than were those in the automatic condition (mean slope = – 1.00; see the left side of Fig. 5); zero was not a credible value when estimating this difference.

The standard regression of the balloon sizes at cash-in are also shown in Table 1 and on the right side of Fig. 5. It is readily apparent that excluding popped trials from the analysis results in a downward-biased estimate of the predicted cash-in balloon size (by 146 units of balloon size at the average power of 1.0). This analysis also significantly underestimated the uncertainty in individual differences, because the unobserved longer tail of the distribution had no influence on the assessment of variability; the 95% credibility interval for the standard deviation of the intercepts across participants was [0.19, 0.24] for the standard analysis of truncated values, versus [0.24, 0.31] for the censored regression. Furthermore, the standard method of estimating wait times does not incorporate the differential likelihoods of popping the balloon across the two conditions—27% for the automatic condition and 36% for the manual condition—as well as variations in the likelihood of popping for the different balloon thicknesses (i.e., values of the power parameter). Because the participants in the manual condition popped the balloon more often, there was more truncation of the upper tail, thus driving down their estimated balloon size at cash-in.

Discussion

The purpose of Experiment 1 was to determine whether manually inflating the balloon would produce less waiting than was observed when the balloon inflated automatically, due to an

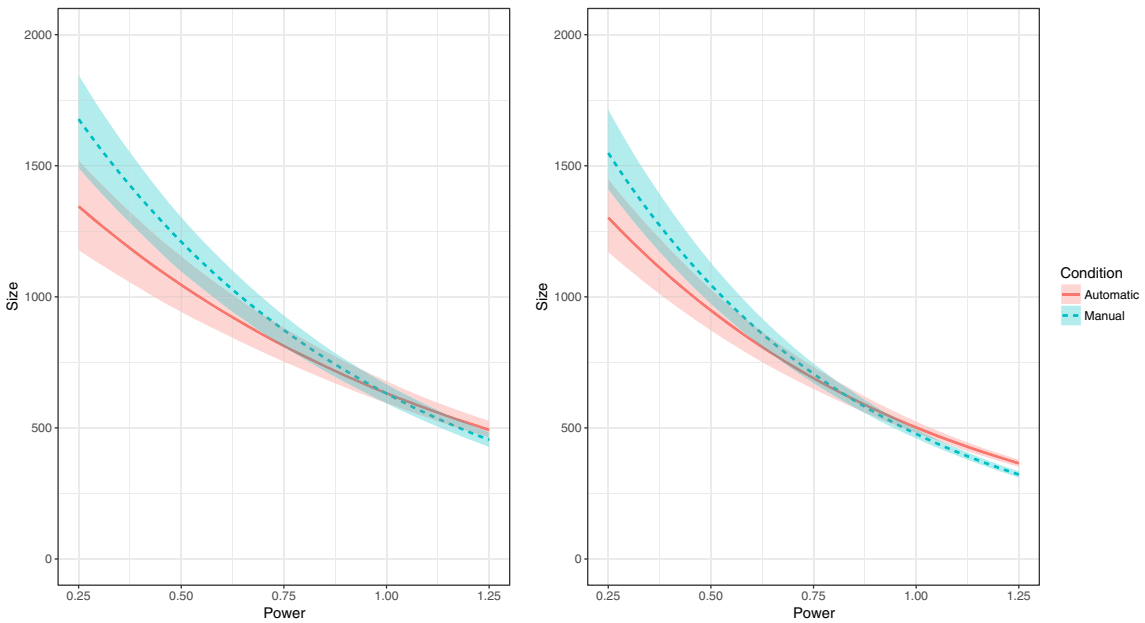

Fig. 5 (Left) Average balloon size at cash-in for each of the conditions in Experiment 1, as estimated by a censored gamma regression. (Right) Average balloon size at cash-in for each of the conditions in Experiment 1, as estimated by a gamma regression of the trials on which the balloon did not pop. Error ribbons show 95% credibility intervals

avoidance of effort. The results suggest that manually inflating did not produce a tendency to cash in earlier, but rather a bit later (equivalent to one pump in size). Thus, we found no evidence that effort plays much of a role in explaining the tendency to cash in a balloon earlier, especially for the thickest balloons, which required even more effort to obtain the greatest reward. The slightly quicker cash-ins for the automatic condition could have been caused by the rapidity at which the balloon inflated when pumping was not required.

A key result was the ability of the censored regression to appropriately estimate the balloon size at which people would cash in. When we analyzed the data using the traditional approach, in which popped trials were omitted, the analysis greatly underestimated the predicted balloon size at cash-in, and overestimated the size of the power slope (see Table 1). However, even with the proper, censored regression, people still cashed in much too early for the thickest balloons. This may have been due to experience with the thinnest balloons carrying over to trials involving the thicker ones—the participants may have defaulted to cashing in faster until they learned that the new balloon color corresponded to a thicker balloon. Previously published work showing early cashing in when only a thick balloon is present suggests that the risk-aversive behavior might persist even without the carryover (Hunt et al., 2005). Unfortunately, these earlier studies used a standard analysis that only considered the number of pumps on trials that did not result in popping. For example, in the Hunt et al. study, 31% of the trials ended in the balloon exploding, resulting in truncation of nearly one-third of the upper tail of the distribution of the number of pumps. Regardless, there is considerable evidence of risk aversion in a related task in which the probability of receiving a reward changed within a trial (Young & Cole, 2012; Young et al., 2014).

The difference in estimates derived form the censored regression and the standard analysis highlights that any assessment of performance on the BART that does not use censored regression must report any differences in popping probability across conditions or participants. It would be easy to assume that the number of pumps and the probability of popping are highly correlated, but participants with higher trial-to-trial variability in their pumping (perhaps reflecting more exploratory behavior) would be more likely to pop the balloon than would a participant with a similar average number of pumps and less variability in that number.

Although the observed tendency to cash in earlier than is optimal could demonstrate risk aversion, especially when larger rewards are more readily available (for the thickest balloons), some degree of underinflation could have been caused by participants wanting to complete the experiment sooner. The differential benefit of cashing in sooner has its biggest effect for the thickest balloons, because they require longer waiting in order to achieve optimal behavior. This explanation was evaluated in Experiment 2.

In Experiment 2 we also evaluated whether the immediate experience of popping contributed to the observed risk aversion. Our laboratory's ultimate goal was to create a variation of the BART in which information about the loss of points would be withheld until the end of the trial, so that the task would be more amenable to use in an electroencephalography (EEG) study of decision making. Popping creates a visual discontinuity that would produce a large effect on the EEG signal that could prove problematic for interpretation. Furthermore, popped trials do not produce a cash-in response for which a response-related EEG signal could be assessed. But, by decreasing the temporal contiguity between a behavior and its consequences (a situation that also exists in the automatic BART; Pleskac et al., 2008), a task without visual popping could decrease the salience of the loss of points associated with the popping, and thus create less risk aversion than would normally be observed.

Experiment 2

In Experiment 2, participants were assigned to one of two conditions, both of which involved automatic inflation of the balloons. In the first condition (*inflate to pop*), when the balloon popped it disappeared as usual, but the screen remained blank until the programmed presentation time of the feedback (3,400 ms after each balloon began inflating). This ensured that each trial had the same duration, and it removed any incentive to cash in earlier to more quickly complete the experiment. The second condition (*inflate to full*) was identical to the first, except that the balloon did not disappear when the balloon popped. Thus, there was no indication of the pop time until after the end of the trial, when the feedback screen was displayed; participants who had responded after the scheduled pop time earned no points on that trial. This condition allowed the participant to respond at any time during the 3,400-ms balloon inflation, and thus also served to prevent any data truncation caused by the popping.

Method

Participants A total of 54 introductory psychology students at Kansas State University received course credit for their voluntary participation. None of these participants took part in Experiment 1. Given the absence of demographic effects in Experiment 1, we used a smaller sample and did not collect demographic information in Experiment 2.

Stimuli and procedure The stimuli and procedures were identical to those of Experiment 1, with the following exceptions. The inflate-to-pop condition was a conceptual replication of the automatic condition from Experiment 1. In this condition, the balloon animation terminated at the preprogrammed pop time or whenever the participant cashed in the balloon. However, this condition was modified from the automatic

version used in Experiment 1, such that the length of the trial was held constant. When the balloon animation terminated, the screen remained solid white until the end of the 3,400-ms interval (the maximum length of time the balloon could have been on the screen), at which point participants were informed of the amount of points earned on the trial. In the inflate-to-full condition, participants were instructed that the balloon would inflate to the same size on every trial, regardless of the cash-in latency and predetermined pop-time; the number of points earned was displayed at the same time as in the other condition. Because the purpose of the study was strictly methodological, we did not collect individual differences variables, but their characteristics would be similar to those from the first experiment, because they were drawn from the same population.

We again ran a generalized linear mixed-effect censored regression on balloon size at cash-in and specified a gamma error distribution (with a log link function). In the inflate-to-full condition, none of the outcome values were censored, by design. The default priors were $N(0.0, 0.5)$ for the regression weights, but we used a weakly informative prior for the intercept, $N(6.0, 1.0)$, and the power slope, $N(-1.0, 0.5)$ (the latter prior was different from that used in Exp. 1 because it was informed by those results), a weakly informative *Cauchy*(0.4, 0.2) for the standard deviations, and *LKJ*(2.0) for the correlations. Otherwise, the analysis was identical to the censored regression used in Experiment 1. We should note that the narrower priors on the power slope and standard deviation had minimal impact on these results—the credibility intervals were nearly identical to those from an analysis using uninformative priors. Regardless, the use of these more informed priors was in the original analysis plan, and thus was adopted.

Results and discussion

All participants completed the task within the allotted time of 1 h. Cash-in latencies below 100 units in size were excluded for being too fast, which resulted in 126 trials being excluded (less than 1% of the data); the metric of balloon size corresponds to the number of milliseconds waited. Four participants stopped responding (after 6, 27, 125, and 141 trials). Because all were in the inflate-to-full condition, their data before they stopped responding were retained, but the data after they stopped was dropped from the analysis.

Plots of the full distribution, geometric means, and optimal size (to maximize the expected value) are shown in the bottom part of Fig. 4. It is again apparent that participants were cashing in much earlier than was optimal for the thickest balloons (powers less than 0.75), but they actually appear to have cashed in later than optimal in the inflate-to-full condition for the thinner balloons (powers of 1.00 and 1.25). These distributions and geometric means are based on collapsing across all participants, regardless of sample size.

The Bayesian estimated posterior credibility intervals are shown in the right column of Table 1, and the model fit is shown in Fig. 6. Participants in the inflate-to-full condition were estimated to cash in at a balloon size of 1,075 units (or 1,075 ms), whereas those in the inflate-to-pop condition cashed in much earlier, at a size of 846 units. Additionally, in the inflate-to-full condition the participants showed weaker sensitivity to the thickness manipulation (mean slope = – 0.62), whereas participants in the inflate-to-pop condition showed stronger sensitivity (mean slope = – 1.06). For the condition differences in average cash-in time (i.e., intercept) and sensitivity to power (i.e., slope), zero was not a credible value (see the bottom two rows in the right column of Table 1).

The observed estimates for the inflate-to-pop condition in which the duration of trials was constant were very similar to those in the Experiment 1's automatic condition (mean balloon sizes of 812 in Exp. 1 and 846 in Exp. 2, mean power slopes of – 1.03 in Exp. 1 and – 1.06 in Exp. 2). This observation suggests that the participants' tendency to cash in earlier than was optimal was not strongly driven by a desire to shorten the experiment. However, the difference between the inflate-to-full and inflate-to-pop conditions suggests that experiencing the popping is a significant contributor to the risk aversion observed in the BART task.

In a final exploratory analysis, we examined how behavior changed across the balloons within each set of ten trials, as well as how it changed across the 25 sets of balloon types (see Fig. 3 for the design). A Bayesian censored regression was fit by including trial (within set) and set (1 to 25). The results are shown in Fig. 7. Not surprisingly, at the beginning of a set of ten trials there is no real sensitivity to the balloon thickness, and sensitivity increased significantly across the ten trials. There was also evidence of learning across the experiment, with the response to the thin balloons in the inflate-to-full condition showing the largest change across the experiment. By the last set of balloons, the behavior in the two conditions was very similar. The slower development of strong sensitivity to balloon thickness in the inflate-to-full condition suggests that participants were slower to learn the task contingencies when the feedback was delayed to the end of the trial, rather than immediate (recall that in the inflate-to-pop condition, the popping provided immediate feedback that the participant would earn zero points at the end of the trial). Finally, at the end of the experiment, participants were still cashing in sooner than was optimal for the two thickest balloons.

General discussion

These two experiments suggest that the tendency to respond earlier than is optimal in the BART is not a by-product of a desire to reduce effort or to complete the task early. The earlier-than-optimal cash-ins were robust for the thickest

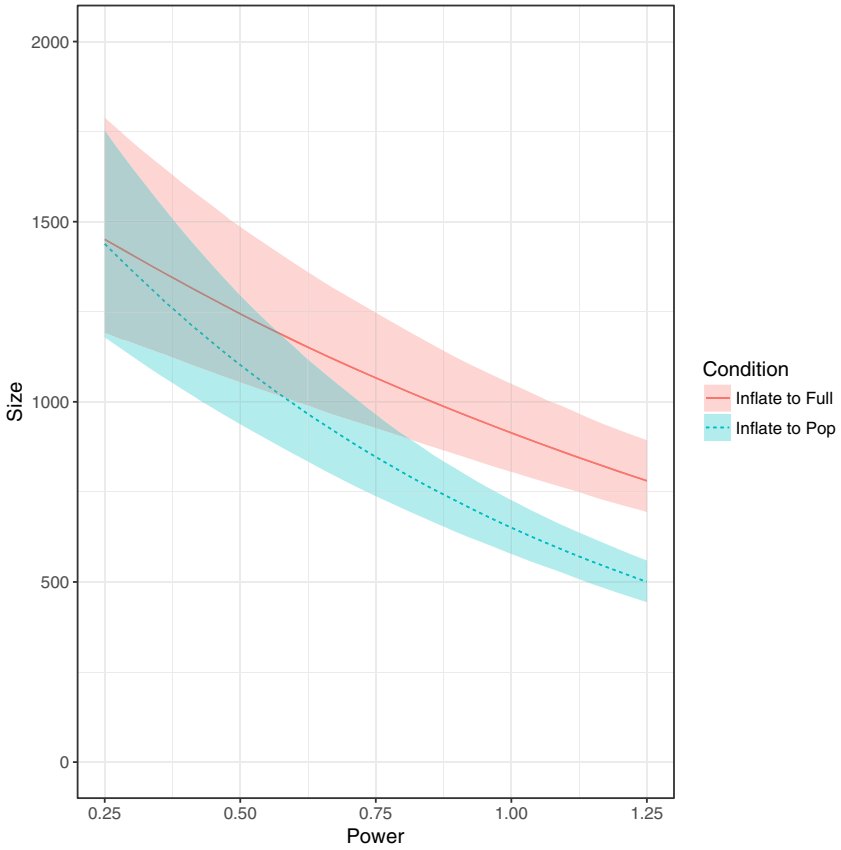

Fig. 6 Average balloon size at cash-in for each of the conditions in Experiment 2, as estimated by a censored gamma regression. Error ribbons show 95% credibility intervals

balloons with the least risk. In contrast, the immediate experience of popping appears to be a major contributor to risk aversion. Because experiencing the popping is more common for the thinner balloons (higher power values), it was not surprising that the condition difference in Experiment 2 was larger for the thinner balloons (see Fig. 6).

The majority of the published studies involving the BART have only used the thickest balloons (e.g., Hopko et al., 2006; Lighthall, Mather, & Gorlick, 2009; Pleskac et al., 2008; Wright & Rakow, 2017), because greater individual differences were observed in this condition by Lejuez et al. (2002). The robust tendency to avoid risk in precisely the conditions in which risk is lowest may be a consequence of more reward being available in those conditions. Acheson, Reynolds, Richards, and De Wit (2006) documented that participants earning 1¢ per pump were less risk-averse (M = 48 pumps) than those earning 5¢ per pump (M = 37 pumps) or 25¢ per pump (M = 31 pumps). For the thinnest balloons in both the present study and the one reported by Lejuez et al. (2002), very little reward was at stake, because these balloons invariably popped before much reward was available. This raises the question of whether the thinner balloons might prove of greater utility to studying risk aversion if the reward rate increased to offset the greater risk. For example, if Lejuez et al.'s (2002) orange balloons (with an optimal pump number of 4) accumulated 16 times as much reward for each pump than did the blue balloons (with an optimal pump number of 64), participants might cash in much earlier and, perhaps, show the larger degree of individual differences that Lejuez et al. (2002) observed for the blue balloons.

Although the present studies strongly indicate that effort reduction and time reduction played little role in the high level of risk aversion for thicker (less risky) balloons, the truncation of wait durations caused by popping clearly has an effect. Regardless, participants in the inflate-to-full condition still showed risk aversion for the thick balloons when a censored regression was used (see Fig. 6), despite their experiencing many fewer pops for those balloons. This behavioral pattern gives rise to an alternative explanation of both of these tendencies: carryover effects. Because participants experienced multiple balloon types within a session, they may have evidenced carryover from the other balloon thickness conditions.

For example, when a new balloon type is encountered, the default behavior may be to assume that the expected pop time will be the typical pop time across all balloon types; only by experiencing many balloons of the same thickness (and behaviorally sampling a range of wait times) would a participant begin to learn the optimal behavior. Indeed, Fig. 7 reveals that

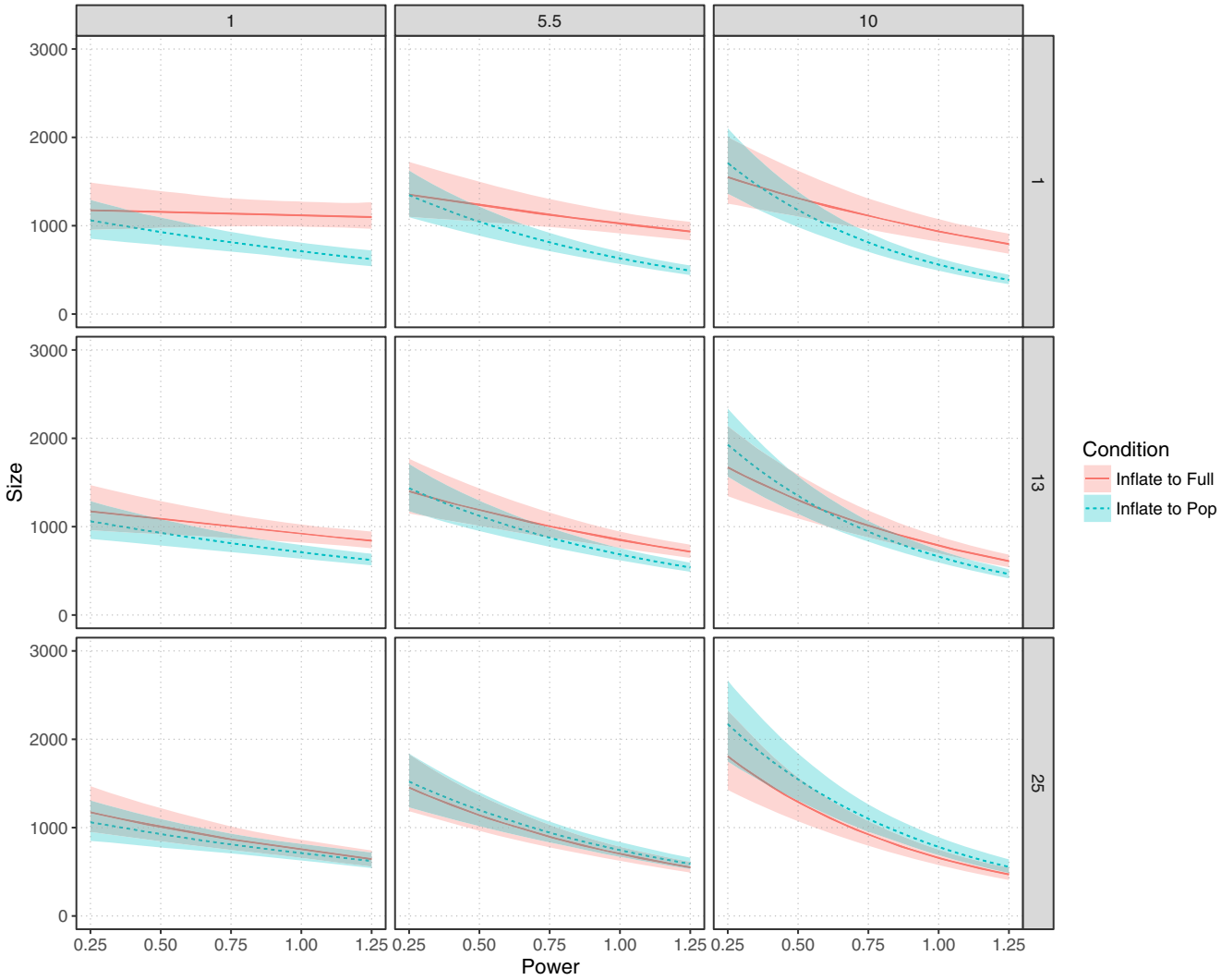

Fig. 7 Average balloon size at cash-in for each of the conditions in Experiment 2, as estimated by a censored regression. Each column represents behavior at the beginning, middle, and end of each set of ten identical balloons. Each row represents the behavior for the 1st, 13th, and 25th set of balloons (i.e., the beginning, middle, or end of the experiment). Error ribbons show 95% credibility intervals

at the beginning of each set of ten balloons, participants tended to cash in at a balloon size of around 1,000, which is the optimal size for the medium-thickness balloon (power = 0.75).

This argument raises the possibility that participant differences in risk aversion may also reflect participant differences in their ability to learn the optimal behavior. A participant who is slow to adjust behavior in response to changing task contingencies would produce weak sensitivity to balloon thickness. Pleskac et al. (2008) found nearly optimal behavior in their automated BART task, but interpreting that result is complicating by an additional methodological change: they informed participants of the optimal number of pumps before beginning the experiment. Indeed, an experiment involving a variation of the automatic BART that did not inform participants of the optimal number of pumps produced many fewer pumps (33 adjusted pumps vs. Pleskac et al.'s 56 adjusted pumps; Wright & Rakow, 2017). Even if behavior approaches optimality under ideal conditions (repeated exposure to the same risk level), individual differences in the rate at which optimality is achieved would need to be examined as an alternative predictor of those behaviors already demonstrated to correlate with risk aversion in the BART. Addressing this question is beyond the scope of the present project but is worthy of study.

The inflate-to-full condition in Experiment 2 is well-designed for use in studies involving EEG and functional magnetic resonance imaging (fMRI). By eliminating the response requirement in the standard BART, the removal of a surfeit of motor activity would make the interpretation of EEG and fMRI more straightforward. Furthermore, by withholding feedback until after the balloon is fully inflated, stimulus confounds are eliminated, because participants experience the same stimulus on each trial. However, the BART's established

relationship to clinically relevant phenomena (Aklin et al., 2005; Hopko et al., 2006; Hunt et al., 2005; Lejuez et al., 2003) is not guaranteed to generalize to the inflate-to-full version, in which behavior indicated less risk aversion by the participants (at least earlier in the experiment). For example, if the established relationships hinge on participants immediately experiencing the pop, then none of the reported correlations would be observed for this new version of the BART. Regardless, the inflate-to-full version does have task characteristics involving risk aversion, sensitivity to changes in risk, and performance feedback that can uncover important results regarding the neural basis of these environmental variables and individual differences (McCoy, 2015).

How people manage the trade-off between magnitude, risk, effort, and time applies to everyday behaviors as varied as eating, drug use, driving, and investing. The BART has inspired a wide range of important investigations that can provide insight into more than just the trade-off between magnitude and risk. We hope that further investigations involving variations on this valuable task will continue this trend, and that a more careful investigation of the individual differences in the behavioral variables at play will provide novel insights regarding risk aversion, sensitivity to changing levels of risk, variability in exploratory behavior, and the ability to learn to behave more optimally. Finally, we encourage researchers to use a censored regression approach to analyzing BART data in order to seamlessly integrate the observed latency to respond with the censoring that occurs when the balloon pops before the participant cashes in.

Acknowledgements This project was supported by a grant from the National Institute of General Medical Science GM113109 of the National Institutes of Health.

References

Acheson, A., Reynolds, B., Richards, J. B., & De Wit, H. (2006). Diazepam impairs behavioral inhibition but not delay discounting or risk taking in healthy adults. *Experimental and Clinical Psychopharmacology*, *14*, 190–198.

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. *Behavior Research and Therapy*, *43*, 215–228. https://doi.org/10.1016/j.brat.2003.12.007

Chatterjee, S., & Heath, T. B. (1996). Conflict and loss aversion in multiattribute choice: The effects of trade-off size and reference dependence on decision difficulty. *Organizational Behavior and Human Decision Processes*, *67*, 144–155. https://doi.org/10.1006/obhd.1996.0070

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, *25*, 257–271. https://doi.org/10.3758/BF03204507

Crowley, M. J., Wu, J., Crutcher, C., Bailey, C. A., Lejuez, C. W., & Mayes, L. C. (2009). Risk-taking and the feedback negativity response to loss among at-risk adolescents. *Developmental Neuroscience*, *31*, 137–148.

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-naïve alcoholics. *Drug and Alcohol Dependence*, *92*, 141–148. https://doi.org/10.1016/j.drugalcdep.2007.07.017.

Hopko, D. R., Lejuez, C. W., Daughters, S. B., Aklin, A. W., Osborne, A., Simmons, B. L., & Strong, D. R. (2006). Construct validity of the Balloon Analogue Risk Task (BART): Relationship with MDMA use by inner-city drug users in residential treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*, 95–101. https://doi.org/10.1007/s10862-006-7487-5

Hunt, M. K., Hopko, D. R., Bare, R., Lejuez, C. W., & Robinson, E. V. (2005). Construct validity of the Balloon Analog Risk Task (BART): Associations with psychopathy and impulsivity. *Assessment*, *12*, 416–428. https://doi.org/10.1177/1073191105278740

Jentsch, J. D., Woods, J. A., Groman, S. M., & Seu, E. (2010). Behavioral characteristics and neural mechanisms mediating performance in a rodent version of the Balloon Analog Risk Task. *Neuropsychopharmacology*, *35*, 1797–1806. https://doi.org/10.1038/npp.2010.47

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., & Read, J. P. (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., . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75–84. https://doi.org/10.1037/1076-898X.8.2.75

Lighthall, N. R., Mather, M., & Gorlick, M. A. (2009). Acute stress increases sex differences in risk seeking in the balloon analogue risk task. *PLoS ONE*, *4*, e6002. https://doi.org/10.1371/journal.pone.0006002

McCoy, A. W. (2015). Factors affecting the amplitude of the feedback-related negativity on the balloon analogue risk task (Master's thesis), Kansas State University, Manhattan, KS.

Pinheiro, J. C., & Bates, D. M. (2004). Mixed-effects models in S and S-PLUS. New York, NY: Springer.

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, *16*, 555–564. https://doi.org/10.1037/a0014245

Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior*, *55*, 233–244.

Rodriguez, G. (2007). Multilevel generalized linear models. In J. de Leeuw & E. Meijer (Eds.), Handbook of multilevel analysis (pp. 337–378). New York, NY: Springer.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*, 94–105. https://doi.org/10.1016/j.jmp.2010.08.010

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. https://doi.org/10.1037/0033-295X.112.4.862

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*, 669–689.

Wright, R. J., & Rakow, T. (2017). Don't sweat it: Re-examining the somatic marker hypothesis using variants of the balloon analogue risk task. *Decision*, *4*, 52–65. https://doi.org/10.1037/dec0000055

Young, M. E., & Cole, J. J. (2012). Human sensitivity to the magnitude and probability of a continuous causal relation in a video game. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 11–22. https://doi.org/10.1037/a0026357

Young, M. E., Webb, T. L., & Jacobs, E. A. (2011). Deciding when to "cash in" when outcomes are continuously improving: An escalating interest task. *Behavioural Processes*, *88*, 101–110.

Young, M. E., Webb, T. L., Rung, J. M., & McCoy, A. W. (2014). Outcome probability versus magnitude: When waiting benefits one at the cost of the other. *PLoS ONE*, *9*, e98996. https://doi.org/10.1371/journal.pone.0098996